2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Novel Method for Generation of *In Silico* Subjects with Type 2 Diabetes

Roberto Visentin, Mattia De Lazzari

***Abstract*— A type 2 diabetes (T2D) simulator has been recently proposed for supporting drug development and treatment optimization. This tool consists of a physiological model of glucose/insulin/C-peptide dynamics and a virtual cohort of T2D subjects (i.e., random extractions of model parameterizations from a joint parameter distribution) well describing both average and variability realistic T2D dynamics. However, the state-of-art procedure to get a reliable virtual population requires some post-processing after subject extraction, in order to discard implausible behaviors. We propose an improved method for virtual subjects' generation to overcome this burdensome task. To do so, we first assessed a refined joint parameter distribution, from which extracting a number of subjects, greater than the target population size. Then, target-size subsets are undersampled from the large cohort. The final virtual population is selected among the subsets as the one maximizing the similarity with T2D data and model parameter distribution, by means of measurement' outcome metrics and Euclidian distance (Δ), respectively. In the final population, almost all the outcome metrics are statistically identical to the clinical counterparts (p-value>0.05) and model parameters' distribution differs by ~5-10% from that derived from data. The methodology described here is flexible, thus resulting suitable for different T2D stages and type 1 diabetes, as well.**

***Clinical Relevance*— A straightforward subjects' generation would ease the availability of tailored *in silico* trials for testing diabetes treatment in a specific population.**

I. Introduction

Diabetes treatments usually requires the administration of oral or injectable drugs. Testing new treatments and medications is often time-consuming and expensive. Moreover, sometimes it is not possible to perform an experiment on human subjects because it cannot be done at all, or it is too difficult, too dangerous, or unethical [1]. For these reasons, the application of computer simulation can assume an extremely important role. In particular, the so-called "*in silico* clinical trials" represent a replacement of animal testing that, allowing testing several conditions in a cost-effective way, helps to cut the time and cost required to develop a medical product or a drug.

Simulation in diabetes field started about 50 years ago, but in a first stage, they did not have a significant impact as they were based on average data [2]. A significant step forward was made in 2008 with the development of the UVA/Padova Type 1 diabetes (T1D) simulator [3]. The crucial innovation of this tool compared to the previous simulators was the availability of a cohort of virtual subjects well spanning the inter-individual variability of a real population. This tool has been constantly updated through the years [4], providing continuous support for testing artificial pancreas prototypes and, more recently, insulin molecules and glucose sensors.

Recently, the same group proposed a type 2 diabetes (T2D) simulator [5] to support the research also for the most common form of diabetes disease. Similarly to T1D, the T2D simulator consists of a simulation model able to describe the dynamics of the glucose, insulin and C-peptide in T2D subjects, and a cohort of virtual subjects, representative of an early stage T2D population.

In both T1D and T2D simulators, virtual subjects are meant as realizations of model parameters, which are randomly extracted from an appropriate joint parameter distribution, i.e., a mean vector and a covariance matrix calculated from the parameters estimated on real subjects. This procedure, already described in details in [4],[5], is rather articulated and, requires post-processing to obtain the final population: for example, outlier subjects with implausible dynamics have to be manually discarded; in addition, if the population does not fit the real average and variability in the data, a new set of subjects has to be generated.

In order to improve this burdensome task, here we propose a novel procedure for the generation of *in silico* subjects, also evaluating possible refinements that can further increase virtual population reliability. The methodology, schematized in Fig. 1, is described in the following section. Here it is applied for generating subjects of the T2D simulator, however, it can be employed in T1D as well.

II. Methods

A. *The Padova T2D Simulator*

The Padova T2D simulator has been presented in early 2020 [5]. It consists of a model of glucose, insulin and C-peptide dynamics during a meal and a population of 100 *in silico* T2D subjects, spanning the inter-individual variability of a real T2D population.

The simulation model shares the core structure proposed by Dalla Man et al. in 2007 [6], with slight modifications including a better description of glucose dynamics in hypoglycemia, a more physiological model of insulin kinetics [7], and models of C-peptide secretion and kinetics [8].

