Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 56-2 (2023) 6518–6524

Model Identification with Incomplete Input Data in Type 1 Diabetes ⋆

Basak Ozaslan * **Eleonora M. Aiello** * **Francis J. Doyle III** * **Eyal Dassau** *

* *Harvard John A. Paulson School Of Engineering And Applied Sciences, Harvard University, Cambridge MA (e-mail: dassau@seas.harvard.edu).*

Abstract: A major challenge in fitting models to glucose metabolism in people with type 1 diabetes is incomplete data as its collection partially relies on self-reporting and does not include all relevant events. We develop a method for identifying optimal input corrections to reestablish a correct input-output relationship in the data while jointly identifying personalized model parameters. The unreported or misreported parts in the data are reconciled by adding sparse corrections via mixed-integer quadratic programming leading to an improved identification of the model parameters. We conduct numerical experiments with incomplete in-silico training data and show that models obtained from our method are able to provide more accurate predictions on test data than models obtained from standard methods. The performance of our methodology is similar to that attained with the standard method when trained on data with complete information.

Keywords: Identification and validation, Quantification of physiological parameters, Grey box modeling, Physiological Model, Type 1 Diabetes

1. INTRODUCTION

Type 1 Diabetes (T1D) is associated with the lack of endogenous insulin production, leading to high blood glucose levels (American Diabetes Association, 2021). Short of a cure, strategic planning of insulin injection and carbohydrate intake are essential to keep blood glucose levels within a safe range. Yet, glucose management is a challenging task in daily life as (i) glucose metabolism is, in general, difficult to predict, and (ii) the insulin effect happens on a timescale that makes timely glucose regulation non-trivial. Therefore, mathematical models that capture the relationship between the main inputs (i.e., insulin and meal) and glucose outcomes are highly sought to guide immediate and long-term treatment decisions for achieving satisfactory glucose management.

With the increasingly available data recorded through Continuous Glucose Monitors (CGM) and insulin pumps, the literature has focused on data-driven methods to obtain personalized models (Toffanin et al., 2017; Colmegna et al., 2020). It has been shown that machine learning methods, such as training neural networks on T1D data may help increase the model performance (Kushner et al., 2020a) but they may also lead to inadvertent outcomes on test data (Narasimhamurthy et al., 2019; Kushner et al., 2020b). Thus, models that are based on physiological knowledge and are parameterized to allow inter-individual differences, while not as flexible as neural networks, are considered safer. Therefore, the identification problem comprises fitting few model parameters to data. However, free-living data are noisy and in many cases incomplete. In particular, as people with T1D self-report their carbohydrate intake, the time and amount of these meals can be wrongly reported or not reported at all (e.g., the person may simply choose not to or forget to report). Such inaccuracies may lead to inaccurate models that result in inappropriate treatment decisions.

The main goal of this work is to identify personalized metabolic parameters in a way that the identification becomes inherently robust against significant inaccuracies. Patek et al. (2016) introduced the concept of continuous correction signal to reconcile the difference between model-predicted and measured outputs. The work in Hughes et al. (2021) extended this method to personalizing the model parameters before adding a continuous correction signal.

Noting that meal intake and insulin injections represent the system inputs, events like unreported or misreported meals and boluses, can be interpreted as input disturbances. Moreover, noting that such events are typically characterized by an impulse nature, we develop a method that automatically recognizes the occurrence of such disturbances and allows for a sparse correction signal in the identification. This effort allows reestablishing a correct input-output relationship in the data leading to improved identification of the model parameters. The reason for choosing sparse corrections is to mainly target the above disturbances, which are usually few and far between in a day. However, as demonstrated in this work, such corrections can also handle other input inaccuracies, which are not strictly sparse by nature, as in the case of insulin pump occlusion (when the insulin pump records injected insulin that in fact is not delivered). Sparsity can be enforced

⋆ Research reported in this article was supported by NIH grant R01DK120358

10.1016/j.ifacol.2023.10.299

in the identification problem by a suitable regularization term in the cost function (James et al., 2013) or with a cardinality constraint in a mixed-integer formulation of the identification problem (Bertsekas, 1997).

The contribution of this work is a sequential identification algorithm that alternates between nonlinear least squares to optimize model parameters and Mixed-Integer Quadratic Programming (MIQP) to optimize the time and amount of input corrections (Section 4). We provide case studies on the performance of the proposed method for ten in-silico subjects under five real-life use-case scenarios (Section 5) using the UVA/Padova T1D Metabolic Simulator (Dalla Man et al., 2014) accepted by the United States Food and Drug Administration (FDA).

Mixed integer optimization has been employed in model identification of hybrid models (Roll et al., 2004), fault detection (Marseglia and Raimondo, 2017), and best subset selection in regression analysis (Bertsimas et al., 2020). To the best of our knowledge, no prior work optimized sparse corrections for incomplete input data while jointly identifying the model parameters of a dynamical system. We argue that model identification in T1D is amenable to such an approach due to the particular nature of its data.

2. BACKGROUND

In this section, we provide the necessary background on T1D treatment and insulin-meal-glucose models.

2.1 Daily Management of T1D

The main treatment is external insulin injections, which can be administered manually or using automated insulin delivery systems - depending on the availability and personal choice (Holt et al., 2021). The goal is to keep blood glucose levels within a desired range. The two main elements of any T1D treatment regimen are (i) monitoring blood glucose levels, and (ii) timing and sizing the amount of insulin injections. In order to guide the decisions on insulin dosing based on data from glucose measurements and patient inputs (such as meal reports), models are sought to predict glucose outcomes.

