2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Detection of Bias Injection Attacks on the Glucose Sensor in the Artificial Pancreas Under Meal Disturbance

Fatih Emre Tosun[1], André Teixeira[2], Anders Ahlén[1] and Subhrakanti Dey[3]

Abstract— **The artificial pancreas is an emerging concept of closed-loop insulin delivery that aims to tightly regulate the blood glucose levels in patients with type 1 diabetes. This paper considers bias injection attacks on the glucose sensor deployed in an artificial pancreas. Modern glucose sensors transmit measurements through wireless communication that are vulnerable to cyber-attacks, which must be timely detected and mitigated. To this end, we propose a model-based anomaly detection scheme using a Kalman filter and a χ^2 test. One key challenge is to distinguish cyber-attacks from large unknown disturbances arising from meal intake. This challenge is addressed by an online meal estimator, and a novel time-varying detection threshold. More precisely, we show that the ordinary least squares is the optimal unbiased estimator of the meal size under certain modelling assumptions. Moreover, we derive a novel time-varying threshold for the χ^2 detector to avoid false alarms during meal ingestion. The results are validated by means of numerical simulations.**

I. INTRODUCTION

Type 1 diabetes (T1D) is an autoimmune disease which causes elevated blood glucose (BG) levels due to the lack of insulin secretion by the pancreas. The artificial pancreas (AP) is an automated insulin delivery system designed to alleviate the daily burden of self-managing T1D while ensuring tight control of BG levels. An AP consists of three main components: an insulin pump, a continuous glucose monitor (CGM), and a controller. In particular, the pump continuously injects a rapid-acting insulin analog (e.g., lispro) at a rate dictated by the controller. A CGM is a glucose sensor deployed in an AP which provides real-time BG measurements. The main control objective of an AP is to ensure the BG levels remain in a certain glycemic target range, typically 70-180 mg/dL. For minimal invasion, the pump and the CGM are inserted in the subcutaneous (SC) tissue (i.e., under the skin) which introduces input and sensing delays. These lags render the BG control problem even more complicated in the face of unknown disturbances such as meals or exercise.

To date, a plethora of closed-loop control strategies including PID, model predictive, and a fuzzy logic controller has been proposed to tackle the relevant challenges [1]. Most of these controllers assume normal operating conditions. However, anomalies such as erroneous CGM readings result in an incorrect insulin administration as the closed-loop controller employed in the AP relies on accurate sensor data. Inadequate insulin delivery may lead to hyperglycemia (i.e., high blood sugar) which has long-term health implications. In contrast, excessive insulin may cause a lethal hypoglycemic coma. This non-negligible problem has been addressed in various notable works within a fault detection framework [2]. Fault detection algorithms in the AP can be broadly divided into two categories: model-based and data-driven methods. In particular, data-driven methods requires a large set of offline CGM measurements to extract statistical features for decision-making (e.g., [3]). On the other hand, model-based methods rely on a process model representing the glucoregulatory dynamics to detect the anomalies in the input-output behavior. In early work [4], an online fault detection method which employs a state-space model in the innovation form and a Kalman filter, was proposed to improve the safety of the patients at sleep. In a more recent work [5], a hybrid fault detector has been proposed which employs an unscented Kalman filter for model-based detection, and applies principal component analysis for the data-driven part of the algorithm.

It is important to note that the traditional fault detectors are designed against the natural malfunctioning of system components, which are random, and not necessarily malignant. However, a data deception attack may also be the source of erroneous CGM readings. In [6], it was shown that an attacker can reverse engineer the radio protocols to launch passive (e.g., eavesdropping) and active attacks (e.g., replay and data injection attacks) on an insulin delivery system with limited resources such as off-the-shelf hardware, and publicly available information regarding the system components. The paper concludes by proposing a lightweight cryptography algorithm as well as emphasizing the benefits of deploying a wireless body area network where the communication range is limited to the immediate proximity of the patient. Thus, a conceivable way to mitigate cyberattacks could be to utilize in-body communications through the fat tissue which offers a high data rate [7]. While we acknowledge the necessity of deploying such a network to improve security, we suggest that employing a model-based anomaly detector would offer an extra security layer.

In this work, we consider bias injection attacks on the CGM where the adversary adds a constant bias to the sensor readings during meal ingestion to remain stealthier. It is a special class of false data injection attacks (FDIAs) that requires minimal model knowledge [8]. The novelty of this

*This work was supported by the Swedish Research Council under grant 2018-04396, and by the Swedish Foundation for Strategic Research.

[1]F.E. Tosun, A. Ahlén are with the Department of Electrical Engineering, Uppsala University, 751 03 Uppsala, Sweden {fatihemre.tosun, anders.ahlen}@angstrom.uu.se

[2]André Teixeira is with the Department of Information Technology, Uppsala University andre.teixeira@it.uu.se

[3]Subhrakanti Dey is with the Department of Electronic Engineering, Hamilton Institute, National University of Ireland, Maynooth, Ireland. He is also with the Department of Electrical Engineering, Uppsala University, 751 03 Uppsala, Sweden subhra.dey@signal.uu.se

work is two-fold:

1) We derive the optimal unbiased estimator for the meal size that minimizes the mean squared estimation error. The *a posteriori* meal estimate is used to enhance the anomaly detection capability.
2) We propose a time-varying threshold for the χ^2 detector to handle the effect of meal disturbances on the postprandial (i.e., after a meal) BG levels.

The rest of this article is organized as follows. Section II presents the BG dynamics model considered in this work. Section III presents the proposed anomaly detection and meal estimation schemes. Section IV provides a numerical example for the validation of our scheme. Finally, Section V concludes the paper.

