

## CONTINUOUS BLOOD GLUCOSE MONITORING: A BAYES-HIDDEN MARKOV APPROACH

Lee H. Dicker<sup>1</sup>, Tingni Sun<sup>2</sup>, Cun-Hui Zhang<sup>1</sup>, D. Barry Keenan<sup>3</sup> and  
Larry Shepp<sup>2</sup>

<sup>1</sup>*Rutgers University*, <sup>2</sup>*University of Pennsylvania*,  
and <sup>3</sup>*Medtronic MiniMed*

*Abstract:* Real-time monitoring of blood glucose density is essential for managing diabetes. Continuous glucose monitoring (CGM) systems have been developed to help address this need. Many CGM systems are built around an electrochemical biosensor that may be inserted into the subcutaneous tissue of an individual and allows for nearly continuous monitoring of an electrical current generated by glucose molecules near the sensor site. This electrical current is correlated with blood glucose density and, in principle, provides a means for real-time monitoring of blood glucose density. One of the major challenges in CGM is developing algorithms for converting sensor measurements into accurate estimates of blood glucose density in real time. In this paper, we describe fundamental statistical problems that arise in developing CGM algorithms. We propose statistical algorithms based on Kalman filtering, nonparametric empirical Bayes methods, and ideas from sequential change-point detection, and apply them to a very rich CGM dataset. The performance of our methods compares favorably to that of an existing widely used CGM algorithm. A simulation study sheds light on other interesting and important aspects of the problem. More broadly, this paper highlights an important application that has received little attention in the statistics literature and our results suggest that the appropriate application of statistical methodology may lead to significant contributions in diabetes technology research.

*Key words and phrases:* Diabetes technology, Kalman filtering, nonparametric empirical Bayes, sequential methods

### 1. Introduction

The U.S. Centers for Disease Control and Prevention estimate that over 25 million people in the U.S. have diabetes mellitus (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (2011)). Furthermore, according to the same source, in the U.S.,

Diabetes is the leading cause of kidney failure, nontraumatic lower-limb amputations, and new cases of blindness among adults... [it] is a major cause of heart disease and stroke... [and it] is the seventh leading cause of death.

Diabetes results from defects in an individual's ability to produce or use insulin, and is characterized by persistent high blood glucose levels. There is no cure for diabetes, but substantial progress has been made in developing tools and techniques for managing the disease. Primary among these tools was the invention of injectable insulin 90 years ago (Bliss (2007)). More recently, insulin pumps, multiple injection regimens, and various insulin analogs have been introduced to help diabetes patients control their blood glucose levels and simplify their health maintenance regimens. The effectiveness of tools like these for treating diabetes has been validated in landmark studies, such as the Diabetes Control and Complications Trial (DCCT), which found that intensive insulin treatment can substantially reduce risks related to kidney failure and blindness that are associated with type 1 diabetes (Reichard, Nilsson, and Rosenqvist (1993)). However, the DCCT also found that implementing more aggressive methods for glycemic control increased the risk for severe hypoglycemia (episodes of dangerously low blood glucose levels) by up to 200%.

A key to successfully utilizing any of the previously mentioned tools for glycemic control, which is highlighted by the results of the DCCT, is the availability of reliable real-time measurements of a patient's blood glucose density. The development of electrochemical glucose biosensors that may be inserted into a patient's subcutaneous tissue provides a means for continuously monitoring blood glucose density (Klonoff (2005); Wang (2008)). Ultimately, this technology (continuous glucose monitoring, or CGM), when used in conjunction with an insulin pump, could lead to the widespread adoption of the "artificial pancreas" – a closed-loop system for glycemic management, wherein glucose levels are tracked by CGM and the results are used to determine the rate at which insulin should be pumped into the body (Bequette (2005); Klonoff (2007)). Such a device has the potential to provide life-altering benefits to some of the millions of diabetics in the U.S. and throughout the world.

Though a great deal of progress towards the artificial pancreas has been made, major challenges remain (Hovorka (2006); Harvey et al. (2010)). Many of the most significant challenges are statistical or algorithmic in nature. For instance, accurate CGM calibration algorithms for estimating blood glucose density in real-time are essential for the viability of an artificial pancreas (Kowalski (2009)). Furthermore, given the many sources of noise and variability that exist in the human body and CGM, it is evident that advanced statistical methods will play a critical role in the development of these algorithms. In this paper, we develop a rigorous statistical framework for continuous blood glucose monitoring and show that methods derived within this framework can lead to significant improvements in the overall accuracy of CGM algorithms. These methods provide online estimates of blood glucose density that, in principle, may be used to make real-time decisions relating to glucose management.

Motivated by biological and physical models for blood glucose biosensors and continuous glucose monitoring, we propose and implement several CGM algorithms. The Kalman filter (Kalman (1960)) plays a key role in most (but not all) of the algorithms proposed in this paper. More broadly, all of the methods described here are closely connected with Bayesian methods and hidden Markov models (HMMs). Ideas from the theory of sequential change-point detection (Siegmund (1985); Lai (1998)) play a key role in our top performing algorithms. In addition to conducting extensive simulation studies, our statistical methods have been developed, tested, and evaluated with the aid of the “Star 1” dataset that is described in Section 2.2. Our analysis shows that the methods described in this paper deliver substantial improvements in CGM accuracy when compared to an existing widely-used algorithm. Additionally, and perhaps more importantly, the methods proposed here may help lay the groundwork for future work on statistical problems in glycemic control.

The rest of this paper is structured as follows. In Section 2, we describe more details of the CGM systems studied in this paper and the Star 1 dataset. In Section 3, we discuss mathematical and statistical models relating blood glucose density to sensor measurements and the data typically available for developing CGM algorithms. This provides a statistical framework for the development of sensor calibration algorithms. As initially formulated, the models are quite general. In order to implement effective, statistically sound calibration algorithms based on these models, further specification is required. In Sections 4-5, we discuss how different distributional assumptions for the model in Section 3 lead to different sensor calibration algorithms and describe several algorithms in detail. Statistical inference and prediction intervals for estimated blood glucose density are discussed in Section 6. Section 7 contains the results of a simulation study. Results from our analysis of the Star 1 dataset are reported in Section 8. Section 9 contains a concluding discussion.

## 2. Additional Background

### 2.1. Blood glucose biosensors

An electrochemical blood glucose biosensor is at the core of most CGM systems currently in use. These sensors, which are inserted into subcutaneous tissue, consist of a single electrode coated with an enzyme which reacts with glucose, such as glucose oxidase (Wang (2008)). When a glucose molecule from the interstitial fluid that permeates subcutaneous tissue contacts the electrode, a current is generated in the sensor. Furthermore, the current generated in the sensor increases with the rate at which glucose molecules contact the sensor. By measuring the current, one obtains a surrogate for blood glucose density: If the blood glucose density is high, the density of glucose molecules in the interstitial

fluid near the sensor will likely be high, leading to a high current measurement in the sensor. The association between blood glucose density and current measurements in the sensor explains the usefulness of the sensor for measuring blood glucose density. However, calibrating the sensor, i.e., converting the current measurements into more interpretable and accurate measures of blood glucose density, is essential for clinical relevance. Indeed, this is the central problem of this paper and a significant obstacle to creating an effective artificial pancreas.

In a review article on blood glucose biosensors, Wang (2008) mentions that in addition to glucose, other electroactive substances (such as ascorbic and uric acid) may be found in the interstitial fluid and contribute to current generated in the biosensor. This is a potential source of noise when using these biosensors to measure blood glucose density. However, extensive research has been conducted to minimize this type of noise and increase selectivity of blood glucose biosensors. Indeed, according to Wang (2008), “remarkably high selectivity... has been obtained.”

## 2.2. The Star 1 dataset

In the Star 1 study (Hirsch et al. (2008)) 137 subjects with type 1 diabetes were followed for 6 months, on average, while using a CGM device (developed by Medtronic MiniMed). For each patient in the study, an electrical current measurement from the blood glucose sensor was recorded every 5 minutes; the measurement (in nano amps, nA) at time  $t$  is denoted  $\text{ISIG}(t)$  throughout this paper (ISIG is an abbreviation for interstitial signal; Table 1 contains a list of all abbreviations commonly used by us). Less frequently – approximately every 6 hours, on average – patients in the study recorded a more accurate measure of blood glucose density, obtained via fingerstick (a more invasive and time consuming, yet prevalent, procedure where the patient pricks their finger to obtain a small droplet of blood; the blood droplet is then analyzed by a blood glucose meter). The fingerstick blood glucose measurement (in mg/dL) at time  $t$  is denoted  $\text{FS} = \text{FS}(t)$ . These measurements are entered into the CGM system and are essential for calibrating CGM algorithms. Ultimately, for an artificial pancreas, one hopes to minimize the number and frequency of fingersticks required for accurate calibration of CGM algorithms. However, we do not address this further.

To summarize, and introduce some additional notation, the observed data consists of discretely sampled observations from  $\text{ISIG}(t)$  and  $\text{FS}(t)$ :

$$\{\text{ISIG}(\tilde{\tau}_k)\}_{k=0}^{\infty} \text{ and } \{\text{FS}(\tau_k)\}_{k=0}^{\infty},$$

where the sequences of times  $\{\tilde{\tau}_k\}$  and  $\{\tau_k\}$  are assumed to be increasing, and  $\{\tau_k\} \subseteq \{\tilde{\tau}_k\}$  (i.e.,  $\text{ISIG}(t)$  is sampled more frequently than  $\text{FS}(t)$ ). We take  $\text{ISIG}_k = \text{ISIG}(\tilde{\tau}_k)$  and  $\text{FS}_k = \text{FS}(\tau_k)$ .

Table 1. Commonly used abbreviations.

Abbreviation	Definition
CGM	An abbreviation (i) for the general phrase “continuous glucose monitoring” and (ii) for the estimated blood glucose density (mg/dL) derived from a previously existing proprietary algorithm. It will be clear from the context which definition is being referenced.
ISIG	Interstitial signal (nA). The current measured by a CGM device. The goal of CGM algorithms discussed here is to convert ISIG into an accurate estimate of blood glucose density.
FS	Fingerstick glucose density (mg/dL). Fingerstick measurements of blood glucose density. These measurements are taken intermittently (approximately every 6 hours by patients in the Star 1 study) and used to calibrate CGM algorithms. FS measurements are taken to be the gold standard for blood glucose density in our analysis of the Star 1 dataset.
IG	Interstitial glucose density (mg/dL). The glucose density in interstitial fluid near the blood glucose biosensor. ISIG increases with IG.
BG	Blood glucose density (mg/dL). The primary goal of the methods considered in this paper is to accurately estimate BG. Glucose diffuses from the blood into interstitial fluid, where it may be measured by the blood glucose biosensor.
MARD	Mean absolute relative difference. Defined in (7.3). A measure of accuracy for CGM algorithms.

The primary sources of information for developing CGM algorithms are FS and ISIG. Previous FS and ISIG values along with the current ISIG value are used to estimate the current blood glucose density. However, other useful information available in the Star 1 dataset includes (i) sensor identification codes and (ii) the output of a previously existing, proprietary CGM algorithm for estimating blood glucose density in real-time. Due to biofouling and other issues, the electrochemical sensor in a CGM system must be replaced periodically (in the Star 1 dataset, every 2.7 days on average). The sensor ID codes allow one to monitor when sensors are replaced and to control for effects of sensor degradation and variability in our algorithms. The output of the previously existing CGM algorithm, denoted  $CGM(t)$ , provides a standard of comparison by which we can judge the performance of the methods proposed here. Fingerstick measurements  $FS(t)$ ,  $ISIG(t)$ , and  $CGM(t)$  are plotted for a representative subject in Figure 1, along with indicators for when sensors were replaced.

### 3. A Model for Blood Glucose Biosensors and Continuous Glucose Monitoring

The electrical current  $ISIG(t)$  is correlated with the glucose density in inter-

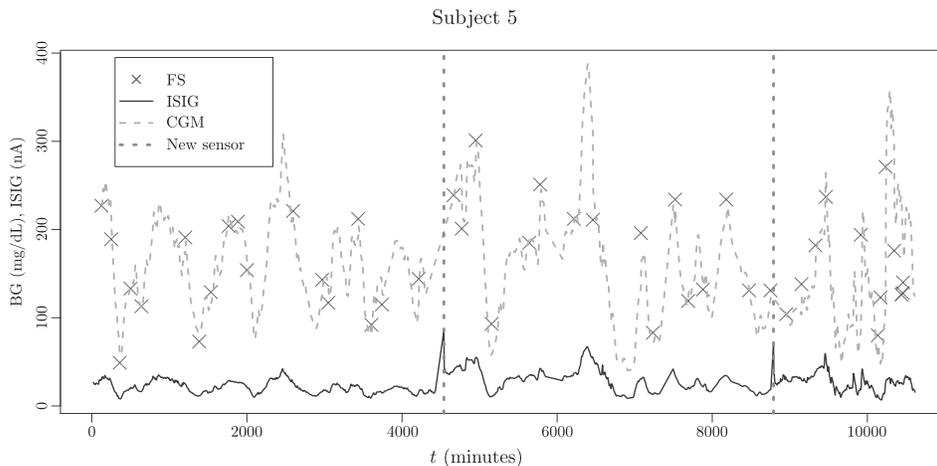


Figure 1.  $FS(t)$ ,  $ISIG(t)$ ,  $CGM(t)$ , and sensor replacement times for Subject 5 in Star 1 dataset.