2.2 Subcutaneous Oral Glucose Minimal Model (SOGMM)

SOGMM is a minimal physiological model for glucose dynamics that relates glucose outputs with oral glucose intake and subcutaneous insulin injections. Its set of differential equations is given by (Patek et al. (2016)):

$$\dot{G}(t) = -(S_g + X_I(t)) \cdot G(t) + S_g \cdot G_b + \frac{R_a(t)}{BW \cdot V_g}, \quad \text{(1a)}$$

$$\dot{X}_I(t) = -p_2 \cdot X_I(t) + p_2 \cdot S_I\big(\frac{I_p(t)}{V_I \cdot BW} - I_b\big), \quad \text{(1b)}$$

$$\dot{Q}_1(t) = -k_\tau \cdot Q_1(t) + u_m(t), \quad \text{(1c)}$$

$$\dot{Q}_2(t) = -k_{\text{abs}} \cdot Q_2(t) + k_\tau \cdot Q_1(t), \quad \text{(1d)}$$

$$\dot{I}_{sc1}(t) = -(k_d) \cdot I_{sc1}(t) + u_i(t), \quad \text{(1e)}$$

$$\dot{I}_{sc2}(t) = -k_d \cdot I_{sc2}(t) + k_d \cdot I_{sc1}(t), \quad \text{(1f)}$$

$$\dot{I}_p(t) = -k_{cl} \cdot I_p(t) + k_d \cdot I_{sc2}(t), \quad \text{(1g)}$$

$$R_a(t) = f_c \cdot k_{abs} \cdot Q_2(t), \quad \text{(1h)}$$

where G (mg/dL) is the plasma glucose concentration, G_b is the basal plasma glucose concentration, X_I (l/min) is the remote insulin compartment that corresponds to the portion of insulin that acts on glucose, Q_1 (mg) and Q_2 (mg) are the two compartments representing the oral glucose transport, I_{sc1} and I_{sc2} (mU) are the interstitial insulin transport compartments, I_p (mU) and I (mU/l) represent the insulin mass in plasma, and R_a (mg/min· kg) represents the glucose rate of appearance. $BW(kg)$ is the person's bodyweight $V_g(kg/dL)$ and $V_I(L/kg)$ represent the distribution volume of glucose and insulin, respectively. p_2, k_d, k_{cl} are transport rates that act on the insulin-related compartments (i.e., I_{sc1}, I_{sc2}, I_p, X_I) while k_τ and k_{abs} are the transport rates that act on meal related compartments (i.e. Q_1, Q_2). Finally, S_g (1/min) is the fractional glucose effectiveness, S_I (1/min per mU/L) is the insulin sensitivity, and f_c is the fractional meal absorption. This model has two inputs: u_m (mg/min) and u_i (mU/min), which are the carbohydrate size in an ingested meal and insulin infusion, respectively. The parameters in (1) are generally assumed to be time-invariant as in Patek et al. (2016), and Hughes et al. (2021).

3. PROBLEM FORMULATION AND APPROACH

3.1 SOGMM Parameters

We adopt SOGMM from Section 2.2. In the original work, Patek et al. (2016) individualized the model through G_{b}, I_{b}, and S_{I} that they obtained a priori based on the person's average glucose over three months, prescribed basal insulin profile, and total daily insulin use, respectively. Garcia-Tirado et al. (2018) showed that the majority of the parameters in SOGMM are structurally identifiable. In this work, we categorize the parameters in (1) into $\theta_p = (G_b, I_b, S_I, k_\tau, k_{abs}, k_d, k_{cl})$ as the set of parameters we aim to identify per individual (see Table 1), and $\theta_f = (S_g, V_g, V_I, p_2, f_c)$ as fixed parameters that are set to their nominal values provided in Patek et al. (2016).

Table 1. Parameters identified per subject (θ_p)

Symbol	Explanation	Nominal value
G_b	Glucose at equilibrium	Average glucose (mg/dL)
I_b	Insulin at equilibrium	Average profile (mU/L)
S_{I}	Insulin Sensitivity	TDI dependent
k_τ	Meal transport rate	0.0018
k_{abs}	Meal absorption rate	0.0182
k_{d}	Insulin utilization rate	0.0100
k_{cl}	Insulin clearance rate	0.1600

Due to discrete-time measurements, we use the system in its discrete-time form by defining k as the discrete time-index and the metabolic state vector as $x(k) = (G(k), X_I(k), Q_1(k), Q_2(k), I_{sc1}(k), I_{sc2}(k), I_p(k))^T$.

3.2 Forward Simulations

By fixing θ_p, we can run numerical simulations of the SOGMM model. First, we discretize time via forward Euler integration and obtain the following:

$$\begin{aligned} x(k+1) &= f(x(k), \theta) + B_i u_i(k) + B_m u_m(k), \\ \hat{y}(k) &= Cx(k), \end{aligned} \quad (2)$$

where $u_i(k)$ and $u_m(k)$ are the insulin and meal inputs, respectively, and $\hat{y}(k) = G(k)$ is the estimated output. We have $B_i = (0, 0, 0, 0, 0, \Delta t, 0)^T$, $B_m = (0, 0, \Delta t, 0, 0, 0, 0)^T$,