II. MEDTRONIC VIRTUAL PATIENT MODEL

The Medtronic Virtual Patient (MVP) model is a control-relevant glucoregulatory model for patients with T1D. This model constitutes the backbone of Medtronic Inc.'s insulin infusion system [9], which has undergone evaluation through clinical trials [10]. The model consists of the following set of ordinary differential equations:

$$\frac{dI_{sc}(t)}{dt} = -\frac{I_{sc}(t)}{\tau_1} + \frac{U_{sc}(t)}{\tau_1\, C_I} \tag{1}$$

$$\frac{dI_p(t)}{dt} = -\frac{I_p(t)}{\tau_2} + \frac{I_{sc}(t)}{\tau_2} \tag{2}$$

$$\frac{dI_e(t)}{dt} = -p_2\, I_e(t) + p_2\, S_I\, I_p(t) \tag{3}$$

$$\frac{dG(t)}{dt} = -(GEZI + I_e(t))\, G(t) + EGP + R_a(t) \tag{4}$$

$$\frac{dG_{sc}(t)}{dt} = -\frac{G_{sc}(t)}{\tau_{sen}} + \frac{G(t)}{\tau_{sen}}. \tag{5}$$

The insulin absorption dynamics are given by (1) and (2) where $U_{sc}(t)$ (m IU/min) is the insulin infusion rate of the pump, $I_{sc}(t)$ and $I_p(t)$ are the insulin levels (m IU/L) in the SC tissue and in plasma, respectively. The interchangeable time constants τ_1 and τ_2 determine the rate of insulin transport from the SC tissue to plasma, and C_I defines the insulin clearance rate.

The insulin-glucose dynamics are given by (3) and (4) where $I_e(t)$ (1/min) is the effect of insulin, and $G(t)$ (mg/dl) is the controlled variable, that is the BG level. The parameter p_2 is the reciprocal of the insulin action time constant, S_I is the insulin sensitivity, $GEZI$ is the glucose effectiveness at zero insulin, and EGP is the endogenous glucose production rate. The term $R_a(t)$ (mg/dl/min) denotes the rate of glucose appearance in plasma following a meal intake. The meal disturbance dynamics are described by the following linear two-compartment model:

$$\begin{aligned} \frac{dD(t)}{dt} &= -\frac{1}{\tau_m} D(t) + C_h(t) \\ \frac{dR_a(t)}{dt} &= -\frac{1}{\tau_m} R_a(t) + \frac{1}{\tau_m^2 V_G} D(t) \end{aligned} \tag{6}$$

where $C_h(t)$ (g/min) is the meal intake, and $D(t)$ (g) is the glucose mass in the input compartment. The parameter τ_m defines the time-to-peak of meal absorption, and V_G is the distribution volume of glucose in plasma. Typically, $C_h(t)$ is modeled as a train of impulses as follows,

$$C_h(t) = \sum_{i \in \mathbb{N}} c_i \delta(t - t_i) \tag{7}$$

where t_i (min) is the time instant of the $i-$th meal intake, c_i (g) is the amount of the carbohydrate (CHO) consumed at t_i, and $\delta(\cdot)$ denotes the Dirac delta function.

Finally, the sensor dynamics are modelled as a first order lag in (5) where the measured variable $G_{sc}(t)$ (mg/dl) is the SC glucose level, and τ_{sen} is the sensor time constant.

III. ANOMALY DETECTION SCHEME

In safety critical systems such as an AP, it is of paramount importance to detect anomalies as early as possible. This section presents the anomaly detection scheme against bias injection attacks on the CGM in the presence of a meal disturbance. In particular, a Kalman filter is employed for residual generation while a χ^2 detector is employed for residual evaluation as shown in Fig.1.

A. Plant Model

To make the analysis tractable, we shall consider a discrete stochastic linear time-invariant (LTI) for the closed-loop AP system based on the MVP model presented in Section II. We consider a discrete model since anomaly detection requires sampling.

1) Insulin-Glucose Dynamics: The MVP model (1-5) can be linearized around the equilibrium point at a given target BG level. The linearized MVP model is then discretized using zero-order hold sampling. We express the state vector of the discrete plant as

$$x(k) \triangleq [\Delta I_{sc}(k)\ \Delta I_p(k)\ \Delta I_e(k)\ \Delta G(k)\ \Delta G_{sc}(k)]^T$$

where $x(k) \in \mathbb{R}^5$ denotes the deviations of the state variables from their equilibrium values at time step k. A more detailed explanation of these steps is provided in Appendix II for better legibility.

We introduce a process noise term $w(k) \in \mathbb{R}^5$ to account for modelling errors. The CGM readings are corrupted by the inherent sensor noise $v(k) \in \mathbb{R}$ as well as possible data injection $y_a(k)$. We assume $\{w(k)\}$ and $\{v(k)\}$ are mutually independent zero-mean white Gaussian processes with covariances $Q \succeq 0$ and $R > 0$, respectively.

The adversary is assumed to be capable of altering the sensor readings at any sampling instant while the insulin pump is immune to FDIAs. Under these assumptions, the plant model can be written as

$$\begin{aligned} x(k+1) &= Ax(k) + B_u u(k) + B_d R_a(k) + w(k) \\ y(k) &= Cx(k) + v(k) + y_a(k). \end{aligned} \tag{8}$$

Here, the control input $u(k) \in \mathbb{R}$ defines the insulin infusion rate relative to the basal insulin rate, and $y(k) \in \mathbb{R}$ denotes the CGM measurements. The dynamics of the discrete meal disturbance $R_a(k)$ are explained in the following section.

