Published in final edited form as:

IEEE Trans Control Syst Technol. 2020 January ; 28(1): 63–78. doi:10.1109/TCST.2018.2871871.

System Identification Approaches For Energy Intake Estimation: Enhancing Interventions For Managing Gestational Weight Gain

Penghong Guo, Daniel E. Rivera [Senior Member, IEEE]

Control Systems Engineering Laboratory (CSEL), School for Engineering of Matter, Transport, and Energy, Arizona State University, Tempe, AZ, 85281 USA.

Penghong Guo: penghong.guo@asu.edu; Daniel E. Rivera: daniel.rivera@asu.edu

Jennifer S. Savage,

Center for Childhood Obesity Research, Pennsylvania State University, University Park, PA, USA

Department of Nutritional Sciences, Pennsylvania State University, University Park, PA, USA.

Emily E. Hohman,

Center for Childhood Obesity Research, Pennsylvania State University, University Park, PA, USA

Abigail M. Pauley,

Exercise Psychology Laboratory, Department of Kinesiology, Pennsylvania State University, University Park, PA, USA

Krista S. Leonard,

Exercise Psychology Laboratory, Department of Kinesiology, Pennsylvania State University, University Park, PA, USA

Danielle Symons Downs

Exercise Psychology Laboratory, Department of Kinesiology, Pennsylvania State University, University Park, PA, USA

Department of Obstetrics and Gynecology, Pennsylvania State College of Medicine, Hershey, PA, USA.

Jennifer S. Savage: jfs195@psu.edu; Emily E. Hohman: eeh12@psu.edu; Abigail M. Pauley: amp34@psu.edu; Krista S. Leonard: kbl5167@psu.edu; Danielle Symons Downs: dsd11@psu.edu

Abstract

Excessive maternal weight gain during pregnancy represents a major public health concern that calls for novel and effective gestational weight management interventions. In *Healthy Mom Zone* (HMZ), an on-going intervention study, energy intake underreporting has been found to be an important consideration that interferes with accurate weight control assessment, and the effective use of energy balance models in an intervention setting. In this paper, a series of estimation approaches that address measurement noise and measurement losses are developed to better understand the extent of energy intake underreporting. These include back-calculating energy intake from an energy balance model developed for gestational weight gain prediction, a Kalman filtering-based approach to recursively estimate energy intake from intermittent measurements in real-time, and an approach based on semi-physical identification principles which features the capability of adjusting future self-reported energy intake by parameterizing the extent of underreporting. The three approaches are illustrated by evaluating with participant data obtained

through the HMZ intervention study, with the results demonstrating the potential of these methods to promote the success of weight control. The pros and cons of the presented approaches are discussed to generate insights for users in future applications.

Keywords

State estimation; system identification; semi-physical identification; Kalman filter; intermittent measurements; underreporting; obesity; weight interventions

Nomenclature

$\alpha_1, \alpha_2, \beta, \xi$	Semi-physical model parameters	
γ	Bernoulli variable to indicate the loss of measurements	
λ_{FFM}	Energy density of fat free mass	771 kcal/kg
λ_{FM}	Energy density of fat mass	9500 kcal/kg
$\mathcal{O}$	Observability matrix	
$\mathcal{R}$	Semi-physical estimation regressor matrix	
$\mathcal{Z}$	Output vector for semi-physical model regression	
μ_s	Mean of signal s	
ν	Measurement noise vector	
ϕ	Empty set	
$\Phi_s(\omega)$	Power spectrum of signal s	
σ_s	Standard deviation of signal s	
θ	Semi-physical regression parameter vector	
ω	Process noise vector	
$\widetilde{EI}_{rept}$	Nominal values of self-reported energy intake	kcal
A, B, C, D	System matrices	
a_p, b_p	Coefficients for the linear functions of total body protein with respect to weight	
a_w, b_w	Coefficients for the linear functions of total body water with respect to weight	
CM	Body mass that remains constant during gestation	kg
e_{pred}	Prediction error	
EE	Daily maternal energy expenditure	kcal
$EI(k)$	Daily maternal energy intake	kcal
EI_{actual}	Maternal actual energy intake	kcal
EI_{est}	Estimated energy intake	kcal
EI_{rept}	Self-reported energy intake	kcal
ES	Daily maternal energy stored	kcal
FFM	Fat free mass	kg
FM	Fat mass	kg
g	Nutrient partitioning constant	0.03
GWG	Gestational weight gain	kg
I	Identity matrix	
i, j, l, m, N, p, q	Indices	

$I_{1,2,\dots,N}$	Intervention components	
k	Discrete time index	day
K_1, K_2	System gain coefficients for energy balance model	kg/kcal/day
n_s	Noise of signal s	
P	Error covariance matrix for Kalman filter	
$PA(k)$	Daily maternal physical activity	kcal
Q	Covariance matrix for process noise, ω	
R	Covariance matrix for measurement noise, ν	
r_{TEF}	Energy expenditure from thermal effect of food stated as a percentage of energy intake	
$RMR(k)$	Daily maternal resting metabolic rate	kcal
RMS	Root mean square	
$RMSE$	Root mean square error	
SD	Standard deviation for estimates	
T	Sampling time	day
t	Continuous time index	day
TBP	Total body protein	kg
TBW	Total body water	kg
TEF	Energy expenditure due to thermic effect of food	kcal
u	Input vector	
$W(k)$	Maternal total weight, obtained daily	kg
W_{actual}	Actual total weight	kg
W_{est}	Estimated total weight	kg
W_{meas}	Measured total weight	kg
x	State vector	
$X_{\mathbb{Z}_1, \mathbb{Z}_2}^-$	A sub-matrix of X by deleting the rows of X indexed in $\mathbb{Z}_1$ and the columns indexed in $\mathbb{Z}_2$, respectively	
y	Output vector	
BMI	Body mass index	kg/m ²
EB	Energy balance	
GA	Gestational age	day
HMZ	Healthy Mom Zone (name of the intervention study)	
KF	Kalman filter	
OB	Obese	
OW	Overweight	

I. Introduction

Over two thirds of childbearing-age females in the US are categorized as overweight (OW; body mass index [BMI] ≥ 25) or obese (OB, BMI ≥ 30) [1]. According to a 2014 study, a majority of the OW/OB women gain weight above the US Institute of Medicine (IOM) guidelines for adequate weight in pregnancy [2], [3]. High gestational weight gain (GWG) elevates the risk of maternal adverse obstetric outcomes, such as gestational diabetes mellitus, hypertension, emergency cesarean delivery, preeclampsia, and postpartum weight retention [3], [4], [5], [6]. Recent studies have also shown that it is associated with

macrosomia and early onset of obesity in offspring [2], [7], [8]. Consequently, there is a pressing need for interventions that can successfully manage GWG within the IOM guidelines, especially for OW/OB women.

Motivated by these needs, *Healthy Mom Zone* (HMZ) [9], a novel intervention conducted at Pennsylvania State University, aims to develop and validate an individually-tailored and intensively adaptive approach to effectively manage GWG in OW/OB women. The novelty of this intervention lies in its potential to shift the focus of weight management from a “one size fits all” approach to a customized intervention strategy that can adapt to individuals’ unique needs. Specifically, a closed-loop control strategy is employed for the design of the intervention to optimize the intervention dosages based on an individual participant’s intervention outcomes, such as diet or physical activity behaviors and weight changes. The system models that are used for the control algorithms involve a first-principles energy balance (EB) model to predict *GWG* from longitudinal measurements of maternal physical activity (*PA*), resting metabolic rate (*RMR*) and energy intake (*EI*) [10],[11]. The discretized form of the model can be expressed as:

$$GWG(k+1) = K_1EI(k) + K_2PA(k) + K_3RMR(k) \quad (1)$$

where the system output, *GWG*, is defined as the daily change of maternal weight (*W*), that is, $GWG(k+1) = W(k+1) - W(k)$; K_1 and K_2 are gain coefficients for the system inputs. Additional details regarding model development are presented in Section II.

In intervention practice, the use of the EB model for accurate weight gain prediction can be compromised due to biased or noise-corrupted input measurements; this has been found to be an issue of concern especially when using self-reported *EI* from the HMZ intervention study for maternal weight predictions. Energy intake misreporting is prevalent in the general adult population, estimated at 40 to 50% for underreporting and 5 to 10% for overreporting [12], [13]; the extent of underreporting can be as much as 59% of their total caloric intake [14]. BMI has been found to be a significant independent predictor of *EI* underreporting: higher extent of underreporting is observed with increasing BMI [15], [16]. Hence, the participants in the HMZ Study as OW/OB pregnant women are more likely to underreport their *EI*. The high prevalence of *EI* underreporting among OW/OB pregnant women has also been previously reported in [17], [18] and [19]. Misreporting of *EI* may relate to participant education, age, psychological status such as depression or poor body image [20]. It also might be due to recall bias or memory lapses, poor awareness of quantities or types of foods eaten, inaccurate portion size estimation, or the inconvenience of reporting [21]. Thus, the measurement bias due to *EI* misreporting can be characterized as both systematic and random, and remains a challenging issue for EB model predictions.

To illustrate underreporting in *EI* self-reports and their influence on weight predictions, Fig. 1 compares the measured weight from two representative HMZ participants with the EB model simulated weight using their self-reported *EI*. As shown in the figure, the discrepancies between the model predictions and weight measurements accumulate along the intervention weeks; similar observations also exist with other participant data that have been collected in the HMZ intervention. While multiple causes could explain this mismatch,

the most significant is the accumulation of errors resulting from underreported EI , which increases substantially over time as a result of the integrating dynamics of the system. From a practical standpoint, self-reported or self-monitored measurements are convenient to obtain from free-living participants, yet found to contain bias and measurement noise, along with data missingness due to lack of participant compliance with the monitors or adherence to interventions. These issues with participant data all pose challenges to reliable model-based estimation, and limit the assessment of intervention outcomes.

Ensuring adequate dietary intake is crucial for optimizing maternal and fetal outcomes during pregnancy [22]. Misreporting of food consumed and dietary calories makes it difficult for clinicians to determine if participants are meeting their energy intake goals, and prevents appropriate health counseling advice to be provided. However, there is a scarcity of literature examining or identifying the characteristics of EI misreporting in pregnancy. The goal of the current work is to accurately calculate participant energy intake, so that timely feedback can be provided to both participants and dietitians; consequently, nutrition counseling as well as suggestions regarding how to adjust physical activity behaviors can be tailored in order for intervention participants to manage their gestational weight gain in line with clinical recommendations. In this particular aspect, real-time estimation approaches that can address noise and measurement losses have significant appeal in real-world intervention settings.

An important end-use application for improved EI estimation is to improve the effectiveness of closed-loop interventions using control engineering principles. A block diagram for a control system incorporating this functionality is shown in Fig. 2. Here a hybrid model predictive control (HMPC) algorithm as described in [23], [24] is used to specify the dosages of intervention components (e.g., healthy eating active learning, physical activity active learning, goal setting) based on the assessments of participant behavior outcomes in real time. EI , as an input to the internal controller EB model, is critical to determine the appropriate control actions; consequently biased EI self-reports will negatively influence controller performance. Therefore, estimated EI measurements described in this work are essential for adequate performance of the closed-loop control system for interventions.