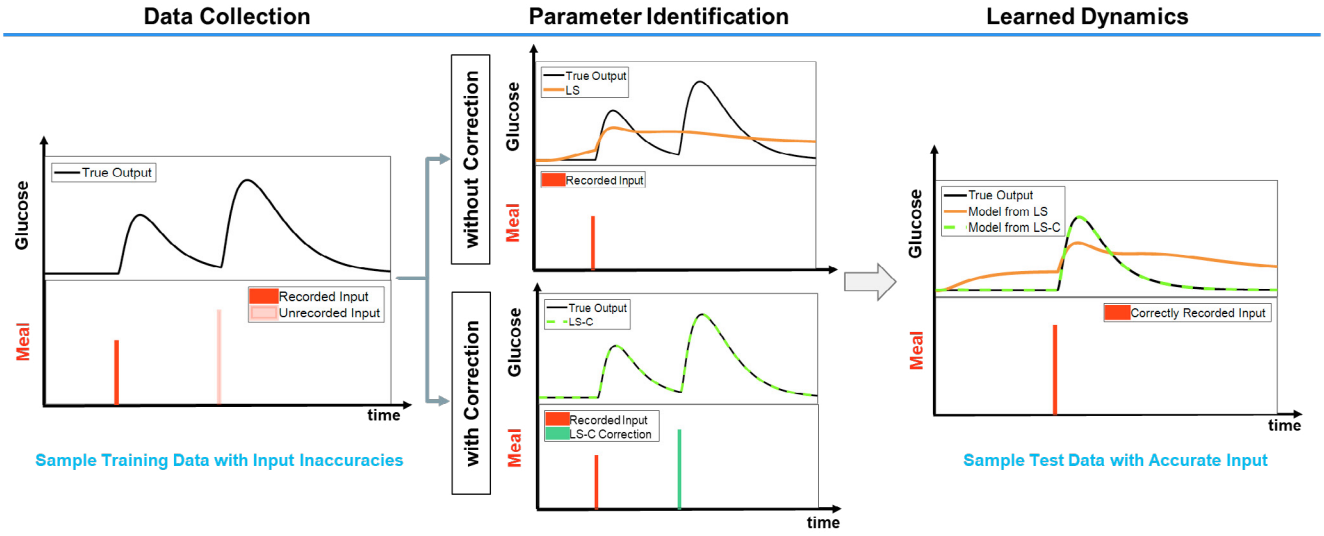

Fig. 1. Schematic example showing how data with input inaccuracies (an unrecorded meal intake) leads to learning an inaccurate model using standard least-squares (LS) while our method that involves sparse input corrections (LS-C) identifies the correct model.

$C = (1, 0, 0, 0, 0, 0, 0)^T$ with Δt being the time-step, and $\theta = \{\theta_p, \theta_f\}$ are the model parameters. We set $\Delta t = 5$ min as in Patek et al. (2016).

All equations in (1) except (1a) are linear in state. As it will be explained later in the paper, we require the model to be linear in state and inputs. We can make a linear approximation to $X_I(k)G(k)$ term in (1a) around the equilibrium point $(X_I = 0, G = G_b)$ and arrive in:

$$\dot{G}(k+1) \approx S_g(G_b - G(k)) - X_I(k)G_b + \frac{R_a(k)}{V_g}. \quad (3)$$

By replacing (1a) with (3) and some algebraic manipulation, we are able to write the state evolution as:

$$\begin{aligned} x(k+1) &= Ax(k) + c + B_i u_i(k) + B_m u_m(k), \\ y(k) &= Cx(k), \end{aligned} \quad (4)$$

where $A \in \mathbb{R}^{7\times 7}$ is derived from packing the terms in (2) (not shown here for brevity) and $c = (S_g G_b, 0, 0, 0, 0, 0, 0)^T$ is the term derived from (3).

Given a time length, N, a vector of insulin inputs $\mathbf{u}_i \in \mathbb{R}^N$, and a vector of meal inputs $\mathbf{u}_m \in \mathbb{R}^N$, one can simulate (4) forward to obtain the simulated outputs $\hat{\mathbf{y}} \in \mathbb{R}^N$ in the following compact form:

$$\hat{\mathbf{y}} = \mathcal{F}(\mathbf{u}_i, \mathbf{u}_m, \theta_p), \quad (5)$$

where the k^{th} values of $\hat{\mathbf{y}}, \mathbf{u}_i, \mathbf{u}_m$ are $\hat{y}(k)$, $u_i(k)$, $u_m(k)$ in (2), respectively. Note that $\mathcal{F}$ is linear in inputs $\mathbf{u}_i$ and $\mathbf{u}_m$ but nonlinear in θ_p. The initial state $x(0)$ plays a negligible role in long-term simulations as its effect is dissipated after few hours into the simulation (warm-up period).

3.3 Model Identification with Incomplete Input Data

The data collected from a person with T1D can be described with the following tuple:

$$\mathcal{D} := (\mathbf{y}, \mathbf{u}_i, \mathbf{u}_m)\,. \quad (6)$$

where $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{u}_i \in \mathbb{R}^N$, and $\mathbf{u}_m \in \mathbb{R}^N$ are the vectors of recorded glucose measurements, insulin injections, and carbohydrate intakes, respectively, and N is the length of data.

As mentioned earlier, the values are recorded every 5 minutes (i.e., one day of data means $N = 288$).