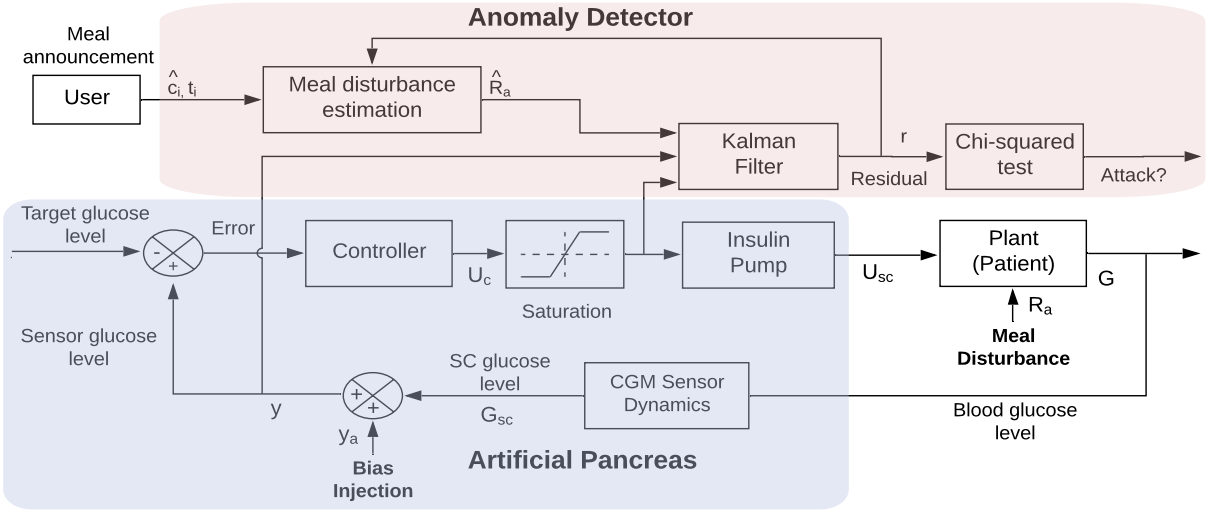


Fig. 1: A block diagram representation of the artificial pancreas along with the proposed anomaly detection scheme. Target glucose level is a set-point. The signs in the feedback loop is reversed to have positive control gains since the insulin must be dispensed when the sensor glucose levels exceed the reference value. The controller output is denoted by U_c which must be saturated below zero as the current pumps cannot remove insulin from the body. The Kalman filter generates the residual signal using the control input U_{sc}, the measured output G_{sc}, and the meal disturbance estimate $\hat{R}_a$. A χ^2 test is applied to decide whether an attack has occured.

2) Meal Disturbance Dynamics: The meal subsystem (6) is discretized by the impulse invariant mapping. Since C_h is assumed to be impulsive, the meal disturbance dynamics may be expressed as an autonomous LTI system with instant changes in the state at meal times as follows,

$$\begin{aligned} x_m(k+1) &= A_m x_m(k) \\ R_a(k) &= C_m x_m(k) \end{aligned} \tag{9}$$

where $x_m(k) = [D(k)\ R_a(k)]^T$. The system (9) starts from zero initial condition, and at meal times, the state is modified as $x_m(t_i) = [c_i\ 0]^T + A_m x_m(t_i - 1)$.

B. Anomaly Detection

An anomaly detector aims to detect any unexpected variation in the plant variables by evaluating the residual signal(s). This section presents the proposed anomaly detection scheme which comprises a Kalman filter, a χ^2 detector with a novel time-varying threshold, and a meal disturbance estimator.

1) Kalman Filter: The state vector $x(k)$ is estimated with a Kalman filter which is assumed to have reached a steady-state; thus, it is a fixed gain estimator. The state estimate at k is defined as $\hat{x}(k) \triangleq \mathbb{E}[x(k)|\{y(0),\ y(1),\ ...,\ y(k-1)\}]$ where $\mathbb{E}[\cdot]$ denotes the expectation operator. Consequently, the state estimate $\hat{x}(k)$ evolve as follows,

$$\hat{x}(k+1) = \bar{A}\hat{x}(k) + B_u u(k) + B_d \hat{R}_a(k) + Ky(k) \tag{10}$$

$$K = APC^T(CPC^T + R)^{-1} \tag{11}$$

with $\bar{A} \triangleq A - KC$. The steady-state filter gain K is given by (11) where P is the steady-state estimation error covariance. The state estimate $\hat{x}(k)$ is given by (10) which includes the known control input $u(k)$ as well as the unknown meal disturbance estimate $\hat{R}_a(k)$ as explained in Section III-C.

We define the state estimation error $\tilde{x}(k)$, the meal disturbance estimation error $\tilde{R}_a(k)$, the output prediction $\hat{y}(k)$, and the residual $r(k)$ as

$$\begin{aligned} \tilde{x}(k) &\triangleq x(k) - \hat{x}(k), & \tilde{R}_a(k) &\triangleq R_a(k) - \hat{R}_a(k) \\ \hat{y}(k) &\triangleq C\hat{x}(k), & r(k) &\triangleq y(k) - \hat{y}(k). \end{aligned}$$

Thus, the state estimation error dynamics evolve as

$$\begin{aligned} \tilde{x}(k+1) &= \bar{A}\tilde{x}(k) + B_d\tilde{R}_a(k) + w(k) - K(v(k) + y_a(k)) \\ r(k) &= C\tilde{x}(k) + v(k) + y_a(k). \end{aligned} \tag{12}$$

The estimation error can be decomposed into 3 parts by invoking the superposition principle as $\tilde{x}(k) = \tilde{x}_s(k) + \tilde{x}_d(k) + \tilde{x}_a(k)$. Here, $\tilde{x}_s(k)$ defines the contribution of the stochastic inputs to the estimation error. Thus, the steady-state error covariance can be defined as $P = \mathbb{E}[\tilde{x}_s(k)\tilde{x}_s(k)^T]$. The state evolution of $\tilde{x}_s(k)$ is described by the following dynamics:

$$\begin{aligned} \tilde{x}_s(k+1) &= \bar{A}\tilde{x}_s(k) + w(k) - Kv(k) \\ r_s(k) &= C\tilde{x}_s(k) + v(k). \end{aligned} \tag{13}$$

Similarly, $\tilde{x}_d(k)$ defines the contribution to $\tilde{x}(k)$ due to meal disturbances, which evolves as

$$\begin{aligned} \tilde{x}_d(k+1) &= \bar{A}\tilde{x}_d(k) + B_d\tilde{R}_a(k) \\ r_d(k) &= C\tilde{x}_d(k). \end{aligned} \tag{14}$$

The last component $\tilde{x}_a(k)$ is the contribution to $\tilde{x}(k)$ due to data injection, which evolves as

$$\begin{aligned} \tilde{x}_a(k+1) &= \bar{A}\tilde{x}_a(k) - Ky_a(k) \\ r_a(k) &= C\tilde{x}_a(k) + y_a(k). \end{aligned} \tag{15}$$

Please note that (15) is valid for generic FDIAs as well as additive sensor faults. However, this work considers only bias injection attacks.

2) χ^2 Detector: In the absence of meals and anomalies, it holds that $r(k) = r_s(k)$. In filtering theory, it is established that $r_s(k)$ is zero-mean white Gaussian with covariance $\sigma_s \triangleq \mathbb{E}[r_s(k)^2] = CPC^T + R$ [11]. Clearly, the distribution of $r(k)$ changes in the presence of meal disturbances or bias injection. Hence, a suitable hypothesis test is required to decide if the residual signal is affected by these factors. To this end, we define a null hypothesis H_0 and an alternative hypothesis H_1 as follows,

H_0 : System is operating normally (i.e., $y_a(k) \equiv 0$).