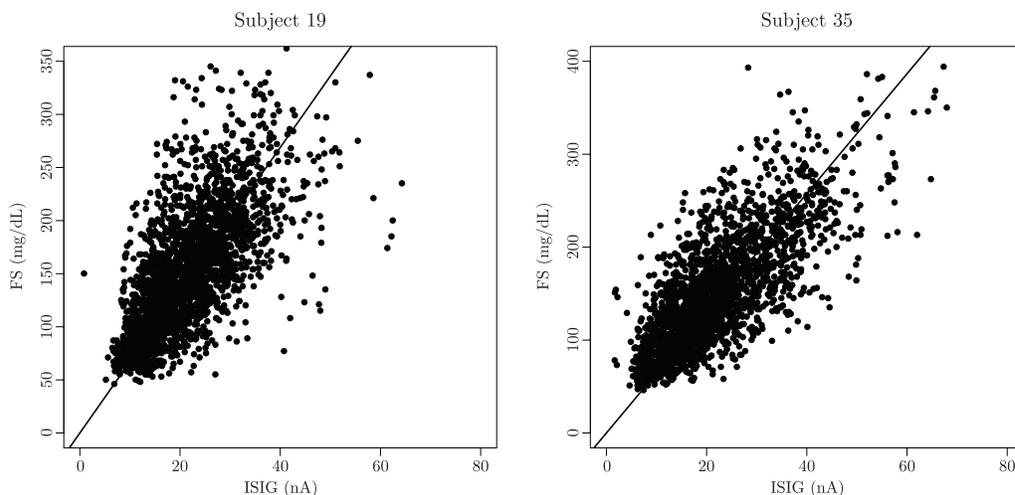


Figure 2.  $FS$  vs.  $ISIG$  for two subjects from the Star-1 study. Sample correlation: 0.66 (Subject 19); 0.76 (Subject 35).

stitial fluid near the sensor site at time  $t$ , which we denote by  $IG(t)$ . Empirical and theoretical evidence suggests that there is an approximately linear relationship between  $ISIG(t)$  and  $IG(t)$  (Koschwanetz and Reichert (2007); Heller and Feldman (2008)). This approximation is supported by Figure 2, which contains plots of  $\{(ISIG_k, FS_k)\}$  for two subjects in the Star 1 study (here,  $FS_k$  is used as a surrogate for  $IG(\tau_k)$  – this is discussed further below). Our basic model

relating ISIG( $t$ ) and IG( $t$ ) is

$$\text{ISIG}(t) = \alpha(t)\text{IG}(t), \quad (3.1)$$

where  $\alpha(t)$  is a slowly varying stochastic process. There is some debate about whether an intercept term should be included in (3.1) (Steil et al. (2005); Wang (2008)). In other words, whether (3.1) or  $\text{ISIG}(t) = \alpha_0(t) + \alpha_1(t)\text{IG}(t)$  for some slowly varying processes  $\alpha_0(t)$  and  $\alpha_1(t)$  is the more appropriate. In our investigations, we have not found a substantial benefit by including an intercept term. Thus, for simplicity, we prefer (3.1). However, all of the methods discussed below can easily be extended to models with an intercept term. Because of issues like sensor degradation, it is known that the relationship between ISIG( $t$ ) and IG( $t$ ) varies with time. Typically, this variation occurs slowly over a relatively large time scale. Allowing for a time-varying process  $\alpha(t)$  is a natural way to account for this time-dependent relationship and helps provide the foundation for a comprehensive statistical framework for the problem. Moreover, we have found that this approach may lead to substantial improvements in the overall performance of CGM algorithms.

Equation (3.1) describes the relationship between ISIG( $t$ ) and IG( $t$ ). On the other hand, we are primarily interested in blood glucose density at time  $t$ , BG( $t$ ), which is never directly observed. Instead, fingerstick measurements of blood glucose density FS( $t$ ) are observed at discrete times  $t = \tau_k$ . These measurements are known to be error prone (Khan et al. (2006)), however, they are generally more reliable than CGM measurements and are the gold standard measurement for blood glucose density in our analysis of the Star 1 data. This suggests the relationship

$$\text{FS}_k = \text{BG}_k + \epsilon_k, \quad (3.2)$$

where  $\text{BG}_k = \text{BG}(\tau_k)$  and  $\epsilon_1, \epsilon_2, \dots$  are mean 0 iid error terms. There have been multiple studies about potential biases in FS $_k$  (Brunner et al. (1998); Cohen et al. (2006); Khan et al. (2006); Kristensen et al. (2008)). Overall, the results of these studies seem inconclusive (Mahoney and Ellison (2007)). However, one might consider incorporating these potential biases in FS $_k$  into future CGM calibration algorithms (i.e., allow for the possibility that  $E(\text{FS}_k) \neq \text{BG}_k$ ). Equations (3.1)–(3.2) link the observed quantities ISIG( $t$ ) and FS( $t$ ) to the unobserved quantities IG( $t$ ) and BG( $t$ ). It remains to describe the relationship between IG( $t$ ) and BG( $t$ ). There is a well-known time lag associated with the diffusion of blood glucose molecules into interstitial fluid (Steil et al. (2005); Wei et al. (2010)). This implies that there may be a lag between interstitial glucose density and the actual blood glucose density. Steil et al. (2005) proposed a two-compartment model to quantify this lag. A simple variant of this model is

$$\text{IG}(t) = \int_0^\infty \text{BG}(t-u)\rho^{-1}e^{-u/\rho} du, \quad (3.3)$$

where  $\rho \geq 0$  is a constant related to the duration of the time lag. Differentiating the previous integral equation yields  $BG(t) = IG(t) + \rho IG'(t)$ . In order to account for variability in the diffusion rate  $\rho$ , we replace it with a slowly varying process  $\rho(t)$  to obtain

$$BG(t) = IG(t) + \rho(t)IG'(t). \quad (3.4)$$

Equations (3.1)–(3.4) determine a model that specifies the relationship between the observed data  $\{ISIG(\tilde{\tau}_k)\}$ ,  $\{FS(\tau_k)\}$  and the unknown quantity to be estimated,  $BG(t)$ . Given this model, the statistical problem is to devise methods for estimating the blood glucose density  $BG(t)$  using the observed data up to time  $t$ ,  $\mathcal{F}_{t-} = \sigma(\{ISIG(\tilde{\tau}_k); \tilde{\tau}_k < t\}, \{FS(\tau_k); \tau_k < t\})$ , and the current sensor measurement  $ISIG(t)$ . In other words, our goal is to develop methods for online estimation of  $BG(t)$ . Additional information, such as sensor age, may also be included in  $\mathcal{F}_{t-}$ . This is discussed further in the next section, where we describe additional distributional assumptions for (3.1)–(3.4) that provide motivation for several novel, high-performing blood glucose estimation methods.

#### 4. Prediction Methods: $\rho(t) \equiv 0$

To simplify things, our initial working assumption is that  $BG(t)$  does not in fact depend on  $IG'(t)$ , i.e.,  $\rho(t) \equiv 0$ . Thus, (3.1)–(3.2) and (3.4) imply

$$FS_k = \beta_k ISIG_k + \epsilon_k, \quad (4.1)$$

where  $\beta_k = \beta(\tau_k)$  and the process  $\beta(t)$  captures the evolution of the relationship between  $FS_k$  and  $ISIG_k$  over time. Despite scientific evidence that suggests  $IG'(t)$  is related to  $BG(t)$  and, thus, the model (4.1) is biased, we believe that (4.1) is a reasonable starting point for our analysis. Indeed, given the noise inherent in CGM data, it is challenging to obtain reliable estimates of the derivative  $IG'(t)$  (or proxies for  $IG'(t)$ ; this is addressed in more detail in Section 5 below) and it seems sensible to begin with a simpler model.

The absence of  $IG'(t)$  in (4.1) simplifies matters considerably and, as seen in Section 8, reasonable blood glucose estimation methods can be developed under this assumption. In Sections 4.1 and 4.2 we discuss two approaches to estimating  $BG(t)$ . The first is highly parametric – strict distributional (normality) assumptions are made. Though these assumptions can not be strictly true in any given real-world situation, the Kalman filtering algorithms derived under these assumptions seem to perform quite well in practice (see Section 8). In our view, the usefulness of these methods is evidenced by their strong practical performance, despite potential discrepancies in the distributional assumptions. However, in practice, the data should be monitored for significant deviations

from distributional assumptions in relationship to their historical behavior, as this can be indicative of performance break-downs. In Section 4.2 we discuss a more nonparametric approach designed to target sensor degradation that may not be adequately captured by the Kalman filtering methods. Additional comments on the methods proposed in Sections 4.1-4.2 may be found in Section 4.3.

#### 4.1. Kalman filtering

Referring to (4.1), assume in this section that  $\epsilon_1, \epsilon_2, \dots \sim N(0, \sigma^2)$  are iid and let  $W(\cdot)$  be an independent standard Brownian motion. Assume further that the errors  $\epsilon_j$  and  $W(\cdot)$  are independent of  $\text{ISIG}(\cdot)$ , and that  $\beta(t) = \beta_0 + v\sigma W(t)$ , where  $\sigma, v > 0$ ,  $\beta_0 \in \mathbb{R}$  are constant. In practice,  $v$  is very small and  $\beta(t)$  resembles a slowly-varying process, despite the fact that it is non-stationary. Under these assumptions, (4.1) can be rewritten as

$$\begin{aligned} \text{FS}_k &= \beta_k \text{ISIG}_k + \epsilon_k, \\ \beta_k &= \beta_{k-1} + \delta_k, \end{aligned} \quad (4.2)$$

where  $\delta_k = \beta_k - \beta_{k-1} \sim N(0, \Delta_k v^2 \sigma^2)$  is independent of  $\beta_{k-1}$  and  $\Delta_k = \tau_k - \tau_{k-1}$ . To estimate  $\text{BG}(t)$ , given  $\mathcal{F}_{t-}$  and  $\text{ISIG}(t)$ , take

$$\widehat{\text{BG}}(t) = \hat{\beta}(t) \text{ISIG}(t),$$

where  $\hat{\beta}(t) = E[\beta(t) | \mathcal{F}_{t-}]$  is the conditional expectation of  $\beta(t)$  given  $\mathcal{F}_{t-}$  (the posterior mean at time  $t$ ). Notice that  $\widehat{\text{BG}}(\tau_k) = \hat{\beta}(\tau_k) \text{ISIG}_k = E[\text{FS}_k | \mathcal{F}_{\tau_k-}, \text{ISIG}_k]$ . In other words, the expected fingerstick value at time  $\tau_k$  (conditional on the observed data and  $\text{ISIG}_k$ ) is used to predict the blood glucose density at time  $t_k$ . Though  $\text{FS}_k$  is a corrupted measurement of blood glucose density (3.2), the fingerstick measurements are, in the present context, the gold standard measurement for blood glucose density. Note that  $\widehat{\text{BG}}(t)$  is the minimum mean squared error predictor for  $\text{FS}(t)$ .

The model (4.2) is a dynamic linear model (West and Harrison (1997)) and Kalman filtering can be used to compute  $E[\beta(t) | \mathcal{F}_{t-}]$ . Indeed, let  $\hat{\beta}_k = \hat{\beta}(\tau_k) = E[\beta_k | \mathcal{F}_{\tau_k-}]$ . Then for  $\tau_k < t \leq \tau_{k+1}$ ,  $\hat{\beta}(t)$  can be efficiently computed by the recurrence

$$\hat{\beta}(t) = \hat{\beta}_{k+1} = \frac{[\Delta_k v^2 + \hat{v}_k^2] \text{FS}_k \text{ISIG}_k + \hat{\beta}_k}{[\Delta_k v^2 + \hat{v}_k^2] \text{ISIG}_k^2 + 1}, \quad (4.3)$$

$$\hat{v}_{k+1}^2 = \frac{\Delta_k v^2 + \hat{v}_k^2}{[\Delta_k v^2 + \hat{v}_k^2] \text{ISIG}_k^2 + 1}, \quad (4.4)$$

where  $\hat{\beta}_0 = \beta_0$  and  $\hat{v}_0 = 0$ . The auxiliary quantity  $\hat{v}_k^2$  used to calculate  $\hat{\beta}(t)$  in (4.3) is the conditional variance,  $\hat{v}_k^2 = \text{Var}(\beta_k | \mathcal{F}_{\tau_k}) / \sigma^2$ , where  $\mathcal{F}_t = \sigma(\{\text{ISIG}(\tilde{\tau}_k);$

$\tilde{\tau}_k \leq t$ ,  $\{\text{FS}(\tau_k); \tau_k \leq t\}$ ). Note that in (4.3)–(4.4) if  $\tau_{k+1} - \tau_k$  is very large, then prior information contained in  $\hat{\beta}_k$  and  $\hat{v}_k$  is weighted less heavily in computing  $\hat{\beta}_{k+1}$  and  $\hat{v}_{k+1}$ . One may estimate  $\beta_0$  and  $v$  using maximum likelihood with historical data, or by using a training dataset and tuning against some criteria, such as the mean absolute relative difference (MARD), defined below at (7.3).