Our goal is to find a personalized model that fits its data. Given $\mathcal{D}$, we can identify the optimal θ_p, denoted by θ_p^*, by solving the following Least-Squares (LS) problem:

$$\boxed{\text{LS}} \quad \min_{\theta_p} \|\mathbf{y} - \mathcal{F}(\mathbf{u}_i, \mathbf{u}_m, \theta_p)\|_2^2, \quad (7)$$

Solving (7) is the standard approach to find θ_p^* but it is known to suffer from misreported/unreported input data. The noise in $\mathbf{y}$ and $\mathbf{u}_i$ are relatively negligible, but the values in $\mathbf{u}_m$ are more prone to errors since the estimation can be subjective and recording is at individual discretion (e.g., a person may forget to report a meal during a busy time). Instead of estimating the initial state $x(0)$, we set it as an equilibrium of (2) with nominal parameters and allow five hours of warm-up before incorporating the corresponding values in $\hat{\mathbf{y}}$ into the cost in (7). We desire the model identification problem to be guarded against incomplete data. Our proposed approach is to *correct* $\mathbf{u}_m$ by adding a correction signal $\mathbf{u}_c$ such that resulting parameter values represent the dynamics of the individual's metabolism with high fidelity as shown in Fig. 1. We also desire $\mathbf{u}_c$ to be sparse such that it only acts on significant differences between the model predicted and observed output. We can now formulate the problem.

Problem. (Least-Squares with Corrections (LS-C)) Given $\mathcal{D}$ as in (6), integer $\gamma \geq 0$, find the optimal model parameters and corrections to meal inputs by minimizing the sum of the squared errors, while the number of non-zero meal correction instances does not exceed γ:

$$\boxed{\text{LS-C}} \quad \begin{aligned} \min_{\theta_p, \mathbf{u}_c} \quad & \|\mathbf{y} - \mathcal{F}(\mathbf{u}_i, \mathbf{u}_m + \mathbf{u}_c, \theta_p)\|_2^2, \\ \text{subject to} \quad & \|\mathbf{u}_c\|_0 \leq \gamma. \end{aligned} \quad (8)$$

where we use $\|\cdot\|_2$ and $\|\cdot\|_0$ to represent Euclidean norm and the number of non-zero elements (zero semi-norm), respectively. The value of γ is determined prior to solving (8) through the heuristic explained in Section 4.2.

An alternative to $\|\mathbf{u}_c\|_0 \leq \gamma$ is adding $\|\mathbf{u}_c\|_2^2$ to the cost in (8) so that adding more and larger corrections would be penalized. We argue that (8) is better suited to deal with incomplete inputs as corrections occur without a cost. The only restriction is sparsity imposed by the cardinality constraint which suits the nature of typical errors in input data in the daily care of T1D (e.g., incorrect meal report).

4. SEQUENTIAL MIXED-INTEGER OPTIMIZATION

Problem (8) falls into the category of nonlinear optimization as $\mathcal{F}$ is nonlinear in θ_p in (1). Furthermore, the cardinality constraint $\|\mathbf{u}_c\|_0 \leq \gamma$ is non-convex and makes the problem combinatorial. Therefore, solving (8) is challenging as (i) gradient-based optimization is not suitable to handle the cardinality constraint and (ii) the problem size (the number of decision variables in θ_p and $\mathbf{u}_c$) is too large for using global search methods such as simulated annealing or genetic algorithm.

By fixing $\mathbf{u}_c$, (8) becomes a standard, smooth, nonlinear least squares problem, which can be solved to local optimality using standard gradient-based methods (Bertsekas, 1997). Furthermore, we can cast $\|\mathbf{u}_c\|_0 \leq \gamma$ as a set of mixed-integer constraints (Section 4.3). While mixed-integer optimization is computationally prohibitive, there exist mature commercial solvers such as (Gurobi Optimization, 2018) that are able to solve moderately sized mixed-integer linear/quadratic programs to near global optimality in reasonable times. However, they cannot handle nonlinear constraints on θ_p. Therefore, we propose an alternating minimization approach to solve (8). In the first step, we fix $\mathbf{u}_c$ and solve for θ_p using a standard nonlinear least-squares solver. In the second step, we fix θ_p and use MIQP to solve $\mathbf{u}_c$. We alternate between these steps until the latest solution cost of (9) is within 5% of the previous iteration and the maximum change in the parameters in θ_p remains within 10% of their values from the previous iteration. The pseudo-code is shown in Algorithm 2. The details are explained in the following sections.

4.1 Nonlinear Least-Squares

Given $\mathbf{u}_c$, we solve the following version of (8):

$$\min_{\theta_p} \|\mathbf{y} - \mathcal{F}(\mathbf{u}_i, \mathbf{u}_m + \mathbf{u}_c, \theta_p)\|_2^2 + (\theta_p - \bar{\theta}_p)'\Lambda(\theta_p - \bar{\theta}_p), \quad (9)$$

where $\bar{\theta}_p$ stands for the nominal parameters and Λ is the regularization matrix. While (9) is a nonlinear program, it is smooth and can be solved to local optimality. When solving (9) in the intermediate steps of Algorithm 2, we warm start (9) with the previous θ_p. In the very first iteration, all elements of $\mathbf{u}_c$ are set to zero and the population-level values for θ_p given in Patek et al. (2016) are used as a reasonable initial guess.