H_1 : System is under attack (i.e., $y_a(k) \not\equiv 0$).

A chi-squared test is widely deployed in linear systems with Gaussian random inputs for such hypothesis testing problems [12]. In particular, the χ^2 test statistic is generated by taking the squared norm of the residual which is then normalized by the steady-state (co)variance σ_s. The test statistic is compared with a suitably large threshold as

$$g(k) = r^T(k)\sigma_s^{-1}r(k) = \sigma_s^{-1}r(k)^2 \underset{H_0}{\overset{H_1}{\gtrless}} \tau. \tag{16}$$

The random variable $g(k)$ follows a χ^2 distribution; hence the name of the test. Please note that the threshold τ is occasionally violated even in the absence of anomalies. The false alarm rate of the test is defined as

$$\alpha \triangleq \mathbb{P}(g(k) > \tau | H_0). \tag{17}$$

The higher values of τ lower the false alarm rate α, but at the expense of higher missed detection which is the fundamental trade-off in choosing τ. In practice, α is chosen as the design variable (typically, 5%), and the corresponding τ is computed either numerically or from a χ^2 table.

In [13], it was proven that for any generic stochastic LTI plant, the alarm rate of a χ^2 detector strictly increases with the magnitude of the constant bias. More precisely, for any two arbitrary vectors $r_1 \in \mathbb{R}^m$ and $r_2 \in \mathbb{R}^m$ that satisfy $\|r_1\|_2 < \|r_2\|_2$, the following statement holds:

$$\mathbb{P}(\|\sigma_s^{-\frac{1}{2}}(r_s(k) + r_1)\|_2^2 > \tau) < \mathbb{P}(\|\sigma_s^{-\frac{1}{2}}(r_s(k) + r_2)\|_2^2 > \tau) \tag{18}$$

where $\|\cdot\|_2$ denotes the 2-norm of a vector.

It is important to note that in case of a *strictly stealthy* FDIA, an attacker with full model knowledge may compute a non-zero attack sequence $y_a(k)$ that does not increase the alarm probability [14]. A trivial example is when $y_a(k) \sim \mathcal{N}(0, \sigma_s)$ which clearly does not alter the distribution of $r(k)$, and thus $g(k)$. Intuitively, the notion of stealthiness restricts the attack space, and thus its impact, but this is beyond the scope of this work. The main challenge we shall address here is to avoid false alarms during meal ingestion while preserving the detection capability. To this end, we propose a meal disturbance estimation algorithm in the next section.

C. Meal Disturbance Estimation

A meal intake is parameterized by the time t_i and the size c_i of the consumed CHO as can be seen from (7). Meal announcement is an established practice in AP systems. In particular, a preemptive insulin bolus proportional to the CHO size must be administered 15-30 minutes before the meal for an optimal control of postprandial BG levels [15]. Thus, an *a priori* estimate of $R_a(k)$ can be obtained from the meal announcement as follows. For convenience, the user is assumed to announce each meal time accurately. However, multiple studies have shown that the manual CHO counting is error prone (e.g., [16]). Hence, we assume that c_i may differ from the user estimate $\hat{c}_i$. We also assume that the time interval between each meal intake is sufficiently large such that $\hat{x}_m(k)$ is practically zero before each meal, that is $\hat{x}_m(t_i - 1) = [0\ 0]^T$. Consequently, we compute the *a priori* meal disturbance estimate $\hat{R}_a^-(k)$ by solving (19) for $\hat{x}_m(t_i) = [\hat{c}_i\ 0]^T$

$$\begin{aligned} \hat{x}_m(k+1) &= A_m\hat{x}_m(k) \\ \hat{R}_a(k) &= C_m\hat{x}_m(k). \end{aligned} \tag{19}$$

Let $\tilde{c}_i \triangleq c_i - \hat{c}_i$ be the meal size estimation error. When $|\tilde{c}_i|$ is too big, the sequence $|r_d(k)|$ assumes large values until the ingestion is complete. Hence, to avoid misclassifying a meal intake as an anomaly, the initial user estimate $\hat{R}_a^-(k)$ must somehow be corrected. Clearly, estimating $R_a(k)$ amounts to estimating c_i which can be formulated in a linear regression fashion as shall be explained in the sequel. First, define an augmented state vector for the meal estimation error as

$$\tilde{x}_m \triangleq x_m - \hat{x}_m, \quad \tilde{x}_{aug}(k) \triangleq [\tilde{x}_d(k)\ \tilde{x}_m(k)]^T. \tag{20}$$

From (14) and (19), the dynamics of this augmented state can be derived as

$$\begin{aligned} \tilde{x}_{aug}(k+1) &= \tilde{A}_{aug}\ \tilde{x}_{aug}(k) \\ r_d(k) &= \tilde{C}_{aug}\ \tilde{x}_{aug}(k) \end{aligned} \tag{21}$$

with

$$\tilde{A}_{aug} \triangleq \left(\begin{array}{c|c} \bar{A} & B_dC_m \\ \hline \mathbf{0}_{2\times5} & A_m \end{array}\right), \quad \tilde{C}_{aug} \triangleq (C\ \mathbf{0}_{1\times2}).$$

When there is no attack, the residual can be split into a zero-mean stochastic and a deterministic part as $r(k) = r_s(k) + r_d(k)$. In particular, $r_d(k)$ is practically zero before a meal due to the assumption of long periods between the intakes. Furthermore, the postprandial values of $r_d(k)$ are determined by (21). Hence, we collect N samples of $r(k)$ from t_i onward to assess the changes in the residual dynamics. Since (21) is an autonomous LTI system, the evolution of the output solely depends on the initial condition, and obeys the following relationship

$$\bar{r}_d \triangleq [r_d(t_i)\ r_d(t_i+1)\ \ldots\ r_d(t_i+N-1)]^T = \mathcal{O}\tilde{x}_{aug}(t_i) \tag{22}$$

where

$$\mathcal{O} \triangleq \begin{bmatrix} \tilde{C}_{aug} \\ \tilde{C}_{aug}\tilde{A}_{aug} \\ \vdots \\ \tilde{C}_{aug}\tilde{A}_{aug}^{N-1} \end{bmatrix}. \tag{23}$$