Continuous glucose monitoring algorithms based on the Kalman filter relationship (4.3)–(4.4) perform reasonably well when applied to the Star 1 dataset (details found in Section 8 below). However, further improvement may be had by incorporating information about the age of the blood glucose sensor into the model. In practice, blood glucose sensors must be replaced periodically (approximately every 3 days in the Star 1 study) because of biofouling and other issues. Let  $T_j$  be the time at which the patient begins using the  $j$ -th blood glucose sensor,  $j = 1, 2, \dots$ . Then  $0 = T_0 < T_1 < T_2 < \dots$ . Now define  $C_j(t) = T_j I\{T_j < t\}$  and let

$$\mathcal{G}_{t-} = \sigma(\{\text{ISIG}(\tilde{\tau}_k; \tilde{\tau}_k < t), \{\text{FS}(\tau_k); \tau_k < t\}, \{C_j(t); j = 1, 2, \dots\}).$$

Assume that  $\{T_1, T_2, \dots\} \subseteq \{\tau_1, \tau_2, \dots\}$  (i.e., a fingerstick measurement is taken whenever a new sensor is inserted, as is typically the case in the Star 1 dataset) and that

$$\beta(t) = \sum_{j=0}^{\infty} I\{T_j \leq t < T_{j+1}\} \beta^{(j)}((t - T_j) \vee 0), \quad (4.5)$$

for processes  $\beta^{(j)}(t)$ ,  $j = 0, 1, 2, \dots$ , defined by

$$\begin{aligned} \beta^{(j)}(t) &= \beta_0^{(j)} + v\sigma W^{(j)}(t), \\ \beta_0^{(j)} &= \beta_0(T_j) = \beta_0 + v_0\sigma \tilde{W}(T_j), \end{aligned}$$

where  $\tilde{W}(t)$ ,  $W^{(0)}(t)$ ,  $W^{(1)}(t)$ ,  $W^{(2)}(t), \dots$  are iid standard Brownian motions and  $v, v_0 > 0$  are constants. Then the process  $\beta(t)$  essentially “restarts” whenever a new sensor is inserted. The constants  $v\sigma$  and  $v_0\sigma$  may be viewed as the “within sensor” and “between sensor” volatility, respectively; both are assumed to be constant and, in practice, are small. Similar to the above method that ignores sensor replacement, we use an estimator of the form  $\widehat{\text{BG}}(t) = \hat{\beta}(t)\text{ISIG}(t)$ , where  $\hat{\beta}(t)$  is computed using a simply modified version of (4.3)–(4.4). In particular, if  $\tau_k < t \leq \tau_{k+1}$  and  $\tau_k \neq T_j$ , then (4.3)–(4.4) are used to compute  $\hat{\beta}(t)$ ; if, on the

other hand,  $\tau_k = T_j$  for some  $j$ , then

$$\hat{\beta}(t) = \hat{\beta}_{k+1} = \frac{[\Delta_k v^2 + (\hat{v}_0^{(j)})^2] \text{FS}_k \text{ISIG}_k + \hat{\beta}_0^{(j)}}{[\Delta_k v^2 + (\hat{v}_0^{(j)})^2] \text{ISIG}_k^2 + 1}, \quad (4.6)$$

$$\hat{v}_{k+1}^2 = \frac{\Delta_k v^2 + (\hat{v}_0^{(j)})^2}{[\Delta_k v^2 + (\hat{v}_0^{(j)})^2] \text{ISIG}_k^2 + 1}, \quad (4.7)$$

$$\hat{\beta}_0^{(j+1)} = \frac{[\Delta_j^{(0)} v_0^2 + (\hat{v}_0^{(j)})^2] \text{FS}_k \text{ISIG}_k + \hat{\beta}_0^{(j)}}{[\Delta_j^{(0)} v_0^2 + (\hat{v}_0^{(j)})^2] \text{ISIG}_k^2 + 1}, \quad (4.8)$$

$$(\hat{v}_0^{(j+1)})^2 = \frac{\Delta_j^{(0)} v_0^2 + (\hat{v}_0^{(j)})^2}{[\Delta_j^{(0)} v_0^2 + (\hat{v}_0^{(j)})^2] \text{ISIG}_k^2 + 1}, \quad (4.9)$$

where  $\Delta_j^{(0)} = T_j - T_{j-1}$ . In other words, at times that do not immediately follow the insertion of a new sensor, one estimates  $\text{BG}(t)$  as in the original setting, where sensor replacement is ignored. However, immediately after a new sensor is inserted, estimates of  $\text{BG}(t)$  are based on the evolution of the sensor process  $\beta_0^{(j)}$ ,  $j = 0, 1, 2, \dots$ , as determined by (4.6)–(4.9).

## 4.2. Nonparametric empirical Bayes

By modeling  $\beta(t)$  as (4.5), which leads to the modified Kalman filter (4.6)–(4.9), we incorporate sensor-specific information into our estimation procedures. We have found that this can lead to substantial improvements in the performance of CGM algorithms (results may be found in Section 8). Given that sensors are known to degrade over time, and that their performance changes as this happens, these improvements are not unexpected. On the other hand, (4.5) only incorporates the fact that a new sensor, when inserted, may perform differently from an older sensor. In fact, one would expect that the sensitivity of the sensor decreases gradually over time, due to biofouling and other causes. In other words, one would expect that  $\beta(T_j + t)$  should be larger than  $\beta(T_j)$  for  $t < T_{j+1}$ . This is confirmed by Figure 3, where the ratios  $\text{FS}_k/\text{ISIG}_k$  are plotted for different sensor ages. Note that the ratio tends to increase with sensor age.

We attempted to incorporate information about the behavior of sensors as they age by using a nonparametric alternative to the Kalman filter. Suppose that (4.1) holds, that  $\epsilon_1, \epsilon_2, \dots \sim N(0, \sigma^2)$  are iid, and that  $\beta(t)$  is an unobserved (hidden) Markov process, with  $\beta(t)|\mathcal{L}_t \sim \pi(\cdot|t)$  for some distribution  $\pi(\cdot)$ . Note that in the previous section  $\beta(t)$  was also an unobserved Markov process (a Brownian motion). The blood glucose  $\text{BG}(t)$  is estimated as above, with  $\widehat{\text{BG}}(t) = \hat{\beta}(t)\text{ISIG}(t)$ . However, here we assume that  $\beta(t)$  is a nonstationary discrete

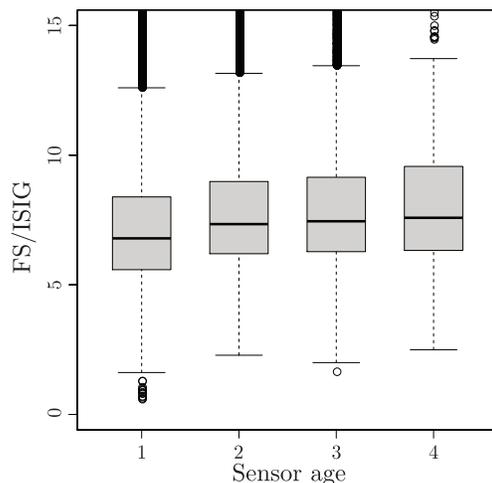


Figure 3. Ratios  $FS_k/ISIG_k$  (mg/dL  $nA^{-1}$ ) by sensor age. 1: Sensor less than 1 day old. 2: Sensor between 1 and 2 days old. 3: Sensor between 2 and 3 days old. 4: Sensor at least 3 days old.

Markov process with a constant transition intensity  $\lambda$  where, at each transition time, a new state for  $\beta(t)$  is generated from a distribution depending on sensor age. More precisely, suppose that  $T_j \leq t < T_{j+1}$ , where  $T_j$  is the time at which the patient begins using the  $j$ -th sensor, and let  $S(t) = t - T_j$  denote the age of the sensor in use at time  $t$ . Additionally assume that

$$0 = S_{0,j} \leq S_{1,j} \leq \cdots \leq S_{N_j,j} = T_{j+1} - T_j$$

are the arrival times of a Poisson process with rate  $\lambda > 0$ . Then

$$\beta(t) = \beta(S_{i,j}), \quad S_{i,j} \leq t < S_{i+1,j}$$

and  $\beta(S_{i,j})$  is generated from a distribution with the probability density function  $g(\cdot|S(t) = S_{i,j})$ , which is determined by the sensor age  $S(t)$ .

For times  $T_j \leq t' < t < T_{j+1}$ , the transition probabilities for the process  $\beta(t)$  are characterized as follows:

$$p_{t',t}(b|b') = e^{-\lambda(t-t')} I\{b = b'\} + \int_{t'}^t g(b|S(r)) de^{-\lambda(t-r)}.$$

The first term here corresponds to the case where there is no transition from time  $t'$  to time  $t$ , while the second term determines the probability of transitioning from  $\beta(t') = b'$  to  $\beta(t) = b \neq b'$  between time  $t'$  and  $t$ . Now let  $\pi(b|t^-)$  be the conditional probability mass function of  $\beta(t)$  given  $\mathcal{G}_{t^-}$ . Then

$$\pi(b|t^-) = e^{-\lambda(t-\tau_k)} \pi(b|\tau_k) + \int_{\tau_k}^t g(b|S(r)) de^{-\lambda(t-r)}$$

for all  $t \in (\tau_k, \tau_{k+1}]$ . Since the data in  $(\tau_k, \tau_{k+1})$  contains no information about the process  $\beta(t)$ , we do not update the density function for  $t \in (\tau_k, \tau_{k+1})$ . At time  $t = \tau_{k+1}$ , the posterior is updated using  $\text{FS}_{k+1}$ . The updating rule is as follows:

$$\pi(b_i|\tau_{k+1}) = \frac{\pi(b_i|\tau_{k+1}^-) \exp(-\{\text{FS}_{k+1} - \text{ISIG}_{k+1}b_i\}^2/(2\sigma^2))}{\sum_j \pi(b_j|\tau_{k+1}^-) \exp(-\{\text{FS}_{k+1} - \text{ISIG}_{k+1}b_j\}^2/(2\sigma^2))}.$$

In this formulation, the distribution  $g(\cdot|s)$  remains unknown. We may use Markov chain Monte Carlo (MCMC) methods to estimate the density functions  $g(\cdot|s)$  for each  $s$ . However, it is computationally costly. For simplicity, we assume that the last transition time is at time  $t$  and thus the approximate density function is

$$\pi(b|t^-) = e^{-\lambda(t-\tau_k)}\pi(b|\tau_k) + (1 - e^{-\lambda(t-\tau_k)})g(b|S(t)),$$

for  $t \in (\tau_k, \tau_{k+1}]$ . The coefficient distribution  $g(\cdot|s(t))$  may be estimated as follows:

$$\hat{g}(\cdot|S(t)) = \frac{1}{|A_\epsilon(t)|} \sum_{\tau_j \in A_\epsilon(t)} \pi(\cdot|\tau_j),$$

where  $A_\epsilon(t) = \{\tau_j; |S(\tau_k) - S(t)| \leq \epsilon \text{ and } \tau_j < t\}$  is the collection of times  $\tau_j < t$  where the sensor age is within  $\epsilon$  of the sensor age at time  $t$ . In practice, we treat the age of sensor as a discrete variable. In the Star 1 dataset, sensors are used for up to 14 days; in our analysis, we take  $\epsilon = 0.5$  and track the averages of posteriors for 14 categories:  $a = 0.5, 1.5, \dots, 13.5$ .