In this paper, a series of estimation approaches have been developed to better understand the issue of EI underreporting. One approach from the literature is to back-calculate EI from the EB model [11], [25]; this conventional approach, while simple to implement, suffers from a number of inherent limitations. To address this problem, a Kalman filter-based approach that can estimate EI in real-time from intermittent and noise-corrupted measurements is developed [26]. Both these two approaches are capable of providing accurate estimates of EI based on measures of weight and energy expenditure terms, but it is important to note that intensive data collection of EB variables is required within the intervention in order for these two methods to be implemented. To overcome these challenges, an approach based on semi-physical identification principles is proposed to model the functional relationship between the self-reports and true EI . Once the model is identified from the past data, it can be used to correct for future self-reports that might be subject to significant bias, and there is no need for measurements of EB variables other than EI or any inputs used in the correction model. In the development of this semi-physical approach, cross-validation procedures are applied

to test the model performance, from which the parsimonious yet accurate models with good predictive ability can be selected for future use. The effectiveness of all the approaches proposed in this paper is assessed from participant data obtained through the HMZ intervention study.

The paper is organized as follows. Section II presents an overview of a dynamical EB model as well as a detailed description of the HMZ intervention study. Section III starts from the description of the conventional back-calculation method, leading to the adaptive algorithm based on Kalman filtering for EI estimation in real-time. In Section IV, the semi-physical estimation approach to adjusting the self-reported EI is developed, and the limitations of this approach due to the input noise are demonstrated with prediction error analysis. Section V gives a summary of our conclusions.

II. Modeling and Background

A. Dynamical Energy Balance Model

The basis for all the three approaches explored in this paper relies on a closed-form EB model which is reformulated from a maternal EB model developed by Thomas *et al.* [10]. In this section, the original EB model will be reviewed, followed by a detailed description of the development of the reformulated EB model.

The EB model from [10] to predict GWG based on EI and energy expenditure (EE) relies on the principle of energy conservation,

$$ES(t) = (1 - g)EI(t) - EE(t) \quad (2)$$

where $ES(t)$ is the energy stored at time t , the parameter $g = 0.03$ is the nutrient partitioning constant. The excess energy is stored and converted into different body tissues (*e.g.* body fat and muscle tissue), leading to a weight change of different body compositions. Hence, ES can be constructed as a sum of the change of different energy storage compartments during pregnancy. Here, a two-compartment model which separates maternal weight (W) into two components, fat free mass (FFM) and fat mass (FM), is applied and can be expressed as,

$$W(t) = FFM(t) + FM(t) \quad (3)$$

With this two-compartment model assumption, the ES term in (2) can be expanded into the sum of the instantaneous weight change of the two components of FM and FFM , multiplied by their respective energy densities λ_{FM} and λ_{FFM} , leading to

$$\lambda_{FFM} \frac{dFFM(t)}{dt} + \lambda_{FM} \frac{dFM(t)}{dt} = (1 - g)EI(t) - EE(t) \quad (4)$$

In this model from [10], EE is estimated as a function of maternal W . In the case of small W changes, the EE quantity remains relatively constant. Despite the effectiveness and simplicity in weight gain predictions, this is not quite suited for an intervention application where individual levels of PA can be significantly increased as a result of intervention

sessions, leading to a substantial increase observed in total EE . Here, modifications are made to compensate for limitations of the modeling in [10] that the different components of maternal EE cannot be adjusted individually.

EE is commonly considered to be composed of PA , RMR , and the thermic effect of food (TEF) [27]. With commercially available accelerometers, PA can be easily measured. RMR is considered as dynamically changing throughout gestation as will be described in Section II-B; it can be estimated or measured with a metabolism device. TEF is usually expressed as a percentage of EI , ranging from 4.0% to 17.1% due to different diets [28]. Regardless of intake nutrients, we assume it to be approximately 7% of daily EI as measured in [29]. Thus, EE can be expressed as,

$$EE(t) = PA(t) + RMR(t) + r_{TEF}EI(t) \quad (5)$$

where $r_{TEF} = 0.07$. Substituting (5) and (3) into (4) gives,

$$\lambda_{FFM} \frac{dFFM(t)}{dt} + \lambda_{FM} \frac{d(W(t) - FFM(t))}{dt} = (1 - g - r_{TEF})EI(t) - PA(t) - RMR(t) \quad (6)$$

Instead of keeping two differential terms in (6), FFM can be expressed to relate to W so that the derivative term is only with respect to W , and the equations become much easier to solve. To begin with, FFM can be considered to be the sum of total body water (TBW), total body protein (TBP), and mass that stays constant during gestation (CM); this latter quantity includes bone mass for example. This leads to,

$$FFM(t) = TBW(t) + TBP(t) + CM \quad (7)$$

where $CM = FFM(0) - TBW(0) - TBP(0)$. TBW and TBP are linear functions of simultaneously measured W based on a participant's BMI, which can be expressed in a generalized form as,

$$TBW(t) = a_W W(t) + b_W \quad (8a)$$

$$TBP(t) = a_P W(t) + b_P \quad (8b)$$

where a_W , b_W , a_P , b_P are the coefficients of the corresponding functions, the values of which can be found in Table I. Note that TBP and W are expressed in kg, TBW in liters. Substituting (8) into (7) leads to,

$$FFM(t) = (a_W + a_P)W(t) + (b_W + b_P + CM) \quad (9)$$

which if substituted into (6) gives the differential equation expressed in terms of the derivative of W as,

$$\frac{dW(t)}{dt} = K_1 EI(t) + K_2 PA(t) + K_2 RMR(t) \quad (10)$$

where

$$K_1 = \frac{1 - g - r_{TEF}}{(a_W + a_P)\lambda_{FFM} + (1 - a_W - a_P)\lambda_{FM}} \quad (11a)$$

$$K_2 = \frac{-1}{(a_W + a_P)\lambda_{FFM} + (1 - a_W - a_P)\lambda_{FM}} \quad (11b)$$

K_1 and K_2 are system gain coefficients, expressed in kg/kcal/day. Table II shows the values of K_1 and K_2 for different categories of BMI. Eqn. (10) forms the basis of the model examined in this paper, from which we are able to predict the system output W using the explicitly measurable inputs EI , PA and RMR . Written in discretized form for daily sampling $T = 1$, the EB model can be expressed as,

$$GWG(k + 1) = K_1 EI(k) + K_2 PA(k) + K_2 RMR(k) \quad (12)$$

where GWG is defined by $GWG(k+1) = W(k+1) - W(k)$.

B. HMZ Intervention

In the HMZ intervention study, maternal W , EI and PA of 27 OW/OB pregnant women (age mean: 30.6; standard deviation (SD): 3.0, and pre-pregnancy body mass index (BMI) mean: 31.6, SD : 7.1) were measured for 22–28 weeks. The developed estimation methods in this paper will be tested with two representative participants shown in Fig. 1: participant A (BMI = 28.6; age = 31; OW) from the intervention group and participant B (BMI = 25.3; age = 37; OW) from the control group. The selection of these two participants aims to embrace the different characteristics or aspects of the two study groups. However, additional participants need to be carefully analyzed to generalize any conclusions on the contrast between the two study groups.

For the measurement of W , the participants weigh themselves daily using Aria Wifi smart digital scales. Participant EI is obtained three days per week from self-reported EI using a dietary intake smartphone app (MyFitnessPal (MFP)); the measurements of PA are obtained using a wrist-worn commercial monitor (Jawbone UP) on a daily basis. RMR is estimated with the quadratic regression formula proposed by [30], which fits the data from [31] as shown below:

$$RMR = 0.1976W^2 - 13.424W + 1457.6 \quad (13)$$

where W is the maternal weight expressed in kg. Eqn (13) captures the slight increase of RMR as women gain weight during gestation. Some special considerations or manipulations of the participant data taken in this paper are described here. The PA signal is considered as having negligible measurement noise for purposes of this paper. Since PA signals are

relatively stationary, mean replacement is employed to impute any missing data points. Considering the participant burden of recording their food consumption, participants are not required to report the energy intake every day. This leads to a substantial number of gaps in the self-reported EI . For any days when the self-reports are not available or missing, linear interpolation is one approach that can be used to impute the missing days so that simulations can be performed. This can also apply to missing data that occurs during the collection of participant W . The missingness rates for the measured EB variables are tabulated for the two selected participants in Table III.

III. Estimation For Underreporting

In this section, energy intake is estimated with a back-calculation method. This is followed by a Kalman filtering approach which can provide real-time estimation while addressing the problem of measurement noise and data missingness simultaneously.

A. Back-calculation Approach

As shown in Fig. 1, bias between the simulated and measured weight is mostly due to the substantial under-reported EI . To test this hypothesis, we back-calculate EI using the reformulated EB model based on the measured W , PA , and the estimated RMR . Numerically approximating the derivative term in (10) using the 2nd order centered difference formula leads to the expression of the estimated $EI(EI_{est})$ as shown below,

$$EI_{est}(k) = \frac{-W(k+2) + 8W(k+1) - 8W(k-1) + W(k-2)}{12TK_1} - \frac{K_2}{K_1}(PA(k) + RMR(k)) \quad (14)$$

The variables are indexed by k with $k = 1, 2, \dots, N$ corresponding to day 1 to day N . T is the sampling time, in this case, $T = 1$ day. The noise in W is small relative to the total W ; however the extent of this noise can significantly affect the numerical calculation of the rate of weight gain per day. Consequently, a 9-day (± 4 days) moving average filter is used to smooth the measured W before an estimate of the daily EI is obtained from (14). The selected length of the smoothing window is in agreement with the experience of behavioral scientists that a seven day window (or longer) is needed to accurately reflect typical daily EI [32].

The back-calculation result is shown in Fig. 3 where the back-calculated EI is generally higher than the reported EI measurements. The difference between the back-calculated EI and the self-reported EI is quantified using the mean and its SD . The mean $\pm SD$ of the EI estimate for participant A is 2660 ± 570 kcal; the mean $\pm SD$ for participant B is 2617 ± 434 kcal. One can also notice that the predicted W using the back-calculated EI follows closely with the measured W , which also provides support for the accuracy of the EI back-calculation from this method.

The method of using algebraic back-calculation for EI is the simplest and quickest approach to implement, and repeated EI estimates can be generated without requiring any EI measurements. Since estimation with this method is sensitive to the noise in weight measurements, pre-processing of measured weight data with techniques such as moving average filters is necessary to reduce variability in the estimates. It is also observed that, while the standard deviation can be significantly reduced by increasing length of the moving average window, the average of the estimates is not affected. The selection of the moving average window is an important adjustable parameter for smoothing, and needs to be determined based on the variance and length of the intervention.