4.2 Choosing the number of corrections (γ)

We conjecture that a "significant difference" between the model prediction and the output implies incomplete input within the few hours leading to the difference. We define significant difference as an Absolute Relative Difference (ARD) value $\geq 20/20\%$. The ARD is calculated as follows:

Algorithm 1: Estimation of γ

Data: $\mathbf{y}$: recorded glucose trace
$\hat{\mathbf{y}}$: output of initial fit from the LS method
Result: Number of corrections γ
Initialize: $\gamma = 0, j = 1$
while $j \leq N$ **do**
  if $ARD(y(j), \hat{y}(j)) > 20$ **then**
    $\gamma \leftarrow \gamma + 1$, $j \leftarrow j + 48$ ▷ Advance 4 hours
  else
    $j \leftarrow j + 1$ ▷ Advance 5 minutes
  end
end
return γ

Algorithm 2: Sequential Optimization for LS-C

Data: Recorded data $\mathcal{D} = (\mathbf{y}, \mathbf{u}_i, \mathbf{u}_m)$
Result: Optimized SOGMM parameters θ_p^*
Initialize: $\mathbf{u}_c = 0$, $i = 0$, $\theta_{p,0} = \bar{\theta}_p$
while *Not Convergence Criteria* **do**
  $\mathbf{u}_m \leftarrow \mathbf{u}_m + \mathbf{u}_c$ and solve (9) to get $\theta_{p,i}^*$
  if $i = 0$ **then**
    Estimate γ from Algorithm 1
  end
  $\theta_p \leftarrow \theta_{p,i}^*$ and solve (11) to get $\mathbf{u}_c$
  $i \leftarrow i + 1$
end
Return $\theta_p^* = \theta_p$

$$\mathrm{ARD}(y, \hat{y}) := \begin{cases} |y - \hat{y}| & y \leq 100 \text{ mg/dL}, \\ 100\dfrac{|y - \hat{y}|}{y} & \text{otherwise.} \end{cases} \quad (10)$$

The selected ARD threshold is about twice the mean ARD of the FDA-cleared CGM devices between 2016 and 2021 (Bailey and Alva, 2021). Thus, regular CGM noise does not increase γ. After solving (9) for the first time, we obtain an estimate of the number of corrections γ (see the reference to Algorithm 1 in Algorithm 2). Note that every time we increase γ, we avoid increasing it again for the subsequent four hours regardless of the ARD value within that period (see Algorithm 1). This is due to the observation that meal intake disturbs glucose levels significantly for three hours and its total absorption takes up to six hours (American Diabetes Association, 2001). Thus, it is safe to assume that each correction can be used to restore the compromised input-output relationship for the subsequent four hours in the identification process.

4.3 Mixed-Integer Quadratic Programming

Using the big-M method (Bertsimas and Weismantel, 2005), we cast $\|\mathbf{u}_c\|_0 \leq \gamma$ as mixed-integer constraints. We obtain the following MIQP:

$$\begin{aligned} \min_{\mathbf{u}_c, \mathbf{v}, \mathbf{b}} \quad & \|\mathbf{y} - \mathcal{F}(\mathbf{u}_i, \mathbf{u}_m + \mathbf{u}_c, \theta_p)\|_2^2, \\ \text{subject to} \quad & u_c(k) - v(k) \geq -M(1 - b(k)), \\ & u_c(k) - v(k) \leq M(1 - b(k)), \\ & -Mb(k) \leq u_c(k) \leq Mb(k), \\ & -M \leq v(k) \leq M, \\ & b(k) \in \{0, 1\}, k = 1, \cdots, N, \\ & \sum_{k=1}^{N} b(k) = \gamma, \end{aligned} \quad (11)$$

where v and b are auxilary continuous and binary variables, respectively, and M is a large number that is an upper bound for any correction magnitude. It is straightforward to check that if $b(k) = 0$ in the constraints, then $u_c(k) = 0$ and if $b(k) = 1$, then $u_c(k)$ can take any value in the range $[-M, M]$.

5. CASE STUDIES

We evaluate the performances of the nominal, LS, and LS-C methods in model identification through in-silico studies using 10 adult in-silico subjects in the UVA/Padova simulator (Dalla Man et al., 2014). We train the models for each subject under scenarios with incomplete input data and then test the performance of the identified models on data with complete information.

Training Scenarios. The experiments are two-day long with three main meals and one snack each day (Day 1: breakfast of 30 g at 7 AM, lunch of 90 g at 2 PM, dinner of 50 g at 7 PM, snack of 40 g at 11 PM; Day 2: breakfast of 40 g at 8 AM, lunch of 60 g at 2 PM, dinner of 30 g at 7 PM, snack of 30 g at 2 AM). Total carbohydrate intake is selected similar to real-life (Aiello et al., 2022). We consider one baseline scenario to be with complete information and four scenarios, each with some unreported or misreported information in $\mathcal{D}$, that can be experienced in real life. As in real life, CGM noise is also included. The scenarios are illustrated in Fig. 2 and summarized below.

I (*Complete Data*): All inputs are recorded accurately.
II (*Unreported snacks*): People tend to not bolus for snacks and hence, the snacks remain unreported (VanderWel et al., 2010). Here, this behavior is implemented for the bedtime snacks on both days.
III (*Miscalculated meals*): Carbohydrate counting is a major challenge in daily T1D care (Lane et al., 2021) and errors are frequent (Brazeau et al., 2013). This is implemented through a 30% overestimation on the biggest meal, lunch on Day 1, and 30% underestimation on the breakfast of Day 2.
IV (*Delays in eating a bolused meal*): People may have delays in eating the meal that they have already bolused for (Briscoe and Davis, 2006). For instance, a delay in the arrival of an ordered meal may cause such situation. As such, this situation combines both misreporting (full meal reported when none is eaten at the time of the insulin bolus) and not reporting meal intake (meal eaten at a later time than reported). Here, this is implemented on the lunch of Day 1. The lunch is reported and bolused for at 2 PM while it is eaten at 3 PM.
V (*Occlusion*): In the case of a "pump occlusion", insulin is delivered by the pump but not received by the user. In this scenario, on Day 2, the insulin bolus for breakfast and the basal infusion in the following hour is not received by the user due to occlusion. Here, insulin is the misreported input as the data suggest all insulin was delivered.