### 4.3. Additional comments

For the methods described in Section 4.1–4.2, estimates of blood glucose density take the form  $\overline{\text{BG}}(t) = \hat{\beta}(t)\text{ISIG}(t)$ , where  $\hat{\beta}(t)$  depends on the data available up to time  $t$ . It is noteworthy that in all of these methods,  $\hat{\beta}(t)$  does not depend on  $\text{ISIG}(\tilde{\tau}_j)$ , for  $\tilde{\tau}_j \notin \{\tau_k\}$ . That is, only ISIG measurements obtained at times  $\tau_k < t$ , where a fingerstick blood glucose measurement is also available are used to compute  $\hat{\beta}(t)$ . (Recall that ISIG is measured at the times  $\{\tilde{\tau}_k\}$ , FS is measured at times  $\{\tau_k\}$ , and that  $\{\tau_k\} \subseteq \{\tilde{\tau}_k\}$ . In the Star 1 dataset the time between ISIG measurements,  $\tilde{\tau}_{k+1} - \tilde{\tau}_k$ , is five minutes, while the time between FS measurements,  $\tau_{k+1} - \tau_k$ , is roughly six hours, on average; thus, ISIG measurements are obtained much more frequently than FS.) Furthermore, in all of the methods considered here,  $\hat{\beta}(t)$  is only updated at times  $\tau_1, \tau_2, \dots$ . In other words,  $\hat{\beta}$  is piecewise-constant, with jumps at  $\tau_1, \tau_2, \dots$ .

It may be reasonable to seek methods for computing  $\hat{\beta}(t)$  that utilize all of the measurements  $\{\text{ISIG}(\tilde{\tau}_k); \tilde{\tau}_k < t\}$  and to update  $\hat{\beta}(t)$  more frequently, using additional information as it becomes available in an attempt to obtain better

estimates of blood glucose density. For example, one could potentially update  $\hat{\beta}(t)$  at each time  $\tilde{\tau}_k$ , in order to reflect the most recent ISIG measurements and sensor age information. However, this is challenging: if ISIG measurements are not paired with FS measurements (or some other direct measurement of blood glucose density), information about the relationship between ISIG and blood glucose density is confounded by changes in the blood glucose density. Indeed, under the statistical models considered in Sections 4.1–4.2,

$$\begin{aligned} E[\beta(t) | \{\text{ISIG}(\tilde{\tau}_k); \tilde{\tau}_k < t\}, \{\text{FS}(\tau_k); \tau_k < t\}] \\ = E[\beta(t) | \{\text{ISIG}(\tau_k); \tau_k < t\}, \{\text{FS}(\tau_k); \tau_k < t\}]. \end{aligned}$$

In our view, more refined statistical and mathematical models for describing the relationship between ISIG and blood glucose density will likely be required in order to successfully develop methods that utilize  $\text{ISIG}(\tilde{\tau}_j)$ ,  $\tilde{\tau}_j \notin \{\tau_k\}$ , for estimating this relationship. This is an area of ongoing research.

### 5. Prediction Methods: $\rho(t) \neq 0$

As demonstrated in Section 8 below, the implementations discussed in Section 4 substantially improve on existing methods when applied to the Star 1 dataset. However, these methods are developed under the assumption that  $\rho(t) \equiv 0$  (i.e., that  $\text{BG}(t)$  does not depend on  $\text{IG}'(t)$ ), even though there is compelling theoretical and empirical evidence that  $\rho(t) \neq 0$  (Steil et al. (2005); Wei et al. (2010)). In this section, we discuss approaches to predicting  $\text{BG}(t)$  that attempt to incorporate this information.

The interstitial glucose density  $\text{IG}(t)$  and its derivative  $\text{IG}'(t)$  are not directly observed by blood glucose biosensors. Rather,  $\text{ISIG}(t)$ , an associated electrical current, is measured. Noise in these measurements, along with the fact that  $\text{ISIG}(t)$  is discretely sampled, make it challenging to estimate  $\text{IG}'(t)$ . Thus, it is unclear how beneficial it is to attempt to incorporate information about  $\text{IG}'(t)$  into methods for predicting  $\text{BG}(t)$ . However, simulation studies and results from the Star 1 dataset indicate that substantial gains may be possible. In fact, our method is to derive estimates  $\widehat{\text{ISIG}}'(t)$  of the derivative  $\text{ISIG}'(t)$ , using  $\{\text{ISIG}(\tilde{\tau}_k); \tilde{\tau}_k \leq t\}$ , and to incorporate this into a model that relates  $\text{FS}_k$  to  $\text{ISIG}_k$  and  $\widehat{\text{ISIG}}'_k = \widehat{\text{ISIG}}'(\tau_k)$ :

$$\text{FS}_k = \beta_k \text{ISIG}_k + \gamma_k \widehat{\text{ISIG}}'_k + \epsilon_k, \quad (5.1)$$

where  $\epsilon_1, \epsilon_2, \dots \sim N(0, \sigma^2)$  are iid,  $\beta(t)$  and  $\gamma(t)$  are slowly varying processes, and  $\beta_k = \beta(\tau_k)$ ,  $\gamma_k = \gamma(\tau_k)$ . Note that we rely on an estimate of  $\text{ISIG}'(t)$ , as opposed to an estimate of  $\text{IG}'(t)$ . This is partially justified by (3.1) and the

assumption that  $\alpha(t)$  is slowly varying, which implies that  $\text{ISIG}'(t) \approx \alpha(t)\text{IG}'(t)$ . Notice that  $\widehat{\text{ISIG}}'_k$  is completely determined by previous values of  $\text{ISIG}(t)$  and does not depend on fingerstick measurements  $\text{FS}_k$ .

The model (5.1) extends (4.1) and suggests a natural way to use information about the derivative  $\text{ISIG}'(t)$  for estimating  $\text{BG}(t)$ . In particular, building on the methodology developed in Section 4, we propose estimators

$$\widehat{\text{BG}}(t) = \hat{\beta}(t)\text{ISIG}(t) + \hat{\gamma}(t)\widehat{\text{ISIG}}'(t), \quad (5.2)$$

where  $\hat{\beta}(t)$  and  $\hat{\gamma}(t)$  are estimates of  $\beta(t)$  and  $\gamma(t)$ , respectively, whose specific form is determined by additional assumptions on the processes  $\beta(t)$  and  $\gamma(t)$ . In fact, the main implementation studied here is an extension of the Kalman filtering approach from Section 4.1. This approach is easily modified to include estimates of the derivative  $\widehat{\text{ISIG}}'_k$ . Assume that (5.1) holds and that  $\beta(t) = \beta_0 + v_\beta\sigma W_\beta(t)$ ,  $\gamma(t) = \gamma_0 + v_\gamma\sigma W_\gamma(t)$ , where  $W_\beta(\cdot)$ ,  $W_\gamma(\cdot)$  are independent Brownian motions. Estimates  $\widehat{\text{BG}}(t) = \hat{\beta}(t)\text{ISIG}(t) + \hat{\gamma}(t)\widehat{\text{ISIG}}'(t)$  are computed using a 2-dimensional Kalman filter. Sensor replacement information may be incorporated as in Section 4.1 by “restarting” the Kalman filter whenever a sensor is replaced. Additionally, nonparametric empirical Bayes methodologies like those discussed in Section 4.2 are easily extended to incorporate  $\widehat{\text{ISIG}}'(t)$ .

### 5.1. Estimating $\text{ISIG}'(t)$

In this section, we discuss specific methods for estimating  $\text{ISIG}'(t)$  using the data  $\{\text{ISIG}(\tilde{\tau}_k); \tilde{\tau}_k \leq t\}$ . Here we use the fact that there are typically many more sensor measurements  $\{\text{ISIG}(\tilde{\tau}_k)\}_{k=0}^\infty$  than fingerstick measurements  $\{\text{FS}(\tau_k)\}_{k=0}^\infty$ . A reasonable initial estimate for  $\text{ISIG}'(t)$  is the first difference

$$\widehat{\text{ISIG}}'_1(t) = \frac{\text{ISIG}(\tilde{\tau}_{\tilde{k}(t)}) - \text{ISIG}(\tilde{\tau}_{\tilde{k}(t)-1})}{\tilde{\tau}_{\tilde{k}(t)} - \tilde{\tau}_{\tilde{k}(t)-1}},$$

where  $\tilde{k}(t) = \max\{k; \tau_k \leq t\}$ . On the other hand, to account for noise in the sensor measurements, it is reasonable to implement smoothing methods that utilize more terms  $\text{ISIG}(\tilde{\tau}_{\tilde{k}(t)-i})$ ,  $i = 1, 2, \dots, j$  for estimating  $\text{IG}'(t)$ . For  $j \geq 1$ , define

$$\overline{\text{ISIG}}_{k,j} = \frac{1}{j+1} \sum_{i=1}^{j+1} \text{ISIG}(\tilde{\tau}_{k-i+1}) \text{ and } \bar{\tau}_{k,j} = \frac{1}{j+1} \sum_{i=1}^{j+1} \tilde{\tau}_{k-i+1}.$$

The local linear regression (Cleveland (1979)) estimator for  $\text{ISIG}'(t)$  is

$$\widehat{\text{ISIG}}'_j(t) = \frac{\sum_{i=1}^{j+1} (\tilde{\tau}_{\tilde{k}(t)-i+1} - \bar{\tau}_{\tilde{k}(t),j}) \left[ \text{ISIG}(\tilde{\tau}_{\tilde{k}(t)} - i + 1) - \overline{\text{ISIG}}_{\tilde{k}(t),j} \right]}{\sum_{i=1}^{j+1} (\tilde{\tau}_{\tilde{k}(t)-i+1} - \bar{\tau}_{\tilde{k}(t),j})^2}. \quad (5.3)$$

The main drawback of  $\widehat{\text{ISIG}}'_j(t)$  is that it may be biased if  $\text{ISIG}'(t)$  is changing rapidly near  $t$ . One of the inherent challenges in successfully implementing  $\widehat{\text{ISIG}}'_j(t)$  is choosing the number of previous observations  $j$ . Other methods to smooth estimates of the derivative  $\text{ISIG}'(t)$  may be used, but these methods typically rely on specification of additional tuning parameters analogous to  $j$ . For example, a weighted local linear regression could be computed where more recent ISIG observations are weighted more heavily. However, specification of the weights further complicates matters and, in our analysis, we have not found that this leads to substantial improvements. Kernel (Gasser, Muller and Mammitzsch (1985)) or spline estimators (Zhou and Wolfe (2000)) for the derivative may also be used, but it is more challenging to implement these methods “on-line,” where the goal is to estimate the derivative  $\text{ISIG}'(t)$  only using data available at times  $s \leq t$ . Moreover, these methods still require specification of tuning parameters (e.g. the kernel and bandwidth, or knot locations for splines). Ultimately, we prefer (5.3) for its simplicity and effectiveness.

The local linear regression estimator  $\widehat{\text{ISIG}}'_j(t)$  with  $j > 1$  offers improvements over the estimator  $\widehat{\text{ISIG}}'_1(t)$ . However, we have found that further improvements may be had by using soft-thresholding. More specifically, define the estimator

$$\widehat{\text{ISIG}}'_{j,\lambda}(t) = \text{sign} \left[ \widehat{\text{ISIG}}'_j(t) \right] \max \left\{ \left| \widehat{\text{ISIG}}'_j(t) \right| - \lambda, 0 \right\},$$

where  $\lambda \geq 0$  is the thresholding level. Soft-thresholding methods have been extensively studied in a variety of statistical settings (Donoho (1995); Tibshirani (1996)). Benefits provided by soft-thresholding typically arise from their variance reduction and parsimony properties. Taken together, (3.4) and (5.2) suggest that, in the present setting, soft-thresholding may be viewed as adaptively identifying situations where  $\text{ISIG}'(t)$  is likely to be nonzero and  $\widehat{\text{ISIG}}'_j(t)$  is likely to be useful for estimating  $\text{BG}(t)$ .

Like some of the smoothing methods discussed above, soft-thresholding requires specification of a tuning parameter,  $\lambda$ . Larger values of  $\lambda$  yield estimators that equal 0 more frequently and have smaller variance, but tend to have larger bias as well. The major benefit that we have observed from soft-thresholding seems to arise in settings where  $\text{ISIG}'(t) \approx 0$ . When  $\text{ISIG}'(t)$  is near zero, estimates of the derivative that do not utilize soft-thresholding tend to be highly variable. Soft-thresholding effectively moderates this, at the cost of some bias in our estimates. This is illustrated in the simulation studies in Section 7. Choosing  $\lambda$  to optimally balance this bias-variance tradeoff is challenging and important. In our analysis of the Star 1 dataset, we choose  $\lambda$  to optimize estimation accuracy (MARD) on a training dataset.

Our use of soft-thresholding shares similarities with existing methods for sequential change-point detection. However, existing methods for sequential change-point detection in dynamic models are typically based on identifying deviations in residuals ( $FS_k - \widehat{BG}_k$ ) or the likelihood ratio (Lai (1995, 1998)). The soft-thresholding procedure implemented here focuses solely on estimates of the derivative  $ISIG'(t)$ . This reflects our underlying assumption that there is a nonzero lag time associated with the diffusion of glucose molecules from the blood into interstitial space where the sensor is located (i.e.,  $\rho(t) > 0$ ), which implies that the relevance of  $ISIG'(t)$  for prediction is determined by whether or not  $ISIG'(t) = 0$ . Moreover, if  $ISIG'(t) = 0$ , then  $\widehat{ISIG}'(t)$  can only contribute noise to estimates of  $BG(t)$  and, in these settings, it is sensible to use estimators  $\widehat{BG}(t)$  that do not depend on  $\widehat{ISIG}'(t)$ . This highlights the importance of being able to identify situations where  $ISIG'(t)$  is relatively large in magnitude and  $\widehat{ISIG}'(t)$  may be useful for estimating  $BG(t)$ . Soft-thresholding addresses this need.