B. Kalman Filtering With Intermittent Measurements

As demonstrated in the previous subsection, back-calculation is simple to implement, but the variation of the estimation results is largely due to the noise in the measured GWG (GWG_{meas}). Data imputation techniques such as linear interpolation are also required to address issue of missing data. In behavior weight management interventions, however, measurement loss is common in data collection. An on-line algorithm that can simultaneously address the issue of missing data and measurement noise to realize real-time estimation can be useful for the success of weight interventions. A recursive Kalman filter (KF) based approach for real-time state estimation from randomly intermittent measurements is developed in this section to address the measurement noise and inevitable partial data loss during interventions. The problem of Kalman filtering with intermittent measurements has been examined in [33], [34]. In practice of such long-duration interventions, it is common that the participant may occasionally forget to complete measures of a single or multiple determinants of the EB model on the same day. Hence, partial random data loss is considered and incorporated in the formulation of the algorithm. The detailed development of this Kalman filter-based algorithm based on intermittent measurements is elaborated in this paper, which demonstrates significant potential for real-time use. Such application has been demonstrated in [26] where the derivation and more details of the KF-based algorithm can be found.

Consider a multiple-input multiple-output (MIMO) discrete time linear system model with q output sensors, the state and measurement equations of which are described as below:

$$x_{k+1} = Ax_k + Bu_k + \varpi_k \quad (15a)$$

$$\begin{bmatrix} y_{1,k} \\ y_{2,k} \\ \vdots \\ y_{q,k} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_q \end{bmatrix} x_k + \begin{bmatrix} \nu_{1,k} \\ \nu_{2,k} \\ \vdots \\ \nu_{q,k} \end{bmatrix} \quad (15b)$$

$\underbrace{\hspace{1.5cm}}_{y_k} \quad \underbrace{\hspace{1.5cm}}_{C_k} \quad \underbrace{\hspace{1.5cm}}_{\nu_k}$

where k is the sampling time; $x_k, \varpi_k \in \mathbb{R}^l$ are the state and noise of the system, respectively;

$u_k \in \mathbb{R}^p$ is the system input; $y_k, \nu_k \in \mathbb{R}^{\sum_{i=1}^q m_i}$ are the measurement and measurement noise

with elements $y_{i,k}, \nu_{i,k} \in \mathbb{R}^{m_i}$, respectively; A , B , and C are the system matrices with appropriate dimensions. We assume that ϖ_k and $\nu_{i,k}$ are uncorrelated zero-mean Gaussian white noise with covariances of $Q > 0$ and $R_{ii} > 0$, respectively, that is $\varpi_k \sim \mathcal{N}(0, Q)$ and $\nu_{i,k} \sim \mathcal{N}(0, R_{ii})$; R is the block-diagonal covariance matrix for ν_k with the matrices R_{ii} on the diagonal, written as $R = \text{diag}\{R_{ii}, i = 1, 2, \dots, q\}$. For the system described per (15), we also assume that (A, B) is completely stabilizable and (A, C) completely detectable so that the error covariance of the Kalman filter converges to a unique value, in the case of no measurement loss.

To incorporate measurement loss in the algorithm, we use the Bernoulli variable $\gamma_{i,k}$ to model the random arrivals of the measurements: $\gamma_{i,k} = 1$ indicates that the measurement of the corresponding element, $y_{i,k}$, has been successfully received at time k , whereas its value of 0 indicates a measurement loss. $\gamma_{i,k}$ is assumed to have known probability distributions. Correlation is allowed to exist among the Bernoulli variables of $\gamma_{i,k}$ at same k , which can be denoted with a joint probability density function of $\Pr(\gamma_{1,k}, \gamma_{2,k}, \dots, \gamma_{q,k})$. Note that $\gamma_{i,k}$ assumes to be independent at different time instants.

Measurement loss can be treated equivalently as receiving a measurement with infinite noise variance. In the presence of measurement loss, the statistical characteristics of measurement noise will change accordingly and cannot be fully described with $\nu_{i,k}$ in (15b). Thus, a second measurement noise term $\nu'_{i,k}$ is introduced and defined with $\nu'_{i,k} \sim \mathcal{N}(0, R'_{ii})$ with $R'_{ii} \rightarrow \infty$. $\nu'_{i,k}$ has the same structure and dimensions as $\nu_{i,k}$. With the augmentation of the variables $\gamma_{i,k}$ and $\nu'_{i,k}$, the measurement equation per (15b) can be redefined for a general case with observation losses as

$$\underbrace{\begin{bmatrix} y_{1,k} \\ y_{2,k} \\ \vdots \\ y_{q,k} \end{bmatrix}}_{y_k} = \underbrace{\begin{bmatrix} \gamma_{1,k}(C_1 x_k + \nu_{1,k}) \\ \gamma_{2,k}(C_2 x_k + \nu_{2,k}) \\ \vdots \\ \gamma_{q,k}(C_q x_k + \nu_{q,k}) \end{bmatrix}}_{\tilde{C}_k} + \underbrace{\begin{bmatrix} (1 - \gamma_{1,k})\nu'_{1,k} \\ (1 - \gamma_{2,k})\nu'_{2,k} \\ \vdots \\ (1 - \gamma_{q,k})\nu'_{q,k} \end{bmatrix}}_{\tilde{C}_k} = \underbrace{\begin{bmatrix} \gamma_{1,k}C_1 \\ \gamma_{2,k}C_2 \\ \vdots \\ \gamma_{q,k}C_q \end{bmatrix}}_{\tilde{C}_k} x_k + \underbrace{\begin{bmatrix} \gamma_{1,k}\nu_{1,k} + (1 - \gamma_{1,k})\nu'_{1,k} \\ \gamma_{2,k}\nu_{2,k} + (1 - \gamma_{2,k})\nu'_{2,k} \\ \vdots \\ \gamma_{q,k}\nu_{q,k} + (1 - \gamma_{q,k})\nu'_{q,k} \end{bmatrix}}_{\tilde{\nu}_k} \quad (16)$$

where $\tilde{\nu}_k \sim \mathcal{N}(0, \tilde{R})$ with $\tilde{R} = \text{diag}\{\tilde{R}_{ii} = \gamma_{i,k}R_{ii} + (1 - \gamma_{i,k})R'_{ii}, i = 1, 2, \dots, q\}$. Note that the measurement equation becomes time-varying and stochastic in nature, due to the time-varying matrix $\tilde{C}_k$ (as defined in (16)) being a function of the random variables $\gamma_{i,k}$.

Following the Kalman filtering approach, we define,

$$\begin{aligned}
\hat{x}_{k|k} &\triangleq \mathbb{E}[x_k | y_0^k, u_0^k, \gamma_0^k] \\
P_{k|k} &\triangleq \mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T | y_0^k, u_0^k, \gamma_0^k] \\
\hat{x}_{k+1|k} &\triangleq \mathbb{E}[x_{k+1} | y_0^k, u_0^k, \gamma_0^k] \\
P_{k+1|k} &\triangleq \mathbb{E}[(x_{k+1} - \hat{x}_{k+1|k})(x_{k+1} - \hat{x}_{k+1|k})^T | y_0^k, u_0^k, \gamma_0^k]
\end{aligned} \tag{17}$$

where x_k , $\hat{x}_{k|k}$ and $\hat{x}_{k+1|k}$ represent the true state, the *a posteriori* and *a priori* state estimate; $\hat{P}_{k|k}$ and $\hat{P}_{k+1|k}$ denotes the *a posteriori* and *a priori* error covariance matrix; $\gamma_k = [\gamma_{1,k} \ \gamma_{2,k} \ \dots \ \gamma_{q,k}]^T$, $\gamma_0^k = \{\gamma_0, \dots, \gamma_k\}$, and $y_0^k = \{y_0, \dots, y_k\}$.

The Kalman filter algorithm based on the time-varying system can be re-derived as described per (15a) and (16). The prediction step of this KF based algorithm to compute $\hat{x}_{k+1}$ and $P_{k+1|k}$ uses the information from the state equation only, so it remains deterministic as in the classical Kalman filter:

$$\begin{aligned}
\hat{x}_{k+1|k} &= A\hat{x}_{k|k} + Bu_k + \varpi_k \\
P_{k+1|k} &= A\hat{P}_{k|k}A^T + Q
\end{aligned} \tag{18}$$

However, the correction step becomes stochastic due to its dependence on the observation process:

$$\begin{aligned}
\hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + K_{k+1}(y_{k+1} - \tilde{C}_{k+1}\hat{x}_{k+1|k}) \\
P_{k+1|k+1} &= P_{k+1|k} - K_{k+1}\tilde{C}_{k+1}P_{k+1|k}
\end{aligned} \tag{19}$$

where $K_{k+1} = P_{k+1|k}\tilde{C}_{k+1}^T(\tilde{C}_{k+1}P_{k+1|k}\tilde{C}_{k+1}^T + \tilde{R})^{-1}$ is the optimal Kalman gain computed by minimizing $P_{k+1|k+1}$.

Given the random set of γ_k at each sampling time k , it can be expected that there exists 2^q possible scenarios for measurement loss: specifically, the number of missing elements can range from 0 to q , with 0 as the measurement set being completely received and q indicating completely lost. The probability of each combination can be calculated from the joint probability function $\Pr(\gamma_{1,k}, \gamma_{2,k}, \dots, \gamma_{3,k})$. Here, we define a notation of $X_{\mathbb{Z}_1, \mathbb{Z}_2}^-$ as a sub-matrix of $X \in \mathbb{R}^{q_1 \times q_2}$ by deleting the rows of X as indexed in $\mathbb{Z}_1$ and the columns indexed in $\mathbb{Z}_2$, respectively. For an arbitrary sequence of measurements defined by $\mathbb{Z} = \{n_1, n_2, \dots\}$ or $\{\phi\}$, where $n_1, n_2, \dots \in \{1, 2, \dots, q\}$, the generalized formulation for the update equations is expressed as,

$$\begin{aligned}
\hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T (\{C_{\mathbb{Z},\phi}^-\} P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T + R_{\mathbb{Z},\mathbb{Z}}^-)^{-1} \\
&\quad (\{y_{k+1}\}_{\mathbb{Z},\phi}^- - \{C_{\mathbb{Z},\phi}^-\} \hat{x}_{k+1|k}) \\
P_{k+1|k+1} &= P_{k+1|k} - P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T (\{C_{\mathbb{Z},\phi}^-\} P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T + R_{\mathbb{Z},\mathbb{Z}}^-)^{-1} \\
&\quad \{C_{\mathbb{Z},\phi}^-\} P_{k+1|k}
\end{aligned} \tag{20}$$

Within the scope of this paper, our objective is to estimate EI underreporting from the measures of W and PA and estimated RMR , assuming negligible noise in PA and RMR . Accordingly, the system can be configured as: $x = [EI]^T$, $y = [GWG]^T$, $u = [PA \ RMR]$, where x , y and u are the state, output and input of the system. The state and measurement equations of the corresponding discrete time linear system are defined as follows:

$$x_{k+1} = Ax_k + Bu_k + \varpi_k \tag{21a}$$

$$y_k = C_k x_k + Du_k + \nu_k \tag{21b}$$

where k is the sampling time; ϖ_k and ν_k are the process and measurement noise of the system, respectively; the system matrices for the state space representation are $A = [1]$; $B = \mathbf{0}$; $C = [K_1]$; $D = [K_2 \ K_2]$. We assume that ϖ_k and ν_k are uncorrelated zero-mean Gaussian white noise with covariances of $Q = 0$ and $R > 0$, respectively, that is $\varpi_k \sim \mathcal{N}(0, Q)$ and $\nu_k \sim \mathcal{N}(0, R)$. In order to be able to use Kalman filter for state estimation, the model set up for the KF algorithm has to be observable, which can be determined by examining whether the rank of the observability matrix of the model $\mathcal{O} = [C^T A^T C^T \dots (A^T)^{(n-1)} C^T]$ equal to n , where n is the number of estimated states. It is easy to confirm that the rank of the observability matrix $\mathcal{O} = [K_1]$ equals 1, which corresponds to the number of the estimated system states. Thus, the model is fully observable and applicable for Kalman filtering. In addition, stabilizability is confirmed by checking (A, B) in case of no observation losses.