There is a linear relationship between the vector $\bar{r}_d$ and the parameter $\tilde{c}_i$ as can be seen from (22). Thus, one can

estimate $\tilde{c}_i$ (not to be confused with c_i) using an ordinary least squares (OLS) regression as follows,

$$\begin{aligned}\hat{\tilde{c}}_i \triangleq \text{argmin } & \frac{1}{N}\sum_{j=t_i}^{t_i+N-1}(r(j)-r_d(j))^2 \\ \text{s.t. } & \bar{r}_d = \mathcal{O}\tilde{x}_{aug}(t_i) \\ & \tilde{x}_{aug}(t_i) = [0\ 0\ 0\ 0\ 0\ \tilde{c}_i\ 0]^T\end{aligned} \tag{24}$$

where $\hat{\tilde{c}}_i$ is the estimate of $\tilde{c}_i$. We suggest to collect minimum seven samples for the static optimization problem (24) as $\tilde{x}_{aug} \in \mathbb{R}^7$. Even though (24) is expressed as a constrained optimization problem for better understandability, it can easily be converted into an equivalent unconstrained optimization problem by substituting the constraint variables in the objective. Since $r_s(k)$ is a zero-mean white homoscedastic sequence, the OLS estimator (24) is the minimum-variance unbiased estimator by the virtue of the Gauss-Markov theorem [17].

Once $\hat{\tilde{c}}_i$ is determined, the *a posteriori* estimate $\hat{R}_a^+(k)$ is computed by solving (19) for the improved estimate $\hat{x}_m(t_i) = [\hat{c}_i + \hat{\tilde{c}}_i\ 0]^T$. After the estimation is complete, the Kalman filter must be reset at the next sampling instant $t_i + N$ to make $r_d(k) \approx 0$. This can be done for example by re-running the filter (10) with $\hat{R}_a^+(k)$ from $k = t_i$. Thus, the meal disturbance estimate can be expressed as follows,

$$\hat{R}_a(k) = \begin{cases}\hat{R}_a^-(k) & \text{if } k \in [t_i,\ t_i+N-1] \\ \hat{R}_a^+(k) & \text{otherwise}\end{cases} \tag{25}$$

Although this method is simple and practical, it takes considerable wait time to collect enough data as the sampling period of a CGM is typically 5 minutes or more. Therefore, during the estimation process, the detection threshold τ must be updated to avoid false alarms as shall be explained in the next section.

D. Time varying threshold

This section derives a time-varying threshold to avoid false detection during meal ingestion. Assuming no attack, the random variable $g(k)$ reads as

$$g(k) = \sigma_s^{-1}[r_s(k)^2 + 2r_s(k)r_d(k) + r_d(k)^2]. \tag{26}$$

The following inequality trivially follows from (26)

$$g(k) \leq \sigma_s^{-1}[r_s(k)^2 + 2|r_s(k)||r_d(k)| + r_d(k)^2]. \tag{27}$$

As noted earlier, $r_d(k)$ is deterministic whereas $r_s(k)$ is stochastic whose best prediction is zero at all times due to the white noise assumption. However, a probabilistic bound on $r_s(k)$ may be obtained from (16) and (17) as follows,

$$\gamma \triangleq \sqrt{\tau_s \sigma_s} \tag{28}$$

$$\alpha = \mathbb{P}(|r_s(k)| > \gamma) \tag{29}$$

where τ_s defines the steady-state threshold that ensures a false alarm rate of α when there is no meal disturbance. The residual statistics change significantly during ingestion unless an accurate meal announcement is made which may not always be possible. To this end, we derive a time-varying detection threshold $\tau(k)$ to account for the effect of the meal during the estimation process, that is $t_i \leq k \leq t_i + N - 1$. We propose the following theorem to derive $\tau(k)$.

Theorem 3.1: Consider the statistical test in (16) with $\tau = \tau_d(k)$ where

$$\tau_d(k) \triangleq \sigma_s^{-1}[\gamma^2 + 2\gamma|r_d(k)| + r_d(k)^2]. \tag{30}$$

The false alarm rate of this test is equal to

$$\mathbb{P}(g(k) > \tau_d(k)) = \frac{\alpha + \mathbb{P}(|r_s(k)| > \gamma + 2|r_d(k)|)}{2} \leq \alpha. \tag{31}$$

Proof: The proof is reported in the Appendix to improve legibility. ∎

Remark 1: The time-varying $\tau_d(k)$ reduces to the constant threshold τ_s in the steady-state. The false alarm rate of the test (31) is strictly less than α during ingestion.

Remark 2: Please note that $\tau_d(k)$ is not an implementable threshold as it is impossible to know the sequence of $r_d(k)$ *a priori*. However, it is reasonable to assume that the initial meal size estimate is off by at most a certain amount $\tilde{c}_{max}$, that is $|\tilde{c}_i| < \tilde{c}_{max}$. Let $r_d^{max}(k)$ denote the output response of (21) for $\tilde{x}_{aug}(t_i) = [0\ 0\ 0\ 0\ 0\ \tilde{c}_{max}\ 0]^T$. Under this assumption, it follows that $r_d^{max}(k) \geq |r_d(k)|$.

Corollary 1.1: Suppose that an implementable time-varying threshold $\tau_{max}(k)$ is defined as

$$\tau_{max}(k) \triangleq \sigma_s^{-1}[\gamma^2 + 2\gamma r_d^{max}(k) + (r_d^{max}(k))^2]. \tag{32}$$

Then, the following probability statement holds:

$$\mathbb{P}(g(k) > \tau_{max}(k)) \leq \mathbb{P}(g(k) > \tau_d(k)) \leq \alpha. \tag{33}$$

Proof: Since $\tau_{max}(k) \geq \tau_d(k)$, the claim of the corollary directly follows from (31). ∎

The sensitivity of the test (16) with $\tau = \tau_{max}(k)$ depends on the magnitude of $\tilde{c}_{max}$. If it is too large, $\tau_{max}(k)$ might be overly conservative in the sense that the detection rate becomes drastically low. However, thanks to the meal estimator presented in Section III-C, this conservative threshold is only necessitated during data collection. Hence, we suggest a piecewise detection threshold as follows,

$$\tau(k) = \begin{cases}\tau_{max}(k) & \text{if } k \in [t_i,\ t_i+N-1] \\ \tau_s & \text{otherwise}\end{cases} \tag{34}$$

After the *a posteriori* meal estimate is obtained, we set $\tau(k)$ back to τ_s without having to wait for the ingestion to be completed.