Test Scenario. After obtaining model parameters for each in-silico subject, we evaluate the prediction capabilities of these models on a separate test scenario where input-output information are reported accurately. If LS-C successfully recovers the effects of inaccurate input

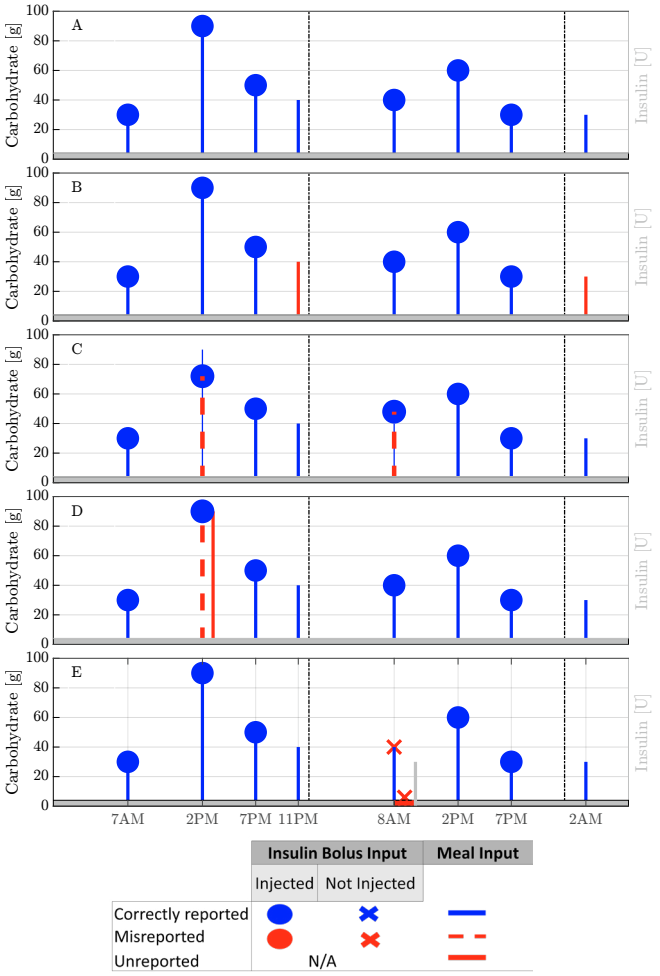

Fig. 2. (A) Scenario I; (B) Scenario II; (C) Scenario III; (D) Scenario IV; (E) Scenario V. Recorded data for meal and insulin are reported with a blue line and a blue dot, respectively. The inaccuracies are shown in red. Gray-colored inputs show meal-independent insulin injections, i.e. basal insulin and correction bolus. Insulin doses are not shown since they vary for each in-silico subject based on the subject-specific insulin treatment parameters.

records in the training data, the resulting model will have improved prediction performance on the test data. The two-day long test scenario has three meals and one snack each day. Meal times, meal sizes, unbolused snack time are different from the training scenario and are designed as follows: Day 1 (breakfast of 40 g at 7 AM, snack of 20 g at 10 AM, lunch of 70 g at 2 PM, dinner of 60 g at 8 PM), Day 2 (breakfast of 30 g at 8 AM, snack of 10 g at 10 AM, lunch of 50 g at 1 PM, dinner of 80 g at 8 PM).

Results. For each model, we compute the root-mean-square error (RMSE) of the glucose predictions in the test scenario. Table 2 shows the median and interquartile range (IQR) of the RMSE across 10 in-silico subjects. The results are presented for each scenario and identification method (i.e., nominal model, the LS method, and the LS-C method). Due to the small sample size, Wilcoxon signed-rank test is used to evaluate the performance differences between LS and LS-C. In each statistical test, the sample

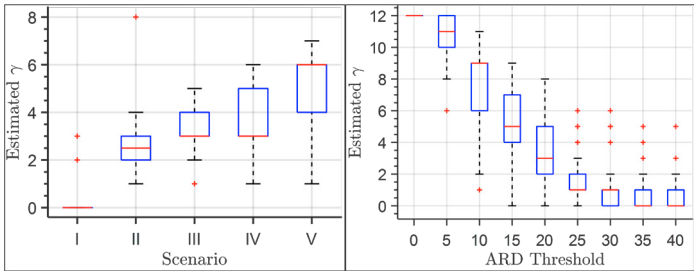

Fig. 3. [Left] Box plots of the estimated number of corrections (γ) across all subjects per scenario for the selected ARD threshold (20/20%). [Right] Estimated γ outcomes across all scenarios for different ARD threshold choices. The results in both plots are based on the two-day long training scenarios.

size was 10, equal to the number of in-silico subjects, as simulations were run for each subject-scenario only once. A two-tailed significance threshold of $\alpha = 0.05$ was used.

Fig.3 [Left] illustrates the distribution of γ across all subjects per scenario. LS-C suggested the lowest number of corrections for Scenario I and the highest number of corrections for Scenario V, when evaluated via median [IQR]: 0 [0] for Scenario I, 2.5 [1] for Scenario II, 3 [1] for Scenario III, 3 [2] for Scenario IV, and 6 [1.75] for Scenario V. We also provide the a posteriori analysis of how estimated γ changes for different thresholds of ARD across all scenarios in Fig.3 [Right]. The trends are as expected: more corrections are needed when data incompleteness is "more severe" or when ARD thresholds are lower. We argue that the selected ARD threshold (20/20%) is around the knee of the curve in Fig. 3 [Right].