The detailed KF-based algorithm based on the system model in (21) is described below, where “hat” denotes the estimate.

- 1) **Initialize:** Set $\hat{x}(0|0) = EI_0$, $P(0|0) = I$. EI_0 is baseline measurements.
- 2) **Predict:** The prediction stage is independent of the update stage and can be expressed as:

$$\begin{aligned}
\hat{x}_{k+1|k} &= A \hat{x}_{k|k} + Bu_k + \varpi_k \\
P_{k+1|k} &= A \hat{P}_{k|k} A^T + Q
\end{aligned} \tag{22}$$

- 3) **Update:** The measurement update equation is the same as (20) and listed again below:

$$\begin{aligned}
\hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T (R_{\mathbb{Z},\mathbb{Z}}^- + \{C_{\mathbb{Z},\phi}^-\} P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T)^{-1} \\
&\quad (\{y_{k+1}\}_{\mathbb{Z},\phi}^- - \{C_{\mathbb{Z},\phi}^-\} \hat{x}_{k+1|k} - D_{\mathbb{Z},\phi}^- u_k) \\
P_{k+1|k+1} &= P_{k+1|k} - P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T (\{C_{\mathbb{Z},\phi}^-\} P_{k+1|k} \{C_{\mathbb{Z},\phi}^-\}^T + R_{\mathbb{Z},\mathbb{Z}}^-)^{-1} \\
&\quad \{C_{\mathbb{Z},\phi}^-\} P_{k+1|k}
\end{aligned} \tag{23}$$

where Q and R are the covariances for ω_k and v_k . Q and R can be used as adjustable parameters to influence the performance of the KF estimation algorithm.

Based on this approach, a real-time estimation of EI can be performed despite the presence of missing data. The results for the two representative participants used in back-calculation are shown in Fig.4, where the prediction bias is demonstrated with root mean square error ($RMSE$). For both participants, $Q = [100000]$ and $R = [0.5]$ are used for obtaining the presented results. The covariance matrix of the process noise, Q , corresponds to the fluctuation in the participant's EI . From examining the standard deviation of the back-calculated EI for multiple participants, the variation in the back-calculated EI is observed to be at the magnitude of 10^2 kcal. Thus, Q is selected to be 100000, which is approximately 300 kcal in terms of standard deviation. On the other hand, R , the covariance matrix of measurement noise that corresponds to the noise in measured weight, is selected based on the magnitude of the standard deviation for measured weights: 10^{-1} kg on average. Hence, R is determined to be 0.5.

It can be seen from the results in Fig.4 that state estimation of EI ‘adapts’ from given initial values and generally keeps above the self-reported EI despite the presence of noise and missing data. When missingness in the measured output GWG_{meas} occurs repeatedly, bias in EI_{est} is inevitably observed, since it is propagated from the previous measurement available. Compared with the noise corrupted GWG_{meas} , an accurate prediction of GWG is obtained from the algorithm as well. Underreporting of EI can be well identified from the estimation results. The mean $\pm$ SD of the EI estimate for participant A is 2581 ± 498 kcal; the mean $\pm$ SD for participant B is 2488 ± 488 kcal. The estimated EI is comparable to back-calculation approach while pre-processing of data to remove missingness or reduce noise is not required. A comparison between the back-calculation method and the Kalman filtering approach based on the estimation results for the two representative participants is tabulated in Table IV, where it can be seen that the back-tracked EI is similar for the two approaches.

Some extensions of the KF approach are elaborated below for the application of multiple state estimation during weight interventions; for example, when there exists a need to filter out the noise in PA and RMR measurements. To include more than one state in the systems model, we re-write our system equations in the form of (24) by reallocating the states to estimate and the state matrices. For two-state systems, EI remains as the first state to be estimated, and either RMR or PA can be used as the second state that needs to be estimated from noisy measurements. For the case of estimating EI and RMR from noise-corrupted measurement of GWG and RMR , the system can be written as: $x = [EI \ RMR]^T$, $y = [GWG$

$RMR]^T$, $u = [PA]$, where PA measurement can be treated as a noise-free input; For estimating EI and PA from uncertain measurement of GWG and PA and noise-free measurement of RMR , the system is written as $x = [EI \ PA]^T$, $y = [GWG \ PA]^T$, $u = [RMR]$. That is, one of the two signals of RMR and PA is used as a noise-corrupted output, while the other used as a noise-free input. For either of these two models, the system matrices can be expressed as:

$$\begin{aligned} A &= IB = \mathbf{0} \\ C &= \begin{bmatrix} K_1 & K_2 \\ 1 & 0 \end{bmatrix}, D = \begin{bmatrix} K_2 \\ 0 \end{bmatrix} \end{aligned} \quad (24)$$

The Kalman filtering approach can be applied to these alternative system models if necessary. An example of the two state estimation with the data from Participant B is presented below where the PA measurements are considered to be noise-corrupted and need to be estimated in addition to the EI . The results are shown in Fig. 5 where the selected covariance matrices are $Q = \begin{bmatrix} 100000 & 0 \\ 0 & 100000 \end{bmatrix}$, $R = \begin{bmatrix} 0.5 & 0 \\ 0 & 100000 \end{bmatrix}$. The mean $\pm SD$ of the EI estimate for participant B is 2323 ± 434 kcal, while the PA estimates are 165.5 ± 55 kcal. It has to be noted that for estimation of more than one state, convergence is a common issue if multiple variables are missing at the same time continuously, leading to instability. Since there is no missing data in PA measurements for this participant, instability is not a concern. Otherwise, data imputation such as linear interpolation can be employed to mitigate missingness. Similar analysis can be extended to three state estimation cases, for example, including both PA and RMR as states to be estimated. However, one needs to be careful about system observability when applying Kalman filtering in this case.

IV. Correction for Underreporting Through Semi-physical Identification

The approaches of back-calculation and Kalman filtering demonstrated in Section III are capable of providing estimates of EI based on the measures/estimation of W , PA and RMR . However, repeated data collection of W and PA is always necessary for these estimation approaches to be implemented. In this section, we aim to develop a model to correct future self-reported EI that contains potential misreporting, but without the need to require for intensive data collection of all measurements as before. Such models can describe the quantitative relationships between the actual energy intake (EI_{actual}) and the EI self-reports ($\widetilde{EI}_{rept}$), or other input variables if necessary, such as participant weight (W_{actual}). When a self-report is available, the model can be used to predict EI_{actual} . Fig. 6 depicts such relationships for modeling purposes, where EI_{actual} is a function of $\widetilde{EI}_{rept}$ and W_{actual} , denoted as

$$EI_{actual} = f(\widetilde{EI}_{rept}, W_{actual}) \quad (25)$$

Here the functional relationship f can be structured differently at the users' request. For simplicity, we only focus on linear or quadratic relationships in this paper. A semi-physical

identification approach based on linear regression from past collected data is used to obtain the models. Since EI self-reports or weight measurements are usually corrupted by noise, the data used as inputs are noise-corrupted signals: EI_{rept} and measured weight (W_{meas}). Accurately estimated $EI(EI_{est})$ from either back-calculation or Kalman filtering can be used to approximate EI_{actual} and to serve as regression outputs.

Once the model is identified, it can be used to adjust future self-reports. Cross-validation procedures are applied to test the models, and the resulting performance is relied on for selecting parsimonious yet accurate models with good predictive ability. The established model is useful to further understand the percentage of EI that is systematically under-reported, enabling health providers to deliver informative health counseling for participants. The effectiveness of this approach is illustrated with participant data evaluated on multiple model structures, demonstrating the ability of correcting biased EI_{rept} in the future.

A. Semi-Physical Estimation

A variety of model structures can be proposed to predict the actual EI from self-reported EI , that is, to correct the self-reported EI from misreporting. For example, a linear formula can be assumed to describe the relationship between EI_{actual} (model output) and $\widetilde{EI}_{rept}$ (model input); this follows according to (26).

$$EI_{actual}(k) = \alpha_1 \widetilde{EI}_{rept}(k) + \xi \quad (26)$$

For clarification, $\widetilde{EI}_{rept}$ represents the nominal values of self-reported EI without noise corruption. This linear relationship describes the deterministic portion of underreporting, which tends to be systematically observed in one's behaviors. For example, an individual may mistake a 320 kcal bagel to be 250 kcal, or repeatedly forgets to report calories from snacks. This consistent behavioral pattern is the target that the proposed relationship tries to capture and model. A challenging aspect of underreporting is associated with possible random variations in the EI self-reports. The effect of such variations can be treated as an input noise signal ($n_{EI_{rept}}$) added to $\widetilde{EI}_{rept}$:

$$EI_{rept}(k) = \widetilde{EI}_{rept}(k) + n_{EI_{rept}}(k) \quad (27)$$

where EI_{rept} is the reported EI from the smartphone app; $n_{EI_{rept}} \sim \mathcal{N}(0, \sigma_{n_{EI_{rept}}}^2)$ and $\sigma_{n_{EI_{rept}}}^2$ is the variance of the white noise $n_{EI_{rept}}$ in self-reports. To form the output of the regression problem, $EI_{actual}(k)$ can be approximated from the model-based back-calculation or Kalman filtering approaches described in Section III. For simplicity, $EI_{actual}(k)$ values computed directly from the EB model in (12) are used here, leading to,

$$EI_{actual}(k) \cong \frac{(GWG(k+1) - K_2(PA(k) + RMR(k)))}{K_1} \quad (28)$$

In the HMZ intervention study, all measurements/estimates are subject to noise, but the noise in GWG_{meas} (n_{GWG}) is relatively more predominant than in other signals, considering daily weight changes that result from the individuals' varying hydration status. Hence, n_{GWG} cannot be neglected and its presence corrupts the model-based estimates of EI_{actual} , leading to the expression of EI_{est} as,

$$\begin{aligned}
 EI_{est}(k) &= \frac{GWG_{meas}(k+1) - K_2(PA(k) + RMR(k))}{K_1} \\
 &= \frac{GWG(k+1) - K_2(PA(k) + RMR(k))}{K_1} + \frac{n_{GWG}(k+1)}{K_1} \\
 &= EI_{actual}(k) + n_{GWG}(k+1)/K_1 \\
 &= EI_{actual}(k) + n_{EI_{est}}
 \end{aligned} \tag{29}$$

where $n_{GWG} \sim \mathcal{N}(0, \sigma_{n_{GWG}}^2)$. Since the magnitude of the noise term $\frac{n_{GWG}(k)}{K_1}$ tends to be quite large compared to EI_{est} , smoothing techniques can be used to reduce variability in the estimation.