## 6. Inference and Prediction Intervals for $BG(t)$ .

Statistical inference for online estimates of blood glucose density has received little attention in the CGM literature. In this section, we discuss basic methods for constructing prediction intervals for  $BG(t)$  that may be used in conjunction with the estimation methods introduced in Sections 4 and 5. We study the empirical performance of these procedures in our analysis of the Star 1 dataset in Section 8. One of the challenges in constructing reliable prediction intervals is that in our analysis, FS is the gold-standard measurement for blood glucose density, yet it is known that fingerstick measurements for blood glucose density are themselves error-prone (Khan et al. (2006)). For the prediction intervals developed in this section, we simply identify BG and FS; alternative approaches to developing prediction intervals that effectively incorporate the errors inherent in FS measurements are the subject of ongoing research.

### 6.1. Normal-based intervals

Under the normality assumptions of the Kalman filter-based methods described in Section 4.1 and Section 5.1, statistical inference via the posterior distribution of  $\beta(t)$  is straightforward. This provides a means for determining, among other things, prediction intervals for  $BG(t)$ . In the following discussion, we restrict our attention to the one-dimensional Kalman filter-based methods of Section 4.1 (i.e., the methods developed under the assumption that  $\rho(t) \equiv 0$ ). The discussion and the proposed intervals are easily adapted to the

two-dimensional Kalman filter from Section 5.1 that incorporates estimates of the derivatives  $\widehat{\text{ISIG}}'(t)$ .

For  $t > 0$ , suppose that  $\tau_k < t \leq \tau_{k+1}$ . Under the model (4.2), we have

$$\beta(t) | \mathcal{F}_{t-} \sim N \left[ \hat{\beta}_k, \sigma^2 \{ (t - \tau_k)v^2 + \hat{v}_k^2 \} \right],$$

where  $\hat{\beta}_k$  and  $\hat{v}_k^2$  are given in (4.3)–(4.4). Thus, the variance of  $\widehat{\text{BG}}(t)$  given the data available up to time  $t$  is

$$\text{Var} \left\{ \widehat{\text{BG}}(t) \mid \mathcal{F}_{t-}, \text{ISIG}(t) \right\} = \sigma^2 \{ (t - \tau_k)v^2 + \hat{v}_k^2 \} \text{ISIG}(t)^2. \quad (6.1)$$

Equation (6.1) quantifies the variability of our estimates for blood glucose density and, in conjunction with (4.1), naturally suggests a  $100(1 - \alpha)\%$  prediction interval for  $\text{BG}(t)$ :

$$\left( \widehat{\text{BG}}(t) - z_{\alpha/2} \sigma \sqrt{\{ (t - \tau_k)v^2 + \hat{v}_k^2 \} \text{ISIG}(t)^2 + 1}, \right. \\ \left. \widehat{\text{BG}}(t) + z_{\alpha/2} \sigma \sqrt{\{ (t - \tau_k)v^2 + \hat{v}_k^2 \} \text{ISIG}(t)^2 + 1} \right). \quad (6.2)$$

Formula (6.2) is also used to obtain prediction intervals for the sensor replacement model of Section 4.1, with  $\hat{\beta}_k$  and  $\hat{v}_k^2$  given by (4.3)–(4.4) and (4.6)–(4.9), as described at the end of Section 4.1.

Note that in order to construct the interval (6.3),  $\sigma^2$  must be estimated (on the other hand, estimating  $\sigma^2$  is not required to compute the estimate  $\widehat{\text{BG}}(t)$ ). We propose estimating  $\sigma^2$  online using the MLE  $\hat{\sigma}_k^2$ , based on the data available at time  $\tau_k$ . The estimator  $\hat{\sigma}_k^2$  is defined recursively by  $\hat{\sigma}_0^2 = 0$  and

$$\hat{\sigma}_k^2 = \left( 1 - \frac{1}{k} \right) \hat{\sigma}_{k-1}^2 + \frac{1}{k} \left\{ \frac{(\text{FS}_k - \hat{\beta}_k \text{ISIG}_k)^2}{(\Delta_k v^2 + \hat{v}_k^2) \text{ISIG}_k^2 + 1} \right\}, \quad k \geq 1.$$

## 6.2. Residual-based intervals

In principle, prediction intervals for  $\text{BG}(t)$  based on the nonparametric empirical Bayes methods of Section 4.2 may be constructed via the posterior distribution of  $\widehat{\beta}(t)$  and the relationship (4.1). However, the relevant distributions are fairly complex and the computational challenges associated with constructing these prediction intervals are nontrivial. Thus, we propose a simple and flexible alternative method for constructing prediction intervals that may be used in conjunction with the estimation methods of Section 4.2: The residual-based  $100(1 - \alpha)\%$  prediction interval for  $\text{BG}(t)$  is

$$\left( \widehat{\text{BG}}(t) + F_t^{-1} \left( \frac{\alpha}{2} \right), \widehat{\text{BG}}(t) + F_t^{-1} \left( 1 - \frac{\alpha}{2} \right) \right), \quad (6.3)$$

where  $F_t^{-1}(\cdot)$  is the inverse CDF of the empirical distribution of the residuals available at time  $t$ ,  $\{\text{FS}(\tau_k) - \widehat{\text{BG}}(\tau_k); \tau_k < t\}$ . Notice that the prediction intervals (6.3) can be constructed using the residuals from any blood glucose estimation method. Indeed, in our analysis of the Star 1 dataset in Section 8, we study the empirical performance of these residual-based intervals when used in conjunction with the nonparametric empirical Bayes estimation methods of Section 4.2, along with the Kalman filter estimation methods of Section 4.1 and Section 5.

## 7. Simulation Studies

In this section, we discuss the results of a simulation study designed to identify situations where estimating  $\text{ISIG}'(t)$  may be useful for predicting  $\text{BG}(t)$ . The model (3.1)–(3.4) forms the basis of our simulation study. However, several simplifying assumptions were made. We generated continuous deterministic signals  $\text{BG}(t)$  (different representative signals are studied in Sections 7.1–7.2) and then generated

$$\text{IG}(t) = \int_0^\infty \text{BG}(t-u)\rho^{-1}e^{-u/\rho} du, \quad (7.1)$$

$$\text{ISIG}(\tau_k) = \alpha\text{IG}(\tau_k) + \epsilon_k, \quad k = 0, 1, 2, \dots, \quad (7.2)$$

where  $\epsilon_0, \epsilon_1, \epsilon_2, \dots \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ ,  $\tau_0, \tau_1, \tau_2, \dots$  were regularly spaced sampling times,  $\rho, \sigma^2 > 0$  were fixed and known, and  $\alpha \equiv 1$  nA mg/dL. Note that in this model, the relationship between  $\text{ISIG}$  and  $\text{IG}$  is not time-varying, the constant  $\alpha \equiv 1$  nA mg/dL is known, and random noise enters the  $\text{ISIG}(\tau_k)$  observations through iid additive errors  $\epsilon_k$ . There are no fingerstick observations  $\text{FS}(t)$  in this model and only the electrical current measurements  $\text{ISIG}(\tau_0), \text{ISIG}(\tau_1), \dots, \text{ISIG}(\tau_k)$  are used to estimate  $\text{BG}(\tau_k)$ .

Two types of estimators for  $\text{BG}(t)$  were considered in this study,

$$\widehat{\text{BG}}_0(t) = \text{ISIG}(t) \quad \text{and} \quad \widehat{\text{BG}}_{j,\lambda}(t) = \text{ISIG}(t) + \rho\widehat{\text{ISIG}}'_{j,\lambda}(t).$$

Clearly, the estimator  $\widehat{\text{BG}}_0(t)$  does not attempt to utilize information about the derivative  $\text{ISIG}'(t)$  to estimate  $\text{BG}(t)$ , similar to the estimators studied in Section 4. Note that (7.1) implies that  $\widehat{\text{BG}}_0(t)$  is necessarily biased. The estimator  $\widehat{\text{BG}}_{j,\lambda}(t)$  makes use of an estimate  $\widehat{\text{ISIG}}'_{j,\lambda}(t)$ . It may be less biased than  $\widehat{\text{BG}}_0(t)$ , but tends to have higher variance. In all of our simulation studies, we used a training dataset to select  $j$  and  $\lambda$  from  $\widehat{\text{BG}}_{j,\lambda}(t)$  in order to minimize MARD (defined in (7.3)).

The overall performance of each estimator was measured by its mean absolute relative difference,

$$\text{MARD}(\widehat{\text{BG}}) = \frac{1}{\#\{\tau_k\}} \sum_k \left\{ 100 \times \frac{|\widehat{\text{BG}}(\tau_k) - \text{BG}(\tau_k)|}{\text{BG}(\tau_k)} \right\}, \quad (7.3)$$

where the sum on the right-hand side above is taken over all sampling times  $\tau_k$  and  $\#\{\tau_k\}$  is the number of sampling times. MARD is a widely used overall measure of performance for continuous blood glucose monitoring (Kovatchev et al. (2008); Keenan, Cartaya, and Mastrototaro (2010)) (in general, if the exact measurement  $\text{BG}(\tau_k)$  is not available in (7.3), then it may be replaced with a surrogate – we use  $\text{FS}(\tau_k)$  in the Star 1 analysis). In our simulation study, we also computed

$$\begin{aligned} \text{MARD}_A(\widehat{\text{BG}}) &= \frac{1}{\#A} \sum_{k \in A} \left\{ 100 \times \frac{|\widehat{\text{BG}}(\tau_k) - \text{BG}(t_k)|}{\text{BG}(\tau_k)} \right\}, \\ \text{MARD}_Q(\widehat{\text{BG}}) &= \frac{1}{\#Q} \sum_{k \in Q} \left\{ 100 \times \frac{|\widehat{\text{BG}}(\tau_k) - \text{BG}(t_k)|}{\text{BG}(\tau_k)} \right\}, \end{aligned}$$

where  $A = \{k; \text{BG}'(\tau_k) \neq 0\}$  and  $Q = \{k; \text{BG}'(\tau_k) = 0\}$  in order to help determine if the estimated derivative  $\widehat{\text{ISIG}}'_{j,\lambda}(t)$  may be more useful when blood glucose density is in an “active” ( $A$ ) or “quiescent” ( $Q$ ) phase. This may have significant clinical implications.

### 7.1. Simulation study 1: Severe hypoglycemia

First consider a blood glucose path  $\text{BG}(t)$  that begins at 180 and drops to 40 in 15 mins. Suppose that the interstitial/blood glucose lag is  $\rho = 15$  minutes, sampling occurs every 5 minutes (i.e.,  $\tau_k - \tau_{k-1} = 5$ ), and  $\sigma = 5$ . Time lags ( $\rho$ ) of 3-15 minutes are referenced in the literature (Boyne et al. (2003); Kulcu et al. (2003); Steil et al. (2005); Wei et al. (2010)). Figure 4 contains plots of  $\text{BG}(t)$ ,  $\text{IG}(t)$ ,  $\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$ ,  $\widehat{\text{BG}}_{1,0}(t)$ , and  $\widehat{\text{BG}}_{2,1}(t)$  for a single sample path  $\{\text{ISIG}(\tau_k)\}$ . Note that in Figure 4 (a)  $\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$  tracks  $\text{IG}(t)$  quite closely, and that when  $\text{BG}(t)$  begins to drop at time  $t = 60$ , there is a substantial lag in  $\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$ . Indeed, when  $\text{BG}(t)$  reaches 40 at time  $t = 75$ , which indicates severe hypoglycemia, the estimate  $\widehat{\text{BG}}_0(t) = \text{ISIG}(t) = 126$  is well within the normal range. If  $\widehat{\text{BG}}_0(t) = \widehat{\text{ISIG}}(t)$  was used to estimate a patient’s blood glucose this could lead to significant clinical consequences. On the other hand, the estimators that attempt to estimate the derivative  $\text{ISIG}'(t)$  (Figure 4 (b)–(c)) track the drop in blood glucose density beginning at time  $t = 60$  much better; at time  $t = 75$ , when  $\text{BG}(t) = 40$ , these methods yield  $\widehat{\text{BG}}_{1,0}(75) = 53$

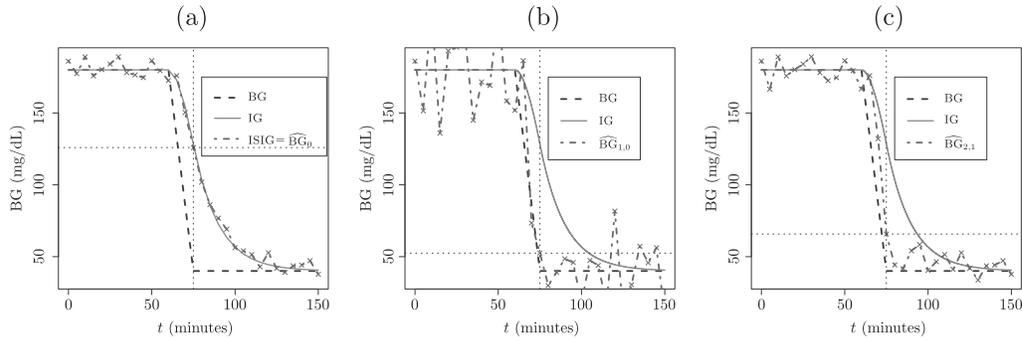


Figure 4. Simulation study 1.  $BG(t)$ ,  $IG(t)$ , and various estimates: (a)  $\widehat{BG}_0(t) = ISIG(t)$ ; (b)  $\widehat{BG}_{1,0}(t)$ ; (c)  $\widehat{BG}_{2,1}(t)$ . Notice that  $BG(75) = 40$ ,  $\widehat{BG}_0(75) = ISIG(75) = 126$ ,  $\widehat{BG}_{1,0}(75) = 53$ , and  $\widehat{BG}_{2,1}(75) = 66$ . The discrepancy between  $\widehat{BG}_0(75) = ISIG(75)$  and  $BG(75)$  could have adverse clinical consequences.