All the analysis presented in this section presumed no attack scenario which seems to defeat the main purpose of this work. Clearly, this anomaly detection scheme is sensitive to the injection attacks that occur outside the time interval $[t_i,\ t_i + N - 1]$. The threshold in (34) can also detect large biases if the attack occurs within this interval. However, a sufficiently small bias can bypass the detector. Nevertheless, the attack should be detected after resetting the filter with the attacked meal disturbance estimate since $\tau(k)$ is also set back to τ_s. The detection time will depend on the size and the rate of the injected bias.

IV. Numerical Simulation

This section presents a validation of the proposed anomaly detection scheme by a numerical simulation. Table I reports the numerical values of the MVP model parameters identified for a particular virtual subject in [18], which will be used in the simulations. The process noise covariance Q is assumed to be a diagonal matrix with entries $10^{-2},\ 10^{-3}, 10^{-8}, 10^{-2},$ and 10^{-1}, respectively. The variance of the sensor noise R is taken as unity.

We employ a discrete causal PID controller with the following pulse transfer function:

$$K_p + \frac{(K_p/T_i)T_s z}{z-1} + \frac{K_p T_d N_f (z-1)}{(1+N_f T_s)z - 1}. \tag{35}$$

The target glucose level is set to 100 mg/dl. The control gains are chosen as follows: $K_p = 0.2$ (mU/min)/(mg/dl), $T_i = 450$ (min), and $T_d = 60$ (min). The filter coefficient N_f is chosen as 0.01. The numerical values of the system matrices in (8) and (9) are obtained from Table I for the sampling period $T_s = 5$ minutes.

We have performed two closed-loop simulations with and without a bias injection for the validation of our results. The simulations started from the steady-state, and ran for an eight hour long interval. We assumed a single meal intake of 75 grams occurring at exactly one hour after the commence of the simulation. The maximum allowable mismatch between the true and the user estimate of the meal size (i.e., $\tilde{c}_{max}$) was chosen to be 20 grams.

The malicious goal of the attacker was to drive the patient into hypoglycemia by adding a bias of 40 (mg/dl) on the CGM. Since this is a very large bias, in order to remain undetected, the attacker was assumed to slowly inject the bias by utilizing the step response of a low-pass filter rather than adding the target bias at once. The filter used for this purpose was a critically damped second order system which is described by the following pulse transfer function:

$$1.21\ 10^{-3}\frac{z+0.9672}{(z-0.9512)^2} \tag{36}$$

To evaluate the worst case scenario, the start of the bias injection was selected to coincide with the time of the meal intake. Fig. 2b visualizes this attack sequence.

TABLE I: The MVP model parameters with their numerical values identified for a certain virtual subject.

Parameter	**Value**	**Value**
C_I	2.01	[L/min]
τ_1	49	[min]
τ_2	47	[min]
p_2	$1.06\ 10^{-2}$	[min^{-1}]
S_I	$8.11\ 10^{-4}$	[L/mU/min]
GEZI	$2.2\ 10^{-3}$	[min^{-1}]
EGP	1.33	[mg/dl/min]
V_G	253	[dl]
τ_m	50	[min]
τ_{sen}	10	[min]

Fig. 2a shows the trajectories of the BG and the measured glucose levels. The sensor glucose readings are lagging behind the actual BG levels as noted in Section I. The positive bias injected by the attacker causes the closed-loop controller to command the pump to deliver more insulin than necessary. This is nicely shown in Fig. 2a where the plasma glucose levels of the patient becomes dangerously low while the sensor glucose readings remain around the target value.

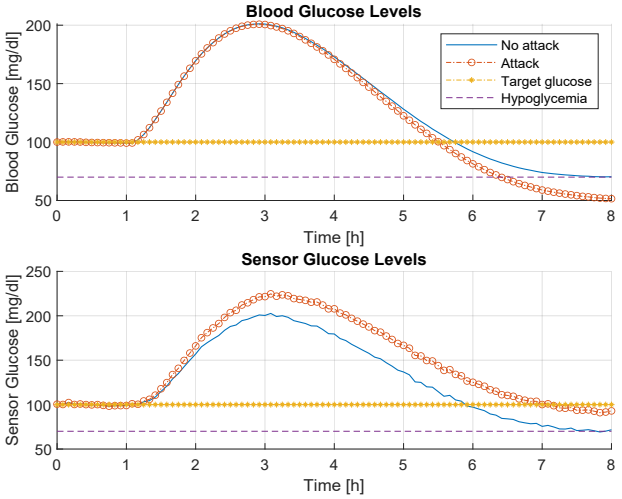
Fig. 2c reports the results of the online meal disturbance estimation algorithm described in Section III-C. We have used the CVX toolbox in Matlab [19] to obtain the *a posteriori* meal size estimate which is given by the convex optimization problem (25). Seven samples of the residual were collected for meal estimation which amounts to a wait time of half an hour. The user estimate of the meal size was 60 g. When there was no attack, the *a posteriori* estimate turned out to be 72 g, which is fairly close to the true value. However, when there was an attack, the algorithm overestimated the meal size as 79 g.

Finally, Fig. 2d demonstrates the performance of the proposed anomaly detection algorithm. In particular, we provide the plots of the χ^2 test results for the attack and no attack scenarios. The false alarm rate α was set to 5 % which stipulates that the steady-state threshold $\tau_s = 3.841$. As can be seen in Fig. 2d, a few false alarms were triggered in the no attack scenario. Moreover, the number of false alarms has agreed with the theoretical rate α even for a relatively small sample size. This indicates that (25) provides a reasonably accurate estimate of the meal size in the absence of attacks.