With the constructed input and output of the correction model, the model parameters α_1 and ξ in (26) can be estimated by solving a regression problem formulated based on measurements as shown below,

$$\begin{aligned}
 \mathcal{Z} &= \mathcal{R} \theta \\
 \begin{bmatrix} EI_{est}(k_1) \\ EI_{est}(k_2) \\ \vdots \\ EI_{est}(k_N) \end{bmatrix} &= \begin{bmatrix} EI_{rept}(k_1) & 1 \\ EI_{rept}(k_2) & 1 \\ \vdots & \vdots \\ EI_{rept}(k_N) & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \xi \end{bmatrix}
 \end{aligned} \tag{30}$$

where $\mathcal{Z} \in \mathbb{R}^N$ is the output vector based on EI_{est} obtained from (29); $\mathcal{R} \in \mathbb{R}^{N \times 2}$ is the regressor that stores input measurements; $\theta \in \mathbb{R}^2$ is the parameter vector that needs to be estimated; $k_1, k_2, \dots, k_N$ are the intermittent days at which the involved measurements are taken. Since EI_{rept} is not obtained daily (in order to minimize participant burden during intervention), the time index of the measurements involved in the regressor is not necessarily consecutive in terms of gestational age (days).

To estimate the parameter vector θ , a least squares cost function is considered:

$$J(k) = \min_{\hat{\theta}} \left\{ \frac{1}{2} [\mathcal{Z} - \mathcal{R}\hat{\theta}]^2 \right\} \tag{31}$$

The solution to (31) will give the estimates of θ , from which α_1 and ξ , the coefficients used to model the under-reported EI in (26) can be calculated. This allows us to estimate the actual EI from the EI_{rept} as shown below,

$$\hat{EI}_{est}(k) = \hat{\alpha}_1 EI_{rept}(k) + \hat{\xi} \quad (32)$$

where “hat” denotes the estimate. Besides the linear structure as shown in (26), other structures that directly relates the EI_{rept} with the output of regression model, EI_{est} , can be considered. These structures may involve different number of parameters or different polynomial orders. Table V summarizes all the evaluated structures in this paper for this approach. For each model structure, the regressor $\mathcal{R}$ and the estimator θ should change accordingly. As seen from this table, nonlinear aspects are incorporated by including quadratic terms with respect to EI_{rept} , while computational complexity is not elevated by maintaining a linear regression solution. It might be noted that, structure F and G use maternal weight as one of the system inputs. During pregnancy, intervention compliance might change as gestation advances to a later stage. Hence, a gestational time dependency or maternal weight dependency might be a potential factor to improve the prediction of women’s underreporting behaviors.

Models A to G in Table V involve increasing number of model parameters, as well as requiring additional pieces of information relevant to the EB model. For example, Model G contains four parameters to be identified, and once estimated, participant weight measurement is required in addition to the EI_{rept} . In comparison, Model A, with one parameter to be identified, needs no measurements. The comparison of how different structures perform in terms of their predictive ability will be discussed in the next section.

B. Results From Cross-Validation

In this section, the semi-physical approach is evaluated against participant data from the HMZ intervention study, and cross-validation techniques are used to test their performance. Cross-validation is a common test procedure in system identification to examine how accurately this predictive model performs on an independent data set. Different procedures for assigning data to estimation and validation sets can be used. Considering the characteristics of the data collected during gestation intervention, an interspersed way of data partitioning is applied by choosing one data point for estimation and every other point for validation. In this manner, the estimation and validation data sets each occupy half of the entire data set respectively, but are spread out uniformly over the intervention span.

For each model structure, estimation based on the assigned data set is performed followed by the validation on the remaining independent data set. To evaluate the performance of model prediction, multiple criteria are examined, including comparing the $\hat{EI}_{est}$ and analyzing the residual from regression. Specifically, the mean and SD of $\hat{EI}_{est}$ as well as the root mean square (RMS) of the residuals are used for analysis. Based on these evaluation criteria, the best model with a good fit can be selected while maintaining a parsimonious structure with minimum inputs. Other measures, such as the Akaike Information Theoretic Criterion (AIC) or Rissanen’s minimum description length (MDL) principle can also be considered but are not as critical in this case where a cross-validation data set is available.

In Figs. 7a and 7b, the results of $\hat{EI}_{est}$ calculated for the two representative study participants are presented based on the model structure C. For Participant A, the residuals remain random and stationary while an increasing drift in the residuals is observed for Participant B. This is caused by the substantial increase observed in the regression output EI_{est} towards late pregnancy due to the increasing rate of maternal weight change. While the increase in the regression output is observed, the regression input EI_{rept} does not reflect such trend of increase but remains stationary. This issue is also found among some of the other participants from the control group, for whom substantial maternal weight gains are observed along with a causal substantial increase in their dietary intake. If this is the case, the relationship between EI_{rept} and EI_{est} cannot be described as linear without introducing other variables. The literature also provides evidence of the time-varying characteristics of EI underreporting status across pregnancy: The level of under reporting was higher in late pregnancy in comparison to early pregnancy [17]. For participants with such characteristics, a time-dependent input such as gestational age or maternal weight, as shown in model structures F and G, will improve the estimates significantly. Fig. 7c shows the estimation results for Participant B using Model F, where the non-stationary trend in the residuals are successfully removed. However, this additional input does not change the results as much for the participants for whom such discrepancy in the increase of EI_{est} and EI_{rept} is not observed.

The estimated results for the two participants using different model structures are tabulated in Table VI. From analyzing the $\hat{EI}_{est}$ time series and the RMS of the residuals (on estimation and validation data sets), the best model for each individual participant can be selected. It should be noted that moderate variation in $\hat{EI}_{est}$ is preferred, as opposed to stationary/"static" estimates. Among all the examined model structures, Model A is the most parsimonious but with the most stationary estimates, while Model B gives the most variable yet least reliable estimates. Therefore, Model C to G are among the best informative models, with a fair balance between the residual RMS and the number of parameters. Estimates from Model C and E show similar RMS magnitudes, but Model C only involves two parameters instead of Model G with three. Hence, Model C is preferred over E. Similarly, Model F produces comparable RMS as Model G but using less parameters, so Model F is preferable to G. This analysis concludes that Model C and F are the best two options for the majority of participants, without overparametrizing the model structures. Depending on the data characteristics for individual participants, these two models can be selected one over the other based on whichever minimizes the averaged RMS . It can be concluded from the results that when there is no substantial increase in the EI_{est} , the 1st order model with two parameters and the 2nd order model with three parameters provide the best (and comparable) estimation results. When substantial increases in the estimates are observed as a result of dramatic gestational weight gain (e.g., Participant B), the dynamics in the energy intake cannot be fully captured with correction models that are only dependent on EI_{rept} . Augmentation of a time-dependent variable in the models, e.g., gestational age or weight, will significantly improve the predictive performance as shown in the residual analysis. It is important to note that the input noise (as shown in Fig. 6) poses a challenge for this

estimation problem due to the errors-in-variables problem [35], [36]. This issue is part of current research and will be elaborated in the next section.

C. Prediction Error Analysis For Semi-physical Estimation

Semi-physical identification approaches can estimate the extent of systematic underreporting and can be used for the prediction and correction of individuals' underreporting. The method requires EI_{rept} leading to point-wise estimates available only at the days with EI self-reports. From the structure of the regression model in Fig. 6, it can be seen that both the input and output signals are corrupted by noise. Traditional system identification considers only noise in the output, treating the input signals as perfectly known and noiseless. The model that we are trying to identify here, however, corresponds to an errors-in-variables (EIV) problem with both uncertain output and input, the noise in the input being unknown and unneglectable.

This estimation can be further understood with the prediction error analysis in the frequency domain. For the example of Model C, the prediction error in time domain can be written as,

$$\begin{aligned} e_{pred}(t) &= \frac{GWG_{meas}(t) - K_2(PA(t) + RMR(t))}{K_1} - \hat{EI}_{est}(t) \Big| t-1 \\ &= (\theta_1 - \hat{\theta}_1)\widetilde{EI}_{rept}(t) + (\theta_2 - \hat{\theta}_2) - \hat{\theta}_1 n_{EI_{rept}}(t) + \frac{n_{GWG}(t)}{K_1} \end{aligned} \quad (33)$$

where the estimated parameters α_1 and ξ are represented as θ_1 and θ_2 for simplicity. From Parseval's theorem, the following relationship between the variance of the prediction error and its power spectrum can be obtained.

$$\lim_{N \rightarrow \infty} \sum_{i=0}^N e_{pred}^2(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_e(\omega) d\omega \quad (34)$$

Suppose $\widetilde{EI}_{rept} \sim \mathcal{N}(\mu_{EI_{rept}}, \sigma_{n_{\widetilde{EI}_{rept}}}^2)$ by assuming $\widetilde{EI}_{rept}$ being a stationary signal composed of a mean ($\mu_{\widetilde{EI}_{rept}}$) and an additive white noise ($n_{\widetilde{EI}_{rept}}$) with variances $\sigma_{n_{\widetilde{EI}_{rept}}}^2$. $\Phi_e(\omega)$ can then be derived as,

$$\begin{aligned}
\Phi_e(\omega) &= \left| \theta_1 - \hat{\theta}_1 \right|^2 \Phi_{\widetilde{EI}_{rept}}(\omega) + \left| \hat{\theta}_1 \right|^2 \frac{\sigma_{nEI_{rept}}^2}{\Phi_{\widetilde{EI}_{rept}}(\omega)} + \left| \theta_2 - \hat{\theta}_2 \right|^2 + \frac{\sigma_{nGWG}^2}{K_1^2} + 2\text{Re} \left\{ \right. \\
&\quad \left. (\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2)^* \mu_{\widetilde{EI}_{rept}} \right\} \\
&= \Phi_{\widetilde{EI}_{rept}}(\omega) \left(\left| \theta_1 - \hat{\theta}_1 \right|^2 + \left| \hat{\theta}_1 \right|^2 \frac{\sigma_{nEI_{rept}}^2}{\Phi_{\widetilde{EI}_{rept}}(\omega)} \right) + \left| \theta_2 - \hat{\theta}_2 \right|^2 + \frac{\sigma_{nGWG}^2}{K_1^2} + 2\text{Re} \left\{ \right. \\
&\quad \left. (\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2)^* \mu_{\widetilde{EI}_{rept}} \right\}
\end{aligned} \tag{35}$$

where $\Phi_{\widetilde{EI}_{rept}}(\omega) = \mu_{\widetilde{EI}_{rept}}^2 + \sigma_{n\widetilde{EI}_{rept}}^2$. The detailed derivation for (35) is presented in the

Appendix. As shown in (35), the estimated parameters are affected by multiple coefficients that are determined by the mean and variance of the signals involved. In the case of the $\widetilde{EI}_{rept}$ signal being zero mean, that is, when $\mu_{\widetilde{EI}_{rept}} = 0$, the expression of $\Phi_e(\omega)$ simplifies to,

$$\Phi_e(\omega) = \sigma_{n\widetilde{EI}_{rept}}^2(\omega) \left(\left| \theta_1 - \hat{\theta}_1 \right|^2 + \left| \hat{\theta}_1 \right|^2 \frac{\sigma_{nEI_{rept}}^2}{\sigma_{n\widetilde{EI}_{rept}}^2} \right) + \left| \theta_2 - \hat{\theta}_2 \right|^2 + \frac{\sigma_{nGWG}^2}{K_1^2} \tag{36}$$

which shows that the presence of bias in θ_1 is influenced by the ratio of $\frac{\sigma_{nEI_{rept}}^2}{\sigma_{n\widetilde{EI}_{rept}}^2}$, while the

estimates of θ_2 should be unbiased. When $\frac{\sigma_{nEI_{rept}}^2}{\sigma_{n\widetilde{EI}_{rept}}^2} \gg 1$, the bias in the estimate of θ_1 will

be negligible. For the case of $\widetilde{EI}_{rept}$ with non-zero mean (which corresponds to real life conditions), bias in the estimation of both parameters can be observed and is affected by the magnitude of the mean and the variance of the signals involved. Despite knowing the fact that removing the mean in EI_{rept} can reduce the bias in θ_2 , the strategy of mean removal is difficult to implement in real data, because most of the model structures proposed in this paper contain a constant term, ξ . Removing the mean of $\widetilde{EI}_{rept}$ results in the removal of one of the to-be-estimated parameters.