Table 2. Simulation study 1. Mean MARD for various estimation methods (1,000 sample paths).

	MARD (all time points)	$MARD_A$ (times $t$ with $BG'(t) \neq 0$ )	$MARD_Q$ (times $t$ with $BG'(t) = 0$ )
$\widehat{BG}_0(t) = ISIG(t)$	30.88	54.00	29.29
$\widehat{BG}_{1,0}(t)$	32.41	22.66	33.08
$\widehat{BG}_{2,1}(t)$	16.88	39.98	15.28

and  $\widehat{BG}_{2,1}(75) = 66$ . Note that the estimator  $\widehat{BG}_{1,0}(t)$  in Figure 4 (b) appears to be more variable than the estimator  $\widehat{BG}_{2,1}(t)$  in Figure 3 (c), which uses local linear regression and soft-thresholding to estimate the derivative  $ISIG'(t)$ .

We generated 1,000 sample paths as depicted in Figure 4 and computed MARD for the estimators  $\widehat{BG}_0(t)$ ,  $\widehat{BG}_{1,0}(t)$ , and  $\widehat{BG}_{2,1}(t)$ . The results are summarized in Table 2. Note that  $\widehat{BG}_{2,1}(t)$  has the smallest MARD (nearly half that of  $\widehat{BG}_0(t) = ISIG(t)$ ) while  $\widehat{BG}_{1,0}(t)$  has the largest MARD. Both  $\widehat{BG}_{2,1}(t)$  and  $\widehat{BG}_{1,0}(t)$  have smaller  $MARD_A$  than  $\widehat{BG}_0(t) = ISIG(t)$  and, in fact,  $\widehat{BG}_{1,0}(t)$  has the smallest  $MARD_A$ . This reflects the potential usefulness of the derivative  $ISIG'(t)$  in situations where blood glucose density is changing. During more stable periods, the benefits of estimating the derivative  $ISIG'(t)$  may be more questionable. Notice that

$$MARD_Q [\widehat{BG}_{2,1}(t)] < MARD_Q [\widehat{BG}_0(t)] < MARD_Q [\widehat{BG}_{1,0}(t)].$$

The large  $MARD_Q$  for  $\widehat{BG}_{1,0}(t)$  is related to instability in  $\widehat{ISIG}'_1(t)$ . On the

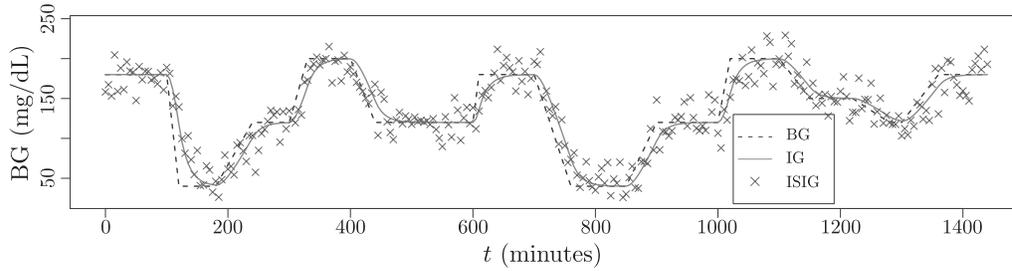


Figure 5. Simulation study 2.  $BG(t)$ ,  $IG(t)$ , and  $ISIG(t)$ ;  $\sigma = 15$ ; twenty-four hour path.

other hand,  $\widehat{BG}_{2,1}(t)$  outperforms  $\widehat{BG}_0(t) = ISIG(t)$  in all situations, whether the blood glucose density is varying ( $MARD_A$ ) or stable ( $MARD_Q$ ). Overall, these results suggest that significant improvements in estimating blood glucose density may be possible during hypoglycemic episodes by utilizing estimates of the derivative  $IG'(t)$ ; however, care must be taken when implementing these estimators.

## 7.2. Simulation study 2: Twenty-four hour monitoring

The blood glucose path depicted in Figure 4 is somewhat specialized and is not representative of the long-term behavior of an individual's blood glucose density that CGM systems are designed to monitor. However, it illustrates the point that, in isolation, certain extreme events can potentially have serious practical implications if the derivative  $ISIG'(t)$  is not included in blood glucose prediction algorithms. In Figure 5, we depict a more complex blood glucose path  $BG(t)$ , over a period of 24 hours (1440 minutes), along with  $IG(t)$  and  $ISIG(t)$ . In this simulation, we assume that interstitial/blood glucose time lag is  $\rho = 15$  minutes and sampling occurs every 5 minutes. We simulated 1,000 sample paths at each of four noise levels,  $\sigma = 20, 15, 10, 5, 0$ , and computed the MARD. Results are summarized in Table 3.

The noise level  $\sigma = 15$  was chosen so that MARD for this simulation was similar to the MARD found in our analysis of the Star 1 dataset; the smaller noise levels are used for illustrative purposes. Notice that  $\widehat{BG}_{j,\lambda}$  outperforms  $\widehat{BG}_0(t) = ISIG(t)$  in all settings, except for  $\sigma = 20$ , where  $\widehat{BG}_0(t) = ISIG(t)$  slightly outperforms  $\widehat{BG}_{j,\lambda}(t)$ . For  $\sigma = 15$ , the noise level with results most closely resembling those from the Star 1 analysis, the estimator  $\widehat{BG}_{j,\lambda}(t)$  only slightly outperforms  $\widehat{BG}_0(t)$ , with an MARD that is 0.38 smaller. Generally speaking, as the noise level  $\sigma$  decreases, the performance of both  $\widehat{BG}_0(t)$  and  $\widehat{BG}_{j,\lambda}(t)$  improves; however, the improvements achieved by  $\widehat{BG}_{j,\lambda}(t)$  are greater than those of  $\widehat{BG}_0(t)$ . This is because of the bias inherent in the estimator

Table 3. Simulation study 2. Mean MARD for various estimation methods and noise levels (1,000 sample paths).

$\sigma$		MARD	MARD <sub>A</sub>	MARD <sub>Q</sub>
20	$\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$	19.03	26.55	17.46
	$\widehat{\text{BG}}_{4,2}$	19.48	26.70	17.97
15	$\widehat{\text{BG}}_0(t) = \alpha\text{ISIG}(t)$	16.10	24.59	14.32
	$\widehat{\text{BG}}_{4,1.5}$	15.72	23.41	14.12
10	$\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$	13.43	23.22	11.38
	$\widehat{\text{BG}}_{4,1}$	11.78	19.64	10.14
5	$\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$	11.15	22.55	8.77
	$\widehat{\text{BG}}_{4,0.5}$	7.58	14.50	6.13
0	$\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$	9.69	22.52	7.00
	$\widehat{\text{BG}}_{1,0}$	1.19	2.25	0.97

$\widehat{\text{BG}}_0(t) = \text{ISIG}(t)$ , which does not account for the derivative  $\text{IG}'(t)$ . The case  $\sigma = 0$  illustrates the extreme case where there is no noise; here  $\widehat{\text{BG}}_0(t) = \text{IG}(t)$  and the error  $\text{MARD}[\widehat{\text{BG}}_0]$  is due entirely to the bias that results from omitting  $\text{ISIG}'(t)$ . On the other hand, when  $\sigma = 0$ ,  $\widehat{\text{BG}}_{1,0}(t)$  provides near perfect reconstruction of  $\text{BG}(t)$ ; the MARD is 0.012 and this could be driven to 0 with a faster sampling rate (so that  $\text{ISIG}'(t)$  could be more accurately estimated).

This simulation study suggests that the effectiveness of estimators  $\widehat{\text{BG}}(t)$  that utilize estimates of  $\text{ISIG}'(t)$  depends largely on the inherent noise level in the data; if the data are less noisy, then these estimators are more likely to perform well. Though the simulation model studied here is greatly simplified, our results suggest that estimating the derivative  $\text{ISIG}'(t)$  may lead to slight improvements in CGM algorithm performance when applied to the Star 1 data (this is borne out in the next section) and when used in other situations with similar noise levels. Our results also suggest that if these noise levels can be reduced – either through an improved biosensor or by other means – then even greater performance gains may be possible.

## 8. Analysis of the Star 1 Dataset

In the Star 1 dataset, 137 subjects using blood glucose sensors were monitored for periods of time spanning 6 days to 948 days (mean 188 days, SD 148). For each subject,  $\text{ISIG}(t)$  was recorded every five minutes with limited exceptions. Fingerstick blood glucose measurements were recorded, on average, every 5.93 hours (mean time between fingersticks). Each individual in the study regularly replaced their blood glucose biosensors, on average every 2.72 days (in the Star1 study, it was recommended that patients replace sensors every 3 days;

more recently available sensors may be used for up to at least 6 days (Keenan et al. (2011)). Sensor replacement times are also contained in the Star 1 dataset, along with estimates of blood glucose density derived from an existing algorithm, denoted  $\text{CGM}(t)$ .

We implemented four methods for estimating  $\text{BG}(t)$  with the Star 1 data.

- (i) The original Kalman filter proposed in Section 4.1 that does not account for sensor replacement.
- (ii) The Kalman filter that “restarts” when a new sensor is inserted (Section 4.1).
- (iii) The nonparametric empirical Bayes method of Section 4.2.
- (iv) The Kalman filtering method that utilizes estimates of  $\widehat{\text{ISIG}}'(t)$  (sensor restarting is not utilized; Section 5).

The mean absolute relative difference MARD was used as the primary metric for comparing the performance of these methods to that of  $\text{CGM}(t)$ . The first 50 patients in the Star 1 dataset were used for training these methods (e.g. identifying  $\beta_0$  and  $v^2$  for the Kalman filtering method, and identifying  $\lambda$  for the nonparametric empirical Bayes method; parameters were chosen to minimize MARD in the training data). The remaining 87 patients were used for validation.

For the Kalman filtering methods described in Section 4.1 (where  $\rho(t) \equiv 0$ ), we took  $\beta_0 = 7$ ,  $v = 0.002$  and  $v_0 = 0.0001$ . These values for  $v$  and  $v_0$  are seemingly very small. However, from (4.3) it is clear that  $\Delta_k v^2 / (\Delta_k v^2 + 1/\text{ISIG}_k^2)$  helps determine the relative significance of recent observations in computing  $\hat{\beta}(t)$ . For the first 50 subjects in the Star 1 dataset, the mean value of  $\text{ISIG}_k^2$  was  $\overline{\text{ISIG}_k^2} = 607$  and the mean time between fingersticks was  $\bar{\Delta}_k = 369$  minutes. Thus, since

$$\frac{\bar{\Delta}_k v^2}{\bar{\Delta}_k v^2 + 1/\overline{\text{ISIG}_k^2}} = \frac{369 \cdot 0.002^2}{607} = 0.53 \quad (8.1)$$

is substantially larger than 0, one can conclude that recent observations have a significant effect on  $\hat{\beta}(t)$ . The relevant quantity for determining the effect of recent observations on  $\hat{\beta}_0^{(j)}$  is

$$\frac{\bar{\Delta}_j^{(0)} v_0^2}{\bar{\Delta}_j^{(0)} v_0^2 + 1/\overline{\text{ISIG}_k^2}} = 0.06,$$

where we have used the fact that the mean sensor life for the first 50 subjects was  $\bar{\Delta}_j^{(0)} = 5110$  minutes. This is significantly smaller than (8.1), suggesting that  $\hat{\beta}_0^{(j)}$  is less variable than  $\hat{\beta}(t)$ . In Figure 6,  $\hat{\beta}(t)$  and  $\hat{\beta}_j^{(0)}$  are plotted for a representative subject in the Star 1 dataset.