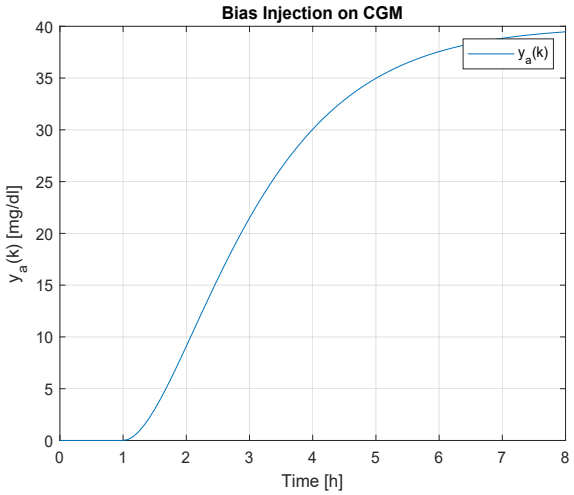
As can be seen in Fig. 2d, the attack was detected soon after the *a posteriori* meal estimation was computed. It is interesting to note that the conservative threshold $\tau_{max}(k)$ is also able to detect persistent attacks. However, the detection time is significantly delayed in this case. Furthermore, the attacker can even bypass the detector by ceasing the data injection towards the end of the meal ingestion. However, this is not possible for $\tau(k)$ where the threshold is set back to the steady-state value after obtaining $\hat{R}_a^+$.

V. Conclusion

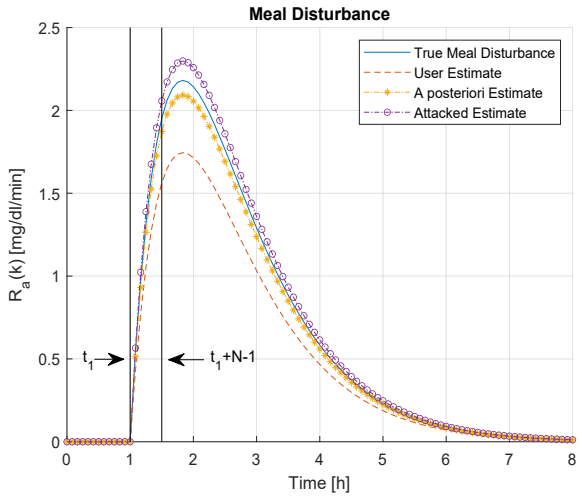
In this work, we have proposed a novel anomaly detection algorithm against bias injection attacks on the glucose sensor deployed in an AP in the presence of meal disturbances. Our algorithm utilizes a χ^2 detector with a time-varying threshold whose false alarm rate has been shown to be bounded by the steady-state alarm rate. The detection capability is further enhanced through an optimal *a posteriori* meal estimator. The efficacy of the proposed algorithm has been demonstrated through a numerical simulation. However, a more in depth investigation for various attack scenarios is needed prior to clinical trials. Future work will include the analysis of more sophisticated attack strategies such as stealthy FDIAs. We will also investigate the benefits of employing a windowed χ^2 detector that uses current and past measurements of the residual.



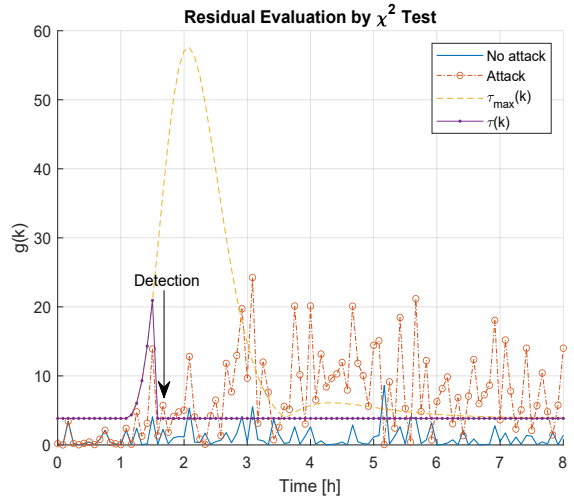
(a) The blood glucose and the sensor glucose levels



(b) Low-pass filtered bias injection



(c) Meal disturbance estimates



(d) The χ^2 detector with a time-varying threshold

Fig. 2: Simulation results of the proposed meal estimation, and the anomaly detection algorithm.

APPENDIX I
PROOF OF THEOREM 3.1

This section presents a proof of Theorem 3.1 with the aid of the following lemmas:

Lemma 1.1: The following statement is valid during a single meal absorption.

$$\mathbb{P}(sgn(r_s(k)r_d(k)) = \pm 1) = 1/2$$

where $sgn(\cdot)$ denotes the sign function.

Proof: Since $r_s(k)$ is a zero-mean Gaussian variable, and $r_d(k)$ does not alter sign during a single meal absorption, it holds that $\mathbb{P}(r_s(k)r_d(k) < 0) = \mathbb{P}(r_s(k)r_d(k) > 0) = 1/2$ which is equivalent to the statement of the lemma. ■

Please note that the claim of the lemma is not guaranteed for multiple meals because $r_d(k)$ may alter sign if, for example, the first meal size is overestimated while the second is underestimated. However, we need not consider this case since the initial user estimate of the meal size is corrected soon after the intake.

Lemma 1.2: The following statements hold true

$$r_d(k) = 0 \iff \mathbb{P}(sgn(r_s(k)r_d(k)) = 0) = 1$$
$$r_d(k) \neq 0 \iff \mathbb{P}(sgn(r_s(k)r_d(k)) = 0) = 0.$$

Proof: The statements trivially follow from the fact that $sgn(x) = 0 \iff x = 0$. ■

Next, we present the proof of Theorem 3.1:

Proof: From (26) and (30), we obtain

$$g(k) > \tau_d(k) \Longleftrightarrow r_s^2(k) - \gamma^2 + 2(r_s(k)r_d(k) - \gamma|r_d(k)|) > 0.$$

Using the identity $x = |x|sgn(x)$, and reordering terms, we can rewrite this inequality as follows,

$$r_s(k)^2 - \gamma^2 + 2|r_d(k)|(|r_s(k)|sgn(r_s(k)r_d(k)) - \gamma) > 0.$$

For convenience of notation, we drop the time argument in $r_s(k)$ and $r_d(k)$. Now that the range of the $sgn(\cdot)$ function

is $\{-1, 0, 1\}$, we define two variables η^+ and η^- as

$$\eta^+ \triangleq r_s^2 - \gamma^2 + 2|r_d|(|r_s| - \gamma) = (|r_s| - \gamma)(|r_s| + \gamma + 2|r_d|)$$
$$\eta^- \triangleq r_s^2 - \gamma^2 - 2|r_d|(|r_s| + \gamma) = (|r_s| + \gamma)(|r_s| - \gamma - 2|r_d|).$$