In support of this error analysis, simulations with hypothetical yet representative data are run under the condition of 10^7 samples to test against such observations in asymptotic condition. Stationary $\bar{EI}_{rept}$ signals are generated with mean $\mu_{\bar{EI}_{rept}}$ and additive white noise signals

$n_{\bar{EI}_{rept}}$ with variances of $\sigma_{n_{\bar{EI}_{rept}}}^2$. Four different scenarios are created by adjusting the ratio

of $\frac{\sigma_{n_{EI_{rept}}}^2}{\sigma_{n_{\bar{EI}_{rept}}}^2}$ and $\mu_{\bar{EI}_{rept}}$. Specifically, two representative $\mu_{\bar{EI}_{rept}}$ are selected to simulate the

real-life conditions: 2500 kcal and 4500 respectively; the ratio of $\frac{\sigma_{n_{EI_{rept}}}^2}{\sigma_{n_{\bar{EI}_{rept}}}^2}$ are manipulated to

be 1/9 or 9/1 with $\sigma_{n_{\bar{EI}_{rept}}}^2$ to be 100 or 900. The estimated parameters under different conditions are tabulated in Tables VII.

It is shown from the simulated cases that bias in the estimation of parameter θ_1 always

exists, but the extent of the bias can be manipulated by the ratio of $\frac{\sigma_{n_{EI_{rept}}}^2}{\sigma_{n_{\bar{EI}_{rept}}}^2}$. When the ratio

decreases to 1/9, bias in θ_1 is negligible. On the other hand, bias in the estimation of the parameter θ_2 is readily observed and affected by the magnitude of the mean and the variance of the signals involved, largely dependent on the magnitude of the mean of $\bar{EI}_{rept}$. The bias in θ_2 decreases with lower $\mu_{\bar{EI}_{rept}}$, while increasing $\mu_{\bar{EI}_{rept}}$ makes the bias in θ_2 more significant. This is also consistent with the results of the prediction error analysis. Even though bias can be observed in both estimated parameters θ_1 and θ_2 (regardless of the ratio of the noise variances), the mean of the estimated EI remains unbiased, indicating that the $\hat{EI}_{est}$ obtained from this semi-physical approach is reliable. The standard deviation of the estimates will vary and is dependent on the ratio of the variances, but as long as the ratio is modest, the variability in the estimates is close to the true signals.

Considering the EIV problem inherent to this system, pursuing approaches developed for EIV model identification is appealing. Unfortunately, the traditional maximum-likelihood approach (or a maximum-likelihood approach with Gaussian latent variables) is not effective for solving the proposed EI correction model, as large errors will result from the estimation [37]. The problem lies in that the relationship between intake and gestational weight gain is modeled as instantaneous, leading to one effective measurement for each unknown value; for any additional data point, there is an additional parameter to be estimated. As noted in [35], [36], such static errors-invariables models are among the most difficult to solve. Other approaches, such as Total Least Squares, require the noise variance ratio between the input and output measurements to be known *a priori*, or the availability of multiple experiments

from the same participant; these are challenging experimental conditions that are not experienced in our study.

V. Summary and Conclusions

Energy intake underreporting is a widespread problem for interventions relying on self-reported measures of EI ; this is reflected in the experience of the authors in the *Healthy Mom Zone* intervention. To address this issue in the context of a gestational weight control intervention, three model-based estimation approaches are developed in this paper and tested against participant data.

The first approach of back-calculating EI from the EB model is the easiest to implement and understand among the three, and may be favored by users seeking low computational complexity. However, it requires *a priori* data smoothing and imputation or interpolation approaches to address data missingness. The second approach based on Kalman filtering enables real-time estimation of EI in the presence of missing data, without demanding *a priori* smoothing. It can be interpreted as a refined way of back-calculating EI , which improves the model-predicted estimates by filtering out the noise in the intermittent weight measurements. It must be noted that the estimation results will depend on the values of the covariance parameters, Q and R , which may be viewed as adjustable parameters. In this paper, the conventional approach for specifying Q and R is used for simplicity. However, algorithms that can adaptively adjust Q and R may be used to improve filtering performance; such methods have been explored in [38]. Alternatively, a time-varying Kalman filter that includes an auxiliary state estimator (so-called double Kalman filtering) is also a possible option to adjust noise covariance based on the change in the system dynamics [39]. Yet, the performance of these algorithms under intermittent measurements remains unclear and needs to be examined; this remains a topic for future research.

The third approach developed based on the semi-physical approach features the ability to correct future EI self-reports by parametrizing the extent of EI underreporting. This approach was illustrated on a variety of model structures that are characterized by different inputs or polynomial orders. In contrast with the first two approaches, intensive measurements of the EB variables (W , PA and RMR) may not be necessary once a correction model is estimated. With this being said, less information from participant measurements is required to realize future estimation. The accuracy of this approach suffers from the intrinsic errors-in-variables problem due to input noise, but a simulation study showed that when the variance of the input noise is modest, this semi-physical approach is still accurate and reliable in terms of estimating the EI . As a future research direction, it may be worthwhile to explore the contributions of certain maternal psychological factors to the underreporting correction models by using the longitudinal measurements of maternal depression, anxiety, visual perception of body image in pregnancy as model inputs.

In conclusion, we have applied a diverse set of approaches that feature varying levels of complexity, novelty, and usefulness. Each approach has pros and cons, and features advantages over the other approaches based on user requirements or data characteristics. Decisions of which approach to be used need to be made carefully considering different

circumstances. Ultimately, approaches presented in this work have played (and will continue to play) an important role in *Healthy Mom Zone*, and related weight control interventions resulting from this research that require judicious determination of energy intake.

Acknowledgment

Support for this work has been provided by the National Heart, Lung, and Blood Institute (NHLBI) through grant R01 HL119212. The opinions expressed in this article are the authors' own and do not necessarily reflect the views of NHLBI.

Appendix

In this section, the derivation for the power spectrum of the prediction error ($\Phi_e(\omega)$) in (35) is presented. Given (33), the covariance of the prediction error can be derived as,

$$\begin{aligned}
 R_{e_{pred}} &= \bar{E}\{e_{pred}(t)^2\} \\
 &= \bar{E}\left\{\left((\theta_1 - \hat{\theta}_1)\bar{EI}_{rept}(t) + (\theta_2 - \hat{\theta}_2) - \hat{\theta}_1 n_{EI_{rept}}(t) + \frac{n_{GWG}(t)}{K_1}\right)^2\right\} \\
 &= \bar{E}\left\{(\theta_1 - \hat{\theta}_1)^2 EI_{rept}^2(t) + 2(\theta_2 - \hat{\theta}_2) \frac{n_{GWG}(t)}{K_1} + 2(\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2)\bar{EI}_{rept}(t) \right. \\
 &\quad \left. + \hat{\theta}_1^2 n_{EI_{rept}}^2(t) - 2\hat{\theta}_1(\theta_1 - \hat{\theta}_1)\bar{EI}_{rept}(t)n_{EI_{rept}}(t) + (\theta_2 - \hat{\theta}_2)^2 + 2(\theta_1 - \hat{\theta}_1) \right. \\
 &\quad \left. \bar{EI}_{rept}(t) \frac{n_{GWG}(t)}{K_1} + \frac{n_{GWG}^2(t)}{K_1^2} - 2\hat{\theta}_1 n_{EI_{rept}}(t) + \frac{n_{GWG}(t)}{K_1} - 2\hat{\theta}_1(\theta_2 - \hat{\theta}_2) \right. \\
 &\quad \left. n_{EI_{rept}}(t)\right\} \tag{37}
 \end{aligned}$$

from which the power spectrum analysis of the prediction error can be derived below,

$$\begin{aligned}
\Phi_e(\omega) &= \left| \theta_1 - \hat{\theta}_1 \right|^2 \Phi_{\widetilde{EI}_{rept}}(\omega) + \left| \theta_2 - \hat{\theta}_2 \right|^2 + \left| \hat{\theta}_1 \right|^2 \sigma_{n_{EI_{rept}}}^2 + \frac{\sigma_{n_{GWG}}^2}{K_1^2} + 2Re \\
&\quad \left\{ (\theta_1 - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2)^* \mu_{EI_{rept}} \right\} + 2Re \left\{ (\theta_1 - \hat{\theta}_1)(-\hat{\theta}_1)^* \mu_{\widetilde{EI}_{rept}} \mu_{n_{EI_{rept}}} \right\} + 2Re \\
&\quad \left\{ (\theta_1 - \hat{\theta}_1) \mu_{\widetilde{EI}_{rept}} \mu_{n_{GWG}} / K_1 \right\} + 2Re \left\{ (\theta_2 - \hat{\theta}_2) \mu_{n_{GWG}} / K_1 + 2Re \left\{ - \right. \right. \\
&\quad \left. \left. \hat{\theta}_1 \mu_{n_{EI_{rept}}} \mu_{n_{GWG}} / K_1 \right\} + 2Re \left\{ (\theta_2 - \hat{\theta}_2)(-\hat{\theta}_1)^* \mu_{n_{EI_{rept}}} \right\} \right. \\
&= \left| \theta_1 - \hat{\theta}_1 \right|^2 \Phi_{\widetilde{EI}_{rept}}(\omega) + \left| \theta_2 - \hat{\theta}_2 \right|^2 + \left| \hat{\theta}_1 \right|^2 \sigma_{n_{EI_{rept}}}^2 + \frac{\sigma_{n_{GWG}}^2}{K_1^2} + 2Re \left\{ (\theta_1 \right. \\
&\quad \left. - \hat{\theta}_1)(\theta_2 - \hat{\theta}_2)^* \mu_{\widetilde{EI}_{rept}} \right\}
\end{aligned} \tag{38}$$

This is the final expression as shown in (35) under the assumption that all the signals are uncorrelated with noise terms being zero mean, and θ_1 and θ_2 are independently parameterized. Note that $n_{\widetilde{EI}_{rept}}$ indicates the variations in the noiseless signal of $\widetilde{EI}_{rept}$, while $n_{EI_{rept}}$ corresponds to the noise and random errors in EI_{rept} . The variance of the prior is expressed with $\sigma_{n_{\widetilde{EI}_{rept}}}^2$ and the latter with $\sigma_{n_{EI_{rept}}}^2$.