As expected, no performance difference was observed for models trained on data with complete input information (Scenario I), while in others the *LS-C algorithm learns a more accurate model* that leads to significantly lower test scenario RMSE ($p < 0.05$). The largest performance improvement is obtained in Scenario II, which represents a critical, yet common, condition of two unreported meals that have a significant impact on glucose levels. A small improvement is reported in Scenario III and incomplete data are successfully recovered in Scenario IV. In Scenario V, we have pump occlusion, which is difficult to detect in practice (Klonoff et al., 2017), and demonstrate that our algorithm can also handle a condition that the inaccuracy does not involve a meal. Fig.4 illustrates an example of correcting an inaccurate input, where the 2 PM meal on day 1 was digested one hour after it was recorded. Compared to the LS method, LS-C yields a superior glucose prediction performance on the training data as well as on test data.

Table 2. RMSE (mg/dL) on test data

	Scenario				
Model	I	II	III	IV	V
Nominal	42.5[4.7]	42.5[4.7]	42.5[4.7]	42.9[4.7]	40.7[5.9]
LS	11.3[3.0]	18.1[2.2]	17.2[1.2]	16.4[1.6]	27.5[4.1]
LS-C	11.3[3.0]	12.6[3.9]	15.1[1.0]	12.5[5.0]	22.6[8.0]
p-value*	0.232	0.010	0.014	0.004	0.027

Values are reported as median[IQR] across 10 in-silico subjects.
* Results from Wilcoxon singed-rank test that compares the paired performances of LS and LS-C across 10 in-silico subjects.

6. DISCUSSION AND FUTURE WORK

Our results show the importance of accurate input information in model identification and suggest that the proposed method, LS-C, is effective in reconciling the unfavorable effects of imperfect input in model identification for T1D. While misreported meals are the primary motivation to apply corrections through meal input, it is not the only one. Phenomena such as different macro-nutrients, acute psychological stress, and physical activities are unreported and unmodeled but have profound effects on insulin-glucose dynamics. These effects can also be approximately modeled by "artificial meals" as first proposed and further elaborated in Patek et al. (2016). For instance, physical exercise is known to cause a sharp decrease in glucose values and hence can be roughly viewed having the person having taken a negative meal. Correcting incompleteness in non-meal input data by adding artificial meals is not ideal, but possible - as exemplified in Scenario V, where the lack of insulin due to the pump occlusion is compensated via additional meal inputs. Note that conditions that rather impact the model parameters, such as sickness or pregnancy, would appear in the model parameter values instead of meal corrections as long as the parameters are sufficient to capture these dynamics. Similarly, our method can be extended to correct significant output inaccuracies affecting model identification outcomes, such as low glucose readings due to compression on the CGM sensor (Mensh et al., 2013). There is value in future studies that explore a variety of corrections (e.g., corrections in the form of step or ramp instead of impulse inputs), the effect of combined inaccuracies, and the performance of the proposed method on clinical data.

REFERENCES

Aiello, E.M., Toffanin, C., Magni, L., and De Nicolao, G. (2022). Model-based identification of eating behavioral patterns in populations with type 1 diabetes. *Control Engineering Practice*, 123, 105128.

American Diabetes Association (2001). Postprandial blood glucose.

American Diabetes Association (2021). Introduction: Standards of medical care in diabetes—2022.

Bailey, T.S. and Alva, S. (2021). Landscape of continuous glucose monitoring (cgm) and integrated cgm: accuracy considerations. *Diabetes Technology & Therapeutics*, 23(S3), S–5.

Bertsekas, D.P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3), 334–334.

Bertsimas, D., Kitane, D.L., Azami, N., and Doucet, F. (2020). Novel mixed integer optimization sparse regression approach in chemometrics. *Analytica Chimica Acta*, 1137, 115–124.

Bertsimas, D. and Weismantel, R. (2005). *Optimization over integers*, volume 13. Dynamic Ideas, Belmont.

Brazeau, A., Mircescu, H., Desjardins, K., Leroux, C., Strychar, I., Ekoé, J., and Rabasa-Lhoret, R. (2013). Carbohydrate counting accuracy and blood glucose variability in adults with type 1 diabetes. *Diabetes Research and Clinical Practice*, 99(1), 19–23.

Briscoe, V.J. and Davis, S.N. (2006). Hypoglycemia in type 1 and type 2 diabetes: physiology, pathophysiology, and management. *Clinical Diabetes*, 24(3), 115–121.