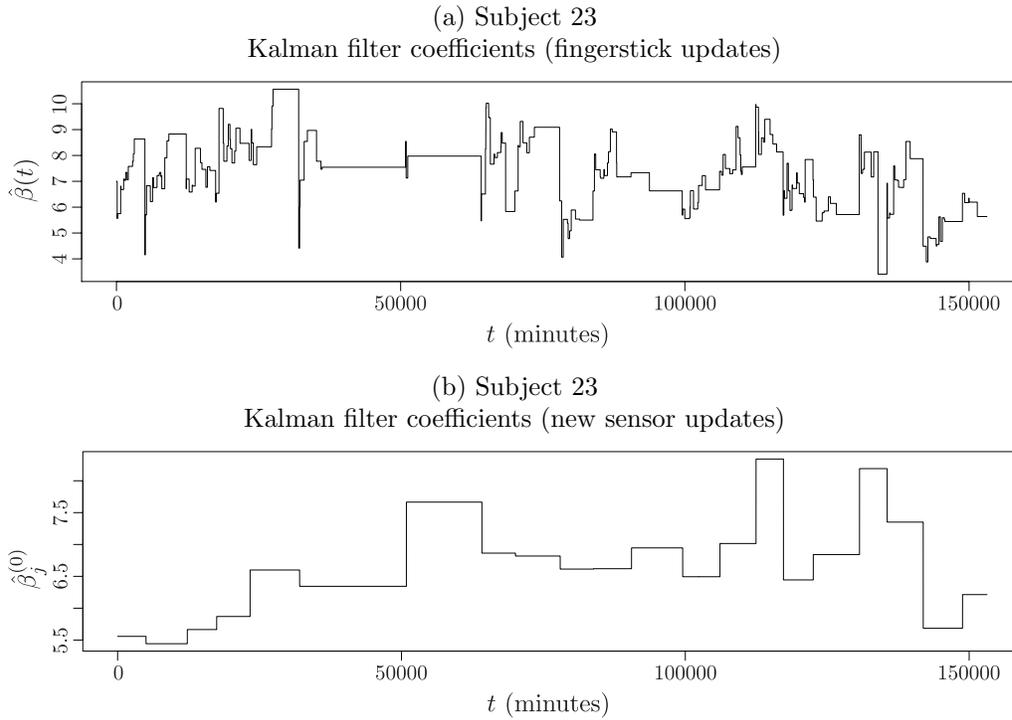


Figure 6. (a)  $\hat{\beta}(t)$  vs. time (in minutes) for Subject 23. (b)  $\hat{\beta}_j^{(0)}$  vs. time (in minutes) for Subject 23. Note that  $\hat{\beta}(t)$  ranges between 3 and 11, while  $\hat{\beta}_0^{(0)}$  ranges between 5 and 9. There are more discontinuities in  $\hat{\beta}(t)$  because it is updated much more frequently than  $\hat{\beta}_0^{(j)}$  (at every fingerstick, as opposed to every time a sensor is replaced). Units for the vertical axes are  $\text{mg/dL nA}^{-1}$ .

For the nonparametric Bayes methods, we took  $\sigma = 20$ ,  $e^\lambda = 0.6$  and let initial distributions  $g(\cdot|a)$  be uniform  $\mathcal{U}[0, 14]$  for each  $a$ . More precisely, the noise level  $\sigma$  for finger stick measurement is first estimated by a ballpark interval  $[10, 25]$  according to (Brunner et al. (1998)), which provides the percentage of measurements within a defined range of the reference values according to different glycemic ranges for various blood glucose meters. The tuning parameter  $e^\lambda$  is a weight for the effect of recent performance and thus between  $(0, 1)$ . We tested various parameters on the training dataset and selected the optimal values in terms of MARD. In practice, it is noticed that the choice of initial distributions does not affect the performance and the results are pretty stable when  $e^\lambda \in [0.5, 0.7]$  and  $\sigma \in [20, 25]$ .

To implement the Kalman filter that utilizes derivative ISIG'(t), we took  $j = 30$  and  $\lambda = 0.01$ , for the derivative estimates  $\widehat{\text{ISIG}}'_{j,\lambda}(t)$ . Notice that  $j = 30$

is much larger than in the simulation studied from the previous section. Since the typical time between ISIG measurements in the Star 1 dataset is 5 minutes, it follows that, on average, ISIG data from the previous 150 minutes is used to estimate  $\widehat{\text{ISIG}}'(t)$  at a given time. We found this to be more effective in the Star 1 dataset than using smaller time windows (i.e., smaller  $j$ ). Other parameter values for our analysis involving  $\widehat{\text{ISIG}}'_{j,\lambda}(t)$  were  $\beta_0 = 7$ ,  $v_\beta = v = 0.002$ ,  $\gamma_0 = 60$ , and  $v_\gamma = 0.001$  (these parameters are defined in Section 5). In Sections 8.1–8.3 we report the results of our analysis.

### 8.1. Estimation accuracy

In Table 4 we report:

- (i) MARD: The average MARD over the 50 subjects in the training data and the 87 subjects in the validation data for the five methods (including CGM) considered here.
- (ii)  $\Delta$ MARD: The difference between the MARD for each method and MARD (CGM).
- (iii) sDARD: The standard deviation of the absolute relative differences

$$100 \times \frac{|\widehat{\text{BG}}(\tau_k) - \text{FS}(\tau_k)|}{\text{BG}(\tau_k)}, \quad (8.2)$$

calculated over all sampling times  $\tau_k$ .

- (iv) MEDARD: The median of (8.2) over all sampling times  $\tau_k$ .
- (v)  $N_{\text{MARD}}$ : The number of individual subjects in the training and validation datasets for which the subject-level MARD of the specified method is smaller than MARD(CGM).

Note that all of the proposed methods substantially outperform CGM. Differences between the four proposed methods are more slight. However, it appears that the Kalman filtering method that utilizes  $\widehat{\text{ISIG}}'(t)$  performs the best, and the original Kalman filter from Section 4.1 (which ignores sensors replacements, and the derivative  $\text{ISIG}'(t)$ ) performs the worst. We point out that though there is a slight degradation in the performance of the four proposed methods when comparing the results for the training data to the results for the validation data, substantial improvements over CGM persist.

### 8.2. Prediction intervals

In Section 5 we proposed methods for constructing prediction intervals for  $\text{BG}(t)$ . We applied these methods to the Star 1 dataset and constructed prediction intervals for  $\text{BG}(t)$  at all times  $\tilde{\tau}_k$  when an ISIG value  $t$  was observed. The

Table 4. Summary statistics for analysis of Star1 dataset.

		MARD	sdARD	MedARD	$\Delta$ MARD	$N_{\text{MARD}}$ (# of subjects in dataset)	
Train.	Kalman	Original	15.51	15.65	11.48	0.99	46 (50)
		Restart w/new sensor	15.37	15.03	11.52	1.13	46 (50)
		$\widehat{\text{ISIG}}'(t)$	15.24	15.38	11.32	1.26	48 (50)
		NP-Bayes	15.39	15.15	11.53	1.11	47 (50)
		CGM	16.50	16.75	12.25		
Valid.	Kalman	Original	15.54	17.01	11.04	0.85	73 (87)
		Restart w/new sensor	15.34	16.44	11.05	1.05	80 (87)
		$\widehat{\text{ISIG}}'(t)$	15.27	16.76	10.85	1.12	80 (87)
		NP-Bayes	15.52	17.05	11.18	0.87	77 (87)
		CGM	16.39	18.31	11.68		

normal-based intervals described in Section 6.1 were constructed in conjunction with the Kalman filter-based estimation methods; the residual-based intervals from Section 6.2 were constructed in conjunction with all estimation methodologies.

Statistics related to the width and coverage of the prediction intervals for the Star 1 dataset are reported in Table 5. Intervals were calculated for nominal coverage levels 80%, 90%, and 95%. The ‘‘Coverage’’ statistics in Table 5 were found by computing the percentage of intervals that contain  $\text{FS}(\tau_k)$ , across all times  $\tau_k$  and all 137 patients in the dataset. The ‘‘Width’’ statistics in Table 5 were found by taking the mean width of intervals at all times  $\tilde{\tau}_k$  across all patients in the dataset.

The results in Table 5 indicate that the normal-based intervals for the Kalman filter methodologies tend to be conservative at the nominal 80% and 90% levels, and are slightly aggressive at the nominal 95% level. On the other hand, the residual-based intervals are somewhat aggressive throughout the range of nominal levels considered here, delivering empirical coverage approximately 2% smaller than the nominal levels in each setting. The mean widths of the intervals constructed for the Star 1 dataset are quite large, ranging from 80.65 mg/dL to 179.67 mg/dL. The prediction intervals for the nonparametric empirical Bayes methods have the smallest mean width of the intervals considered here. Though most of the prediction intervals considered are reasonably accurate (in terms of coverage), they may be too large to serve as the basis for many clinical decisions. On the other hand, this does not negate the improvements in MARD delivered by the proposed methods. Further questions relating to the construction of accurate and useful prediction intervals are the subject of ongoing research.

Table 6 contains quantiles for the empirical distribution of residuals  $\text{FS}(\tau_k) - \widehat{\text{BG}}(\tau_k)$  for all times  $\tau_k$  and all 137 patients in the Star 1 dataset. The residuals

Table 5. Coverage and width of prediction intervals for the Star 1 dataset. “Coverage” is the percentage of intervals that contain  $FS(\tau_k)$ , across all times  $\tau_k$  and all 137 patients in the dataset; “Width” is the mean width of intervals, across all times  $\tilde{\tau}_k$  and all patients in the dataset. Note that the coverage statistics are determined by intervals constructed at times when a FS measurement is available (approximately every six hours), while the width statistics are computed using intervals constructed whenever an ISIG measurement is available (every five minutes).

Estimation methodology	Statistic	Interval type	Nominal coverage		
			80% ( $\alpha = 0.20$ )	90% ( $\alpha = 0.10$ )	95% ( $\alpha = 0.05$ )
Kalman (Original)	Coverage	Normal	86.98%	91.98%	94.49%
		Residual	78.12%	88.04%	93.20%
	Width	Normal	116.30	149.27	177.86
		Residual	83.17	125.11	178.66
Kalman (Restart w/new sensor)	Coverage	Normal	86.95%	91.93%	94.47%
		Residual	78.19%	88.06%	93.20%
	Width	Normal	116.28	149.25	177.84
		Residual	82.65	123.84	176.85
Kalman ( $\widehat{ISIG}'(t)$ )	Coverage	Normal	86.92%	92.02%	94.56%
		Residual	78.17%	88.06%	93.23%
	Width	Normal	112.29	144.12	171.74
		Residual	82.00	124.29	179.67
NP-Bayes	Coverage	Residual	78.08%	88.05%	93.23%
	Width	Residual	80.22	116.70	158.05

Table 6. Quantiles of the empirical distribution of residuals  $FS(\tau_k) - \widehat{BG}(\tau_k)$  (mg/dL) for all time points  $\tau_k$  and all 137 patients in the dataset.

Estimation Methodology	Quantile						
	2.5%	5%	10%	90%	95%	97.5 %	
Kalman (Original)	-95.45	-60.86	-36.94	45.93	68.60	92.90	
Kalman (Restart w/new sensor)	-89.95	-56.51	-34.33	48.30	71.38	95.80	
Kalman ( $\widehat{ISIG}'(t)$ )	-95.99	-61.11	-36.80	44.90	67.10	92.00	
NP-Bayes	-75.64	-51.63	-33.63	47.06	68.79	93.22	

for the Kalman filter-based methods indicate that the residual-based prediction intervals at the 80% and 90% level are somewhat skewed to the right, but the 95% intervals are more symmetric. (Note that all of the normal-based prediction intervals are symmetric by construction.) The residual distribution for the NP-Bayes method is more substantially right-skewed; prediction intervals at the 80%, 90%, and 95% level all tend to be skewed to the right. This is consistent with the fact that while all of the residual-based prediction intervals have similar coverage in the Star 1 dataset, the mean width of the prediction intervals for the

(a) Valid. data: ROC curve for detecting hypoglycemia (b) Valid. data: ROC curve for detecting hyperglycemia

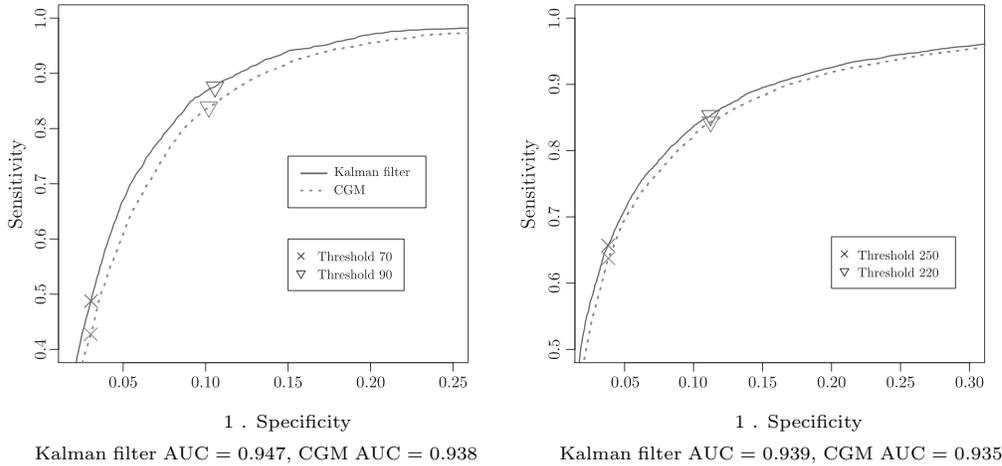


Figure 7. Partial ROC curves (not all thresholds depicted) for detecting hyperglycemia and hypoglycemia using CGM and Kalman filter estimates that utilize  $\widehat{\text{ISIG}}'(t)$ , based on validation data (last 87 subjects from Star1 dataset). Sensitivity and 1 - specificity for specific threshold levels are indicated by the symbols specified in the figure legend. Reported AUC is area under entire ROC curve (all thresholds – so sensitivity and specificity range from 0 to 1).

nonparametric empirical Bayes method is notably smaller than the mean widths of the other residual-based prediction intervals.