Biography

Penghong Guo received her B.E. degree in Pharmaceutical Engineering from Liaoning University, China and her M.S. degree in Chemical Engineering from Carnegie Mellon University, USA, in 2012 and 2013, respectively. She is currently a Chemical Engineering Ph.D. candidate at Arizona State University. Her research interests include applying engineering process control concepts to the analysis, design, and implementation of adaptive, time-varying interventions in the field of behavioral medicine. In addition to optimized interventions for prevention and treatment in behavioral medicine, she also maintains an ongoing interest in the topics of chemical process control and supply chain management.

Daniel E. Rivera (M'91-SM'05) received the B.S. degree in chemical engineering from the University of Rochester, New York in 1982, the M.S. degree in chemical engineering from the University of Wisconsin-Madison in 1984, and the Ph.D in chemical engineering from the California Institute of Technology, Pasadena, California in 1987.

He is a Professor of chemical engineering in the School for Engineering of Matter, Transport, and Energy at Arizona State University in Tempe, Arizona. Prior to joining ASU he was a member of the Control Systems Section of Shell Development Company in Houston, Texas. His research interests span the topics of system identification, robust process control, and applications of control engineering to problems in supply chain management and behavioral medicine.

Since 2014, Dr. Rivera has chaired the IEEE Control System Society's (CSS's) Outreach Task Force. He was the inaugural chair of the IEEE CSS technical committee on healthcare and medical systems (2013 – 2014) and has also chaired the IEEE CSS technical committee on system identification and adaptive control (2007 – 2012). He is past associate editor for the *IEEE Transactions in Control Systems Technology* (2003 – 2010) and the *IEEE Control*

Systems Magazine (2003 – 2007). In 2007 he was awarded a K25 Mentored Quantitative Research Career Development Award from the National Institutes of Health to study the application of control systems engineering principles to optimize personalized prevention and treatment interventions for drug abuse.

Jennifer Savage Williams received a Ph.D. from The Pennsylvania State University in Nutritional Sciences in 2008. She is the Director for the Center of Childhood Obesity Research at The Pennsylvania State University and also an Assistant Professor within the Department of Nutritional Sciences. Her research focuses on developing effective behavioral interventions to prevent maternal and childhood obesity. Dr. Savage's research spans the spectrum of translation from observational studies to efficiency trials. Her recent work uses principles of system science to develop individualized, adaptive prenatal and postnatal interventions to promote health across the lifespan. Dr. Savage has conducted and participated in NIH, HRSA & USDA funded research in the area of childhood obesity for over 10 years and coauthored about 65 papers in international journals and conference proceedings.

Emily E. Hohman received a B.S. in nutritional sciences from Cornell University, Ithaca, NY, USA in 2008 and a PhD in nutrition science from Purdue University, West Lafayette, IN, USA in 2014.

She is currently an Assistant Research Professor at the Center for Childhood Obesity Research, Pennsylvania State University, University Park, PA, USA. Her research interests focus on the interactions between diet and behavioral factors on growth and chronic disease risk in infants, children, adolescents, and pregnant women.

Dr. Hohman is a member of the American Society for Nutrition and the Obesity Society. She was a recipient of the NIH Matilda White Riley Early Career Investigator paper award in 2017.

Abigail M. Pauley received the B.S. degree in Exercise Science from Shenandoah University and the M.S. degree in Exercise Science from Blooms-burg University in 2012 and 2014, respectively. She is currently a Kinesiology Ph.D. candidate at The Pennsylvania State University. Her current research interests include promoting exercise and positive sleep behaviors during pregnancy as well as the relationships between exercise, sleep and psychosocial factors during pregnancy.

Krista S. Leonard received her B.A. degree in exercise science and psychology from Willamette University and her M.S. degree in Kinesiology from The Pennsylvania State University in 2015 and 2017, respectively. She is currently a Kinesiology Ph.D. candidate at The Pennsylvania State University. Her research interests include promoting exercise behavior during and after pregnancy and the use of mobile health tools to promote healthy lifestyle behavior change.

Danielle Symons Downs received a Ph.D. from The University of Florida in Health and Human Performance with a specialization in Exercise/Sport Psychology in 2002, and a Masters in Applied Behavior Change Psychology from the State University of New York at Brockport in 1998. She is a Professor of Kinesiology and Obstetrics and Gynecology, Director of the Exercise Psychology Laboratory in the Department of Kinesiology, College of Health and Human Development, and Associate Director of the Social Science Research Institute at The Pennsylvania State University. Dr. Downs' research expertise is in understanding motivational determinants of exercise and designing behavioral interventions to promote exercise/health correlates. Her recent interventions have promoted exercise for managing gestational diabetes, used mHealth technology to improve women's health before

pregnancy, and developed an individually-tailored, adaptive intervention to manage gestational weight gain in overweight/obese pregnant women. Dr. Downs has published over 100 papers in international journals and conference proceedings.

References

- [1]. Ogden CL, Carroll MD, Kit BK et al., "Prevalence of childhood and adult obesity in the United States, 2011–2012," JAMA, vol. 311, no. 8, pp. 806–814, 2014. [PubMed: 24570244]
- [2]. Rasmussen KM and Yaktine AL, Eds., Weight Gain During Pregnancy: Reexamining The Guidelines. National Academies Press, 2009.
- [3]. Haugen M et al., "Associations of pre-pregnancy body mass index and gestational weight gain with pregnancy outcome and postpartum weight retention: a prospective observational cohort study," BMC Pregnancy and Childbirth, vol. 14, no. 1, p. 201, 2014. [PubMed: 24917037]
- [4]. Chu SY, Callaghan WM, Kim SY, Schmid CH, Lau J, England LJ, and Dietz PM, "Maternal obesity and risk of gestational diabetes mellitus." Diabetes Care, vol. 30, no. 8, pp. 2070–2076, 2007. [PubMed: 17416786]
- [5]. Carreno CA et al., "Excessive early gestational weight gain and risk of gestational diabetes mellitus in nulliparous women," Obstetrics and Gynecology, vol. 119, no. 6, pp. 1227–1233, 2012. [PubMed: 22617588]
- [6]. Nehring I, Schmolli S, Beyerlein A, Hauner H, and Von Kries R, "Gestational weight gain and long-term postpartum weight retention: a meta-analysis," The American Journal of Clinical Nutrition, vol. 94, no. 5, pp. 1225–1231, 2011. [PubMed: 21918221]
- [7]. Gilmore LA, Klempel-Donchenko M, and Redman LM, "Pregnancy as a window to future health: Excessive gestational weight gain and obesity," in Seminars in Perinatology, vol. 39, no. 4, 2015, pp. 296–303. [PubMed: 26096078]
- [8]. Schack-Nielsen L, Michaelsen KF, Gamborg M, Mortensen EL, and Sørensen TI, "Gestational weight gain in relation to offspring body mass index and obesity from infancy through adulthood," International Journal of Obesity, vol. 34, no. 1, pp. 67–74, 2010. [PubMed: 19918246]
- [9]. Symons Downs D, Savage JS, Rivera DE, Smyth JM, Rolls BJ, Hohman EE, McNitt KM, Kunselman AR, Stetter C, Pauley AM, Leonard KS, and Guo P, "Individually tailored, adaptive intervention to manage gestational weight gain: protocol for a randomized controlled trial in women with overweight and obesity," JMIR Res Protoc, vol. 7, no. 6, pp. e150, DOI: 10.2196/resprot.9220, 2018. [PubMed: 29884603]
- [10]. Thomas DM, Navarro-Barrientos JE, Rivera DE et al., "Dynamic energy balance model predicting gestational weight gain," The American Journal of Clinical Nutrition, vol. 95, no. 1, pp. 115–122, 2012. [PubMed: 22170365]
- [11]. Guo P, Symons Downs D, and Savage JS, "Semi-physical identification and state estimation of energy intake for interventions to manage gestational weight gain," in Proceedings of 2016 American Control Conference (ACC), 2016, pp. 1271–1276.
- [12]. Johansson L, Solvoll K, Bjørneboe GE, and D. CA, "Under-and overreporting of energy intake related to weight status and lifestyle in a nationwide sample," The American Journal of Clinical Nutrition, vol. 68, no. 2, pp. 266–74, 1998. [PubMed: 9701182]
- [13]. Poslusna K, Ruprich J, de Vries JH, Jakubikova M, and van't Veer P, "Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice," British Journal of Nutrition, vol. 101, no. S2, pp. S73–85, 2009. [PubMed: 19594967]
- [14]. Lichtman SW et al., "Discrepancy between self-reported and actual caloric intake and exercise in obese subjects," New England Journal of Medicine, vol. 327, no. 27, pp. 1893–1898, 1992. [PubMed: 1454084]
- [15]. McGowan CA and McAuliffe FM, "Maternal nutrient intakes and levels of energy underreporting during early pregnancy," European Journal of Clinical Nutrition, vol. 66, no. 8, pp. 906–913, 2012. [PubMed: 22378227]