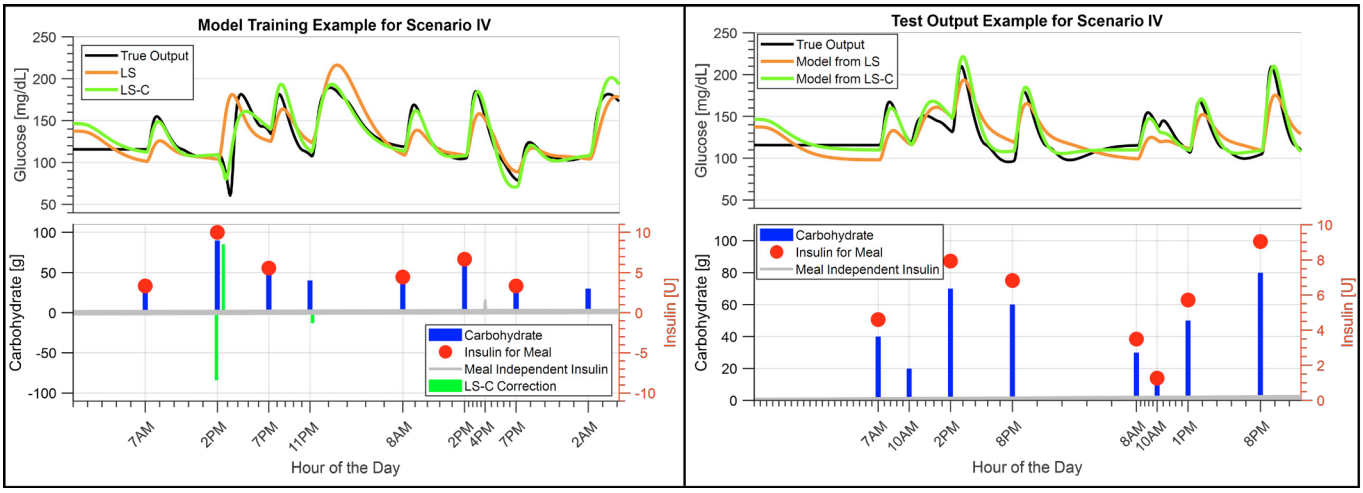

Fig. 4. [Left] Sample training output for in-silico Subject#6 under Scenario IV. 2 PM meal was digested one hour after it was recorded. Top panel: Model fits for identification with LS (orange line) versus with LS-C (green line) are compared with the original output (black line); Bottom panel: Meal and insulin input data [Right] Top panel: Sample test output for in-silico Subject#6 with the models identified with the LS (orange line), with the LS-C algorithm (green line), and the original output (black line). Bottom panel: Meal and insulin input data are reported with a blue line and a red dot, respectively. Meal-independent insulin corresponds to basal or correction insulin.

Colmegna, P., Wang, K., Garcia-Tirado, J., and Breton, M.D. (2020). Mapping data to virtual patients in type 1 diabetes. *Control Engineering Practice*, 103, 104605.

Dalla Man, C., Micheletto, F., Lv, D., Breton, M., Kovatchev, B., and Cobelli, C. (2014). The uva/padova type 1 diabetes simulator: new features. *Journal of Diabetes Science and Technology*, 8(1), 26–34.

Garcia-Tirado, J., Zuluaga-Bedoya, C., and Breton, M.D. (2018). Identifiability analysis of three control-oriented models for use in artificial pancreas systems. *Journal of Diabetes Science and Technology*, 12(5), 937–952.

Gurobi Optimization, L. (2018). Gurobi optimizer reference manual. `https://www.gurobi.com`.

Holt, R.I., DeVries, J.H., Hess-Fischl, A., Hirsch, I.B., Kirkman, M.S., Klupa, T., Ludwig, B., Nørgaard, K., Pettus, J., Renard, E., et al. (2021). The management of type 1 diabetes in adults. a consensus report by the american diabetes association (ada) and the european association for the study of diabetes (easd). *Diabetes Care*, 44(11), 2589–2625.

Hughes, J., Gautier, T., Colmegna, P., Fabris, C., and Breton, M.D. (2021). Replay simulations with personalized metabolic model for treatment design and evaluation in type 1 diabetes. *Journal of Diabetes Science and Technology*, 15(6), 1326–1336.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer, New York.

Klonoff, D.C., Freckmann, G., and Heinemann, L. (2017). Insulin pump occlusions: for patients who have been around the (infusion) block. *Journal of Diabetes Science and Technology*, 11(3), 451–454.

Kushner, T., Breton, M.D., and Sankaranarayanan, S. (2020a). Multi-hour blood glucose prediction in type 1 diabetes: A patient-specific approach using shallow neural network models. *Diabetes Technology & Therapeutics*, 22(12), 883–891.

Kushner, T., Sankaranarayanan, S., and Breton, M. (2020b). Conformance verification for neural network models of glucose-insulin dynamics. In *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control*, 1–12.

Lane, W., Lambert, E., George, J., Rathor, N., and Thalange, N. (2021). Exploring the burden of mealtime insulin dosing in adults and children with type 1 diabetes. *Clinical Diabetes*, 39(4), 347–357.

Marseglia, G.R. and Raimondo, D.M. (2017). Active fault diagnosis: A multi-parametric approach. *Automatica*, 79, 223–230.

Mensh, B.D., Wisniewski, N.A., Neil, B.M., and Burnett, D.R. (2013). Susceptibility of interstitial continuous glucose monitor performance to sleeping position. *Journal of Diabetes Science and Technology*, 7(4), 863–870.

Narasimhamurthy, M., Kushner, T., Dutta, S., and Sankaranarayanan, S. (2019). Verifying conformance of neural network models. In *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1–8. IEEE.

Patek, S.D., Lv, D., Ortiz, E.A., Hughes-Karvetski, C., Kulkarni, S., Zhang, Q., and Breton, M.D. (2016). Empirical representation of blood glucose variability in a compartmental model. In *Prediction Methods for Blood Glucose Concentration*, 133–157. Springer.

Roll, J., Bemporad, A., and Ljung, L. (2004). Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1), 37–50.

Toffanin, C., Del Favero, S., Aiello, E.M., Messori, M., Cobelli, C., and Magni, L. (2017). Mpc model individualization in free-living conditions: a proof-of-concept case study. *IFAC-PapersOnLine*, 50(1), 1181–1186.

VanderWel, B.W., Messer, L.H., Horton, L.A., McNair, B., Cobry, E.C., McFann, K.K., and Chase, H.P. (2010). Missed insulin boluses for snacks in youth with type 1 diabetes. *Diabetes Care*, 33(3), 507–508.