### 8.3. Detecting hypoglycemia and hyperglycemia

In addition to achieving good overall accuracy (as measured by MARD, for instance), it is important for CGM algorithms to reliably detect hypoglycemia (low blood glucose density) and hyperglycemia (high blood glucose density). Following (Bode et al. (2004)), any timepoint with  $\text{FS}_k \leq 70$  mg/dL is defined to be a hypoglycemic period, and any timepoint with  $\text{FS}_k \geq 250$  is defined to be a hyperglycemic period. Threshold rules (with thresholds  $s_{\text{hypo}}, s_{\text{hyper}}$ ) are a simple class of rules for detecting hypoglycemia or hyperglycemia based on an estimate  $\widehat{\text{BG}}(\tau_k)$ : If  $\widehat{\text{BG}}(\tau_k) \leq s_{\text{hypo}}$ , declare hypoglycemia; if  $\widehat{\text{BG}}(\tau_k) \geq s_{\text{hyper}}$ , declare hyperglycemia. Using the validation data from the Star 1 dataset, we computed the sensitivity and specificity of these threshold rules for CGM and the Kalman filter estimates for  $\text{BG}(t)$  that utilize  $\widehat{\text{ISIG}}'(t)$ . The results are summarized in the ROC curves in Figure 7. Notice in Figure 7 that the AUC (area under the ROC curve) is larger for the Kalman filtering method than CGM, both for detecting hypoglycemia and detecting hyperglycemia. Moreover, the ROC curve for the

Table 7. Sensitivity, specificity, and false positive rate for hypoglycemia and hyperglycemia detection with different threshold rules.  $s_{hyp}$  denotes threshold for detecting hypoglycemia;  $s_{hyper}$  denotes threshold for detecting hyperglycemia.

		Hypoglycemia		Hyperglycemia	
		$s_{hyp}=70$	$s_{hyp}=90$	$s_{hyper}=250$	$s_{hyper}=220$
Sensitivity	Kalman $\widehat{ISIG}'(t)$	0.4876	0.8752	0.6568	0.8529
	CGM	0.4273	0.8388	0.6374	0.8432
Specificity	Kalman $\widehat{ISIG}'(t)$	0.9695	0.8943	0.9619	0.8877
	CGM	0.9696	0.8981	0.9617	0.8876
False positive rate	Kalman $\widehat{ISIG}'(t)$	0.0294	0.1025	0.0370	0.1101
	CGM	0.0295	0.0989	0.0373	0.1103

Kalman filtering method appears to dominate the ROC curve for CGM across the entire plotted range in Figures 7 (a)–(b) (the Kalman filtering ROC curve lies above and to the left of the CGM ROC curve). Table 7 contains the sensitivity, specificity, and false positive rate for the threshold rules for specific values of  $s_{hyp}$  and  $s_{hyper}$ . Overall, when used in conjunction with a threshold rule for detecting hypoglycemia and hyperglycemia, the Kalman filtering method seems to perform favorably when compared to CGM. These results could potentially be improved by considering alternatives to simple threshold rules for detecting hypoglycemia and hyperglycemia (e.g. by considering rules which account for trends in  $ISIG(t)$  – note that the Kalman filtering method that incorporates  $\widehat{ISIG}'(t)$  already does this implicitly). However, this is not pursued here.

## 9. Discussion

In this paper, we showed that an array of statistical techniques may be effectively brought to bear on an important problem that has received little attention in the statistical literature. However, there are many important open questions about statistical aspects of continuous glucose monitoring. Ongoing research goals include (i) establishing more refined mathematical and statistical models for the relationship between  $ISIG$  and  $BG$ , (ii) developing more accurate and robust methods for statistical inference in continuous glucose monitoring, and (iii) precisely identifying situations where estimates of the derivative  $ISIG'(t)$  may be effectively utilized and identifying optimality properties of estimators for  $ISIG'(t)$ . Another important challenges lies in developing a more thorough understanding of the tradeoffs between parametric (e.g. Kalman filtering) and nonparametric methods for continuous glucose monitoring. Other interesting statistical problems in continuous glucose monitoring include real-time identification

of hyper- and hypoglycemia, and control algorithms for insulin pumps. Statistical tools from control theory, sequential analysis, and nonparametric function estimation are likely to play important roles in addressing critical problems in this important field.

### Acknowledgements

The authors thank an associate editor and two referees for their many helpful comments. This work was supported in part by the Rutgers Faculty Research Grant Program.

### References

- Bequette, B. (2005). A critical assessment of algorithms and challenges in the development of a closed-loop artificial pancreas. *Diabetes Tech. Therapeutics* **7**, 28-47.
- Bliss, M. (2007). *The Discovery of Insulin*. University of Chicago Press; 25th anniversary edition.
- Bode, B., Gross, K., Rikalo, N., Schwartz, S., Wahl, T., Page, C., Gross, T. and Mastrototaro, J. (2004). Alarms based on real-time sensor glucose values alert patients to hypo- and hyperglycemia: the guardian continuous monitoring system. *Diabetes Technology and Therapeutics* **6**, 105-113.
- Boyne, M., Silver, D., Kaplan, J. and Saudek, C. (2003). Timing of changes in interstitial and venous blood glucose measured with a continuous subcutaneous glucose sensor. *Diabetes* **52**, 2790.
- Brunner, G., Ellmerer, M., Sendlhofer, G., Wutte, A., Trajanoski, Z., Schaupp, L., Quehenberger, F., Wach, P., Krejs, G. and Pieber, T. (1998). Validation of home blood glucose meters with respect to clinical and analytical approaches. *Diabetes Care* **21**, 585-590.
- Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829-836.
- Cohen, M., Boyle, E., Delaney, C. and Shaw, J. (2006). A comparison of blood glucose meters in australia. *Diabetes Research and Clinical Practice* **71**, 113-118.
- Donoho, D. (1995). De-noising by soft-thresholding. *Information Theory, IEEE Transactions on* **41**, 613-627.
- Gasser, T., Muller, H., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47**, 238-252.
- Harvey, R., Wang, Y., Grosman, B., Percival, M., Bevier, W., Finan, D., Zisser, H., Seborg, D., Jovanovic, L., Doyle, F., et al. (2010). Quest for the artificial pancreas: combining technology with treatment. *Engineering in Medicine and Biology Magazine, IEEE* **29**, 53-62.
- Heller, A. and Feldman B. (2008). Electrochemical glucose sensors and their applications in diabetes management. *Chemical Rev.* **108**, 2482-2505.
- Hirsch, I., Abelseh, J., Bode, B., Fischer, J., Kaufman, F., Mastrototaro, J., Parkin, C., Wolpert, H. and Buckingham, B. (2008). Sensor-augmented insulin pump therapy: results of the first randomized treat-to-target study. *Diabetes Technology and Therapeutics* **10**, 377-383.
- Hovorka, R. (2006). Continuous glucose monitoring and closed-loop systems. *Diabetic Medicine* **23**, 1-12.

- Kalman, R. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engineering* **82**, 35-45.
- Keenan, D., Cartaya, R. and Mastrototaro, J. (2010). Accuracy of a new real-time continuous glucose monitoring algorithm. *Journal of Diabetes Science and Technology* **4**, 111.
- Keenan, D., Mastrototaro, J., Zisser, H., Cooper, K., Raghavendhar, G., Lee, S., Yusi, J., Bailey, T., Brazg, R. and Shah, R. (2011). Accuracy of the enlite 6-day glucose sensor with guardian and veo calibration algorithms. *Diabetes Technology and Therapeutics* **14**, 225-231.
- Khan, A., Vasquez, Y., Gray, J., Wians Jr, F. and Kroll, M. (2006). The variability of results between point-of-care testing glucose meters and the central laboratory analyzer. *Archives of Pathology and Laboratory Medicine* **130**, 1527.
- Klonoff, D. (2005). Continuous glucose monitoring. *Diabetes Care* **28**, 1231.
- Klonoff, D. (2007). The artificial pancreas: how sweet engineering will solve bitter problems. *J. Diabetes Sci. Tech.* **1**, 72.
- Koschwanetz, H. and Reichert, W. (2007). In vitro, in vivo and post explantation testing of glucose-detecting biosensors: current methods and recommendations. *Biomaterials* **28**, 3687-3703.
- Kovatchev, B., Anderson, S., Heinemann, L. and Clarke, W. (2008). Comparison of the numerical and clinical accuracy of four continuous glucose monitors. *Diabetes Care* **31**, 1160.
- Kowalski, A. (2009). Can we really close the loop and how soon? accelerating the availability of an artificial pancreas: a roadmap to better diabetes outcomes. *Diabetes Technology and Therapeutics* **11**, 113-119.
- Kristensen, G., Mosen, G., Skeie, S. and Sandberg, S. (2008). Standardized evaluation of nine instruments for self-monitoring of blood glucose. *Diabetes Technology and Therapeutics* **10**, 467-477.
- Kulcu, E., Tamada, J., Reach, G., Potts, R. and Lesho, M. (2003). Physiological differences between interstitial glucose and blood glucose measured in human subjects. *Diabetes Care* **26**, 2405.
- Lai, T. (1995). Sequential changepoint detection in quality control and dynamical systems. *J. Roy. Statist. Soc. Ser. B* **57**, 613-658.
- Lai, T. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *Information Theory, IEEE Transactions on* **44**, 2917-2929.
- Mahoney, J. and Ellison J. (2007). Assessing the quality of glucose monitor studies: a critical evaluation of published reports. *Clinical Chemistry* **53**, 1122.
- Reichard, P., Nilsson, B. and Rosenqvist, U. (1993). The effect of long-term intensified insulin treatment on the development of microvascular complications of diabetes mellitus. *New England Journal of Medicine* **329**, 304-309.
- Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer.
- Steil, G., Rebrin, K., Hariri, F., Jinagonda, S., Tادros, S., Darwin, C. and Saad M. (2005). Interstitial fluid glucose dynamics during insulin-induced hypoglycaemia. *Diabetologia* **48**, 1833-1840.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.
- U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (2011). *National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States*, 2011. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

- Wang, J. (2008). Electrochemical glucose biosensors. *Chemical Reviews* **108**, 814-825.
- Wei, C., Lunn, D., Acerini, C., Allen, J., Larsen, A., Wilinska, M., Dunger, D. and Hovorka, R. (2010). Measurement delay associated with the guardian® rt continuous glucose monitoring system. *Diabetic Medicine* **27**, 117-122.
- West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. 2nd edition. Springer Verlag.
- Zhou, S. and Wolfe, D. (2000). On derivative estimation in spline regression. *Statistica Sinica* **10**, 93-108.

Department of Statistics and Biostatistics, Rutgers University, 477 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, U.S.A.

E-mail: ldicker@stat.rutgers.edu

Department of Statistics, 400 Wharton School, University of Pennsylvania, U.S.A.

E-mail: tingni@wharton.upenn.edu

Department of Statistics and Biostatistics, Rutgers University, 504 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, U.S.A.

E-mail: czhang@stat.rutgers.edu

Medtronic MiniMed, 18000 Devonshire Street, Northridge, CA 91325, USA.

E-mail: barry.keenan@medtronic

Department of Statistics, 462 Wharton School, University of Pennsylvania, U.S.A.

E-mail: shepp@wharton.upenn.edu

(Received March 2012; accepted September 2012)