- [16]. Trabulsi J and Schoeller DA, "Evaluation of dietary assessment instruments against doubly labeled water, a biomarker of habitual energy intake," *American Journal of Physiology-Endocrinology And Metabolism*, vol. 281(5), no. 5, pp. E891–E899, 2001. [PubMed: 11595643]
- [17]. Moran LJ, McNaughton SA, Sui Z, Cramp C, Deussen AR, Grivell RM, and Dodd JM, "The characterisation of overweight and obese women who are under reporting energy intake during pregnancy," *BMC Pregnancy and Childbirth*, vol. 18, no. 204, 2018.
- [18]. Nowicki E, Siega-Riz AM, Herring A, He K, Stuebe A, and Olshan A, "Predictors of measurement error in energy intake during pregnancy," *American Journal of Epidemiology*, vol. 173, no. 5, pp. 560–568, 2011. [PubMed: 21273398]
- [19]. Mullaney L, O'higgins AC, Cawley S, Doolan A, McCartney D, and Turner MJ, "An estimation of periconceptional under-reporting of dietary energy intake," *Journal of Public Health*, vol. 37, no. 4, pp. 728–736, 2014. [PubMed: 25355686]
- [20]. Poslusna K, Ruprich J, de Vries JH, Jakubikova M, and van't Veer P, "Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice," *British Journal of Nutrition*, vol. 101, no. S2, pp. S73–85, 2009. [PubMed: 19594967]
- [21]. Lutomski JE, van den Broeck J, Harrington J, Shiely F, and Perry IJ, "Sociodemographic, lifestyle, mental health and dietary factors associated with direction of misreporting of energy intake," *Public Health Nutrition*, vol. 14, no. 3, pp. 532–541, 2011. [PubMed: 20707944]
- [22]. Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein and Amino Acids. National Academies Press, 2005.
- [23]. Dong Y, Rivera DE, Symons Downs D, Savage JS, Thomas DM, and Collins LM, "Hybrid model predictive control for optimizing gestational weight gain behavioral interventions," in *Proceedings of 2013 American Control Conference (ACC)*, 2013, pp. 1970–1975.
- [24]. Rivera DE, Hekler EB, Savage JS, and Symons Downs D, "Intensively adaptive interventions using control systems engineering: Two illustrative examples," in *Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics*, Collins LM and Kugler KC, Eds. Springer, in press.
- [25]. Hall KD and Chow CC, "Estimating changes in free-living energy intake and its confidence interval," *The American Journal of Clinical Nutrition*, vol. 94, no. 1, pp. 66–74, 2011. [PubMed: 21562087]
- [26]. Guo P, Rivera DE, Symons Downs D, and Savage JS, "State estimation under correlated partial measurement losses: Implications for weight control interventions," in *Proceedings of 20th World Congress, The International Federation of Automatic Control (IFAC)*, 2017, pp. 14 074–14 079.
- [27]. Ravussin E and Bogardus C, "A brief overview of human energy metabolism and its relationship to essential obesity," *The American Journal of Clinical Nutrition*, vol. 55, no. 1, pp. 242S–245S, 1992. [PubMed: 1728837]
- [28]. Westerterp KR, "Diet induced thermogenesis," *Nutrition & Metabolism*, vol. 1, no. 1, p. 5, 2004. [PubMed: 15507147]
- [29]. Piers LS, Diggavi SN, Thangam S, Raaij JMV, Shetty PS, and Hautvast JG, "Changes in energy expenditure, anthropometry, and energy intake during the course of pregnancy and lactation in well-nourished indian women," *The American Journal of Clinical Nutrition*, vol. 61, no. 3, pp. 501–513, 1995. [PubMed: 7872213]
- [30]. Thomas DM, "A dynamical fetal-maternal model of gestational weight gain," *Personal Communication*, 2009.
- [31]. Butte NF, Wong WW, Treuth MS et al., "Energy requirements during pregnancy based on total energy expenditure and energy deposition," *The American Journal of Clinical Nutrition*, vol. 79, no. 6, pp. 1078–1087, 2004. [PubMed: 15159239]
- [32]. Willett W, *Nutritional Epidemiology*. Oxford University Press New York, 1990.
- [33]. Sinopoli B, Schenato L, Franceschetti M et al., "Kalman filtering with intermittent observations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1453–1464, 2004.
- [34]. Liu X and Goldsmith A, "Kalman filtering with partial observation losses," in *43rd IEEE Conference on Decision and Control (CDC)*, vol. 4, 2004, pp. 4180–4186.

- [35]. Söderström T, “System identification for the errors-in-variables problem,” Transactions of the Institute of Measurement and Control, vol. 34(7), no. 7, pp. 780–792, 2012.
- [36]. Söderström T, Error-in-Variables Methods in System Identification. Springer International Publishing, 2018.
- [37]. Risuleo RS (KTH Royal Institute of Technology, Sweden), Personal Communication, 2017.
- [38]. Akhlaghi S, Zhou N, and Huang Z, “Adaptive adjustment of noise covariance in Kalman filter for dynamic state estimation,” Submitted to arXiv, 2017.
- [39]. Johansen TA and Fossen TI, “Nonlinear filtering with exogenous Kalman filter and double Kalman filter,” in 2016 European Control Conference (ECC), 2016, pp. 1722–1727.

Fig. 1:

Weight predictions from the energy balance model according to (1) using data from two representative participants in the *Healthy Mom Zone* intervention study. These show evidence of significant underreporting of energy intake. Participant A is an OW woman from the intervention group and Participant B (OW) from the control group. The self-reported EI are obtained from a smartphone app (MyFitnessPal) and PA is objectively monitored with a wrist-worn accelerometer (Jawbone). Additional information on data collection is described in Section II.B.

Fig. 2:

Block diagram depicting a closed-loop intervention for gestational weight gain, and how estimation approaches to energy intake as developed in this paper (indicated in the blue box) can be incorporated within the system. Energy intake estimates as well as the filtered weight measurements resulting from these estimators can be used by a hybrid model predictive control (HMPC) algorithm to determine optimized intervention dosages of intervention components (such as healthy eating active learning, physical activity active learning, goal setting), as described in [23], [24].

(a) Participant A

(b) Participant B

Fig. 3:

EI back-calculation results for the two representative participants from the HMZ intervention study. Comparing the measured W (circles) with the simulated W (dashed) using back-calculated EI , it provides support for the validity of the EI estimates. BMI: body mass index; GA: gestational age.

(a) Participant A

(b) Participant B

Fig. 4:

Results of estimating the EI using Kalman filtering for two HMZ participants. The results indicate that underreporting of EI can be identified for most of the time; however, estimation accuracy is compromised when the weight measurement is missing for multiple consecutive days. The prediction bias is indicated with root mean square error ($RMSE$). Vertical black lines in the GWG plot indicate the days of missing GWG measurements.

Fig. 5:

Kalman filtering performance for Participant B for two-state estimation. In this case, the estimation of the *EI* is implemented simultaneously with noise filtering for the *PA* measurements. Vertical black lines in the *GWG* plot indicate the days of missing *GWG* measurements. *RMSE* stands for root mean square error.

Fig. 6:

Block diagram of the regression model used for the development of the semi-physical identification approach. n_W , $n_{EI_{rept}}$ and $n_{EI_{est}}$ indicate the noise in the measured W , self-reported EI , and estimated EI , respectively.

(a) Participant A, Model C

(b) Participant B, Model C

(c) Participant B, Model F

Fig. 7:

Results of semi-physical identification of EI for two HMZ participants. (a) Estimation results based on Model C for an intervention Participant A on validation data set only. (b) Results for a control Participant B based on Model C (residual demonstrating non-stationarity). (c) Results based on Model F for the control Participant B (non-stationary trend in the residuals removed).

TABLE I:

Tabulation of the coefficients for the linear functions of total body water (TBW) and total body protein (TBP) with respect to maternal weight (W).

BMI Category (kg/m ²)	Weight Range (kg)	a_w	b_w	a_P	b_P
Low BMI (< 19.8)	$W \leq 52$	0.489	3.875	-0.04762	9.28
	$52 < W \leq 57.7$			0.105263	1.33
	$W > 57.7$			0.075472	3.05
Normal BMI (19.8–26)	$W \leq 60.2$	0.4836	2.853	-0.667	47.533
	$60.2 < W \leq 65.1$			0.0204	6.17
	$W > 65.1$			0.0724	3.05
High BMI (> 26)	$W \leq 81.8$	0.503	4.885	-0.03226	10.4387
	$81.8 < W \leq 85.8$			0.1	0.38
	$W > 85.8$			0.098765	0.27407

TABLE II:

System gain parameters K_1 and K_2 for low, normal and high BMI.

BMI Category (kg/m ²)	Weight Range (kg)	$K_1 \times 10^4 (\frac{kg}{kcal \cdot day})$	$K_2 \times 10^4 (\frac{kg}{kcal \cdot day})$
Low BMI (< 19.8)	$W \leq 52$	1.59	-1.77
	$52 < W \leq 57.7$	2.09	-2.32
	$W > 57.7$	1.97	-2.19
Normal BMI (19.8–26)	$W \leq 60.2$	0.81	-0.90
	$60.2 < W \leq 65.1$	1.76	-1.96
	$W > 65.1$	1.94	-2.15
High BMI (> 26)	$W \leq 81.8$	1.67	-1.85
	$81.8 < W \leq 85.8$	2.12	-2.36
	$W > 85.8$	2.12	-2.35

TABLE III:

Rates of missing measurements for the self-monitored or self-reported EB variables. Note: Missingness % is compute by $\frac{\text{Number of missing measurements (days)}}{\text{Total number of days for data collection}} \times 100\%$.

Participant	<i>W</i>	<i>PA</i>	<i>EI</i>
A	10%	11%	26%
B	16%	0%	59%

TABLE IV:

EI estimation results for Participants A and B with back-calculation method and Kalman filtering approach.

	<i>EI_{est}</i> (kcal)	
Participant	Back-calculation	Kalman filtering
A	2660±570	2581±498
B	2617±434	2488±488

TABLE V:

Summary of structures proposed to correct self-reported EI . Each structure is characterized by the number of model parameters and the number of pieces of required information (longitudinal measurements of EB variables). The estimator corresponding to each regression model is a subset of $[a_1, \xi, a_2, \beta]$.

Model Structure		Para #	Info #
A	$EI_{est}(k) = \xi$	1	0
B	$EI_{est}(k) = a_1 EI_{rep}(k)$	1	1
C	$EI_{est}(k) = \xi + a_1 EI_{rep}(k)$	2	1
D	$EI_{est}(k) = a_1 EI_{rep}(k) + a_2 EI_{rep}(k)^2$	2	1
E	$EI_{est}(k) = \xi + a_1 EI_{rep}(k) + a_2 EI_{rep}(k)^2$	3	1
F	$EI_{est}(k) = \xi + a_1 EI_{rep}(k) + \beta W_{meas}(k)$	3	2
G	$EI_{est}(k) = \xi + a_1 EI_{rep}(k) + a_2 EI_{rep}(k)^2 + \beta W_{meas}(k)$	3	2

TABLE VI:

Estimation Results for Participants A and B using the semi-physical approach for all the proposed model structures (A to G in Table V). *RMS* represents the root mean square of the residuals.

Model Structure	Participant A				Participant B			
	Estimation		Validation		Estimation		Validation	
	$\hat{E}I_{est}$	RMS	$\hat{E}I_{est}$	RMS	$\hat{E}I_{est}$	RMS	$\hat{E}I_{est}$	RMS
A	2665±3E-12	654	2665±3E-12	675	2671±1E-12	503	2671±1E-12	483
B	2455±699	1008	2257±771	1160	2592±500	648	2628±410	577
C	2665±39	653	2676±43	673	2671±90	495	2677±74	477
D	2594±457	769	2540±536	938	2646±289	549	2680±98	470
E	2665±91	648	2677±69	687	2671±90	495	2677±79	478
F	2665±50	652	2674±50	665	2671±198	463	2674±183	409
G	2665±95	647	2675±74	680	2671±203	460	2678±189	406

Estimated parameters and energy intake under different scenarios. Note: % error is compute by $\frac{(\text{True Value} - \text{Estimated Value})}{\text{True Value}} \times 100\%$.

TABLE VII:

	Scenarios					
	2500 kcal			4500 kcal		
$\mu_{EI_{rept}}$	$\frac{1}{9}$			$\frac{1}{9}$		
$\sigma_{nEI_{rept}}^2$	$\frac{9}{1}$			$\frac{9}{1}$		
$\sigma_{nEI_{rept}}^2$	$\frac{1}{9}$			$\frac{1}{9}$		
$\theta_1 (\alpha_1)$	True	Estimated	% error	True	Estimated	% error
$\theta_2 (\xi)$	True	Estimated	% error	True	Estimated	% error
EI (kcal)	True	Estimated	% error	True	Estimated	% error
	3150±110	3150±35	3150±313	5350±110	5350±35	5350±313
	400	2875 (619%)	400	400	4854 (1114%)	895 (124%)
	1.1	0.1 (-90%)	1.1	1.1	0.1 (-90%)	1.0 (-10%)