Next, we partition the probability $\mathbb{P}(\chi^2(k) > \tau_d^2(k))$ as

$$\mathbb{P}(\eta^+ > 0 | sgn(r_s r_d) = 1)\ \mathbb{P}(sgn(r_s r_d) = 1) +$$
$$\mathbb{P}(r_s^2 > \gamma^2 | sgn(r_s r_d) = 0)\ \mathbb{P}(sgn(r_s r_d) = 0) +$$
$$\mathbb{P}(\eta^- > 0 | sgn(r_s r_d) = -1)\ \mathbb{P}(sgn(r_s r_d) = -1).$$

Since $\eta^+ > 0 \Longleftrightarrow |r_s| > \gamma$ and $\eta^- > 0 \Longleftrightarrow |r_s| > \gamma + 2|r_d|$ the proof follows from Lemmas 1.1, 1.2, and (29). ■

APPENDIX II
LINEARIZATION OF THE MVP MODEL

The MVP model (1-5) has five states as

$$x(t) = [I_{sc}(t)\ I_p(t)\ I_e(t)\ G(t)\ G_{sc}(t)]^T$$

The equilibrium point of this nonlinear model in the absence of meal disturbances (i.e., $R_a(t) = 0$) is obtained by solving $\dot{x}(t) = \mathbf{0}$. The equilibrium values of the state variables and the control input are denoted by the asterisk superscript. The equilibrium point (x^*, u^*) is easily computed as

$$x^* = [\frac{I_e^*}{S_I}\ \frac{I_e^*}{S_I}\ I_e^*\ G^*\ G_{sc}^*]^T, \quad u^* = \frac{C_I I_e^*}{S_I} \qquad (37)$$

with $I_e^* = EGP/G^* - GEZI$. Then, a continuous linear model for the insulin-glucose dynamics is obtained from the Jacobian linearization of equations (1-5) around $(x^*\ u^*)$. The state vector of the linear system defines the deviation of the original state from the equilibrium as $\Delta x(t) = x(t) - x^*$. Similarly, for the control input the following relationship holds $u(t) = U_{sc}(t) - u^*$. The target glucose G^* is chosen as 100 mg/dl in the numerical example presented in Section IV. With the selection of the target glucose level, the rest of the numerical values of the linearized system directly follows from (37) and Table I. The simulations are then performed on the discretized plant model as explained in Section III-A.

REFERENCES

[1] A. Haidar, "The artificial pancreas: How closed-loop control is revolutionizing diabetes," *IEEE Control Systems Magazine*, vol. 36, no. 5, pp. 28–47, 2016.

[2] K. Kölle, A. L. Fougner, K. A. Frelsøy Unstad, and Øyvind Stavdahl, "Fault detection in glucose control: Is it time to move beyond cgm data?" *IFAC-PapersOnLine*, vol. 51, no. 27, pp. 180–185, 2018, 10th IFAC Symposium on Biological and Medical Systems BMS 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896318333627

[3] X. Yu, M. Rashid, J. Feng, N. Hobbs, I. Hajizadeh, S. Samadi, M. Sevil, C. Lazaro, Z. Maloney, and A. Cinar, "Fault detection in continuous glucose monitoring sensors for artificial pancreas systems," *IFAC-PapersOnLine*, vol. 51, no. 18, pp. 714 – 719, 2018, 10th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896318319566

[4] A. Facchinetti, S. Del Favero, G. Sparacino, and C. Cobelli, "Detecting failures of the glucose sensor-insulin pump system: Improved overnight safety monitoring for type-1 diabetes," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 4947–4950.

[5] K. Turksoy, A. Roy, and A. Cinar, "Real-time model-based fault detection of continuous glucose sensor measurements," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1437–1445, 2017.

[6] Chunxiao Li, A. Raghunathan, and N. K. Jha, "Hijacking an insulin pump: Security attacks and defenses for a diabetes therapy system," in *2011 IEEE 13th International Conference on e-Health Networking, Applications and Services*, 2011, pp. 150–156.

[7] N. B. Asan, E. Hassan, J. Velander, S. R. Mohd Shah, D. Noreland, T. J. Blokhuis, E. Wadbro, M. Berggren, T. Voigt, and R. Augustine, "Characterization of the fat channel for intra-body communication at r-band frequencies," *Sensors*, vol. 18, no. 9, p. 2752, 2018.

[8] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st International Conference on High Confidence Networked Systems*. New York, NY, USA: Association for Computing Machinery, 2012, p. 55–64. [Online]. Available: https://doi.org/10.1145/2185505.2185515

[9] S. S. Kanderian and G. M. Steil, "Apparatus and method for controlling insulin infusion with state variable feedback," U.S. Patent 8 777 924B2, Oct, 2010. [Online]. Available: https://patents.google.com/patent/US8777924B2/en

[10] S. S. Kanderian, S. A. Weinzimer, and G. M. Steil, "The identifiable virtual patient model: comparison of simulation and clinical closed-loop study results," *Journal of diabetes science and technology*, vol. 6, no. 2, pp. 371–379, 2012.

[11] B. D. Anderson and J. B. Moore, *Optimal filtering*. Mineola, N.Y: Dover Publications, 2005.

[12] R. Mehra and J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, vol. 7, no. 5, pp. 637–640, 1971. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005109871900288

[13] J. Miloševič, T. Tanaka, H. Sandberg, and K. H. Johansson, "Analysis and mitigation of bias injection attacks against a kalman filter," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 8393–8398, 2017, 20th IFAC World Congress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896317321560

[14] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs," *Automatica*, vol. 82, pp. 251–260, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005109817302418

[15] F. J. Doyle, L. M. Huyett, J. B. Lee, H. C. Zisser, and E. Dassau, "Closed-loop artificial pancreas systems: engineering the algorithms," *Diabetes care*, vol. 37, no. 5, pp. 1191–1197, 2014.

[16] A. Brazeau, H. Mircescu, K. Desjardins, C. Leroux, I. Strychar, J. Ekoé, and R. Rabasa-Lhoret, "Carbohydrate counting accuracy and blood glucose variability in adults with type 1 diabetes," *Diabetes research and clinical practice*, vol. 99, no. 1, pp. 19–23, 2013.

[17] Y. Dodge, *The concise encyclopedia of statistics*. Springer Science & Business Media, 2008.

[18] S. S. Kanderian, S. Weinzimer, G. Voskanyan, and G. M. Steil, "Identification of intraday metabolic profiles during closed-loop glucose control in individuals with type 1 diabetes," 2009.

[19] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.