The 100 *in silico* T2D subjects are generated from an appropriate joint distribution of model parameters. More precisely, the simulation model has been decomposed in its single processes; each process has been identified on triple-

R. Visentin, is with the Department of Information Engineering, University of Padova, Via G. Gradenigo 6/B, 35131 Padova, Italy (phone: +39 049 827 7636; fax: +39 049 827 7699; email: visentin@dei.unipd.it).

Mattia De Lazzari is with the Department of Electrical Engineering, Chalmers University of Technology, Hörsalsvägen 11, 412 58, Göteborg, Sweden (email: lazzari@chalmers.se).

tracer data of 51 T2D [9]-[11] and 204 healthy subject data [12], allowing a reliable estimation of all the model parameters, with which building up the joint parameter distribution; then, each *in silico* subject has been generated as a single realization of model parameters vector, randomly extracted from the joint distribution. For more details on the model structure and the process for generation of the *in silico* population we refer to [5].

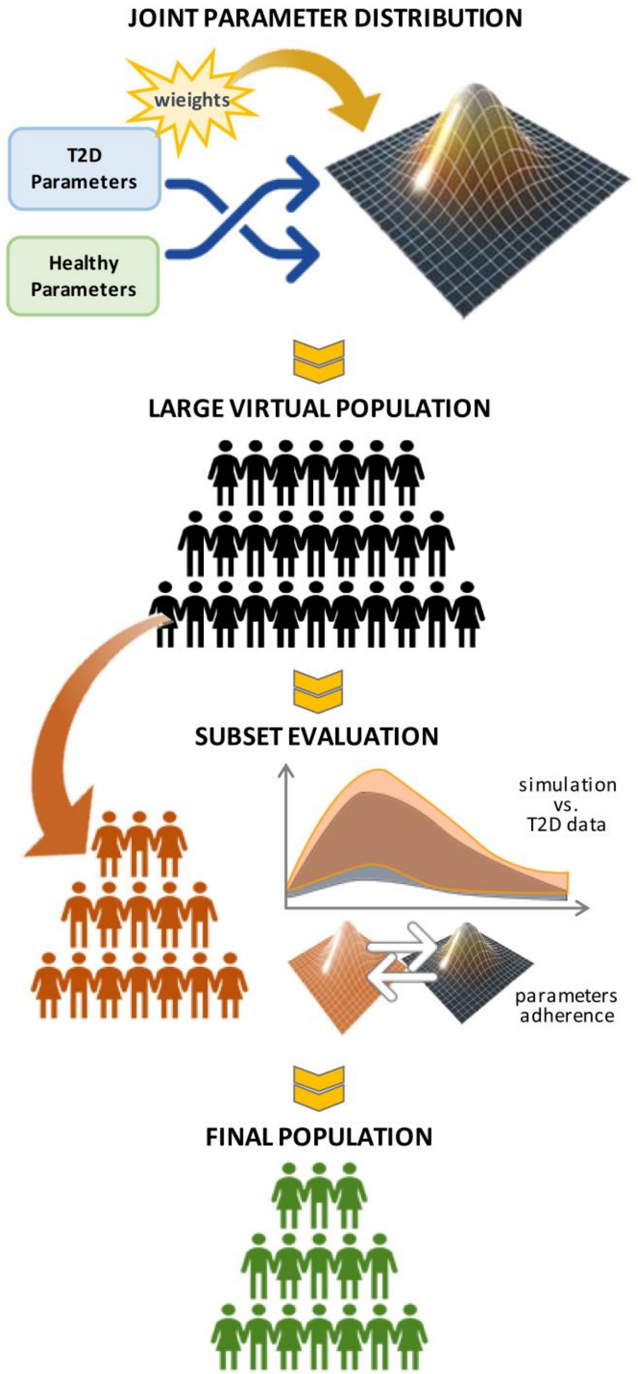

Fig. 1. Flowchart describing the process for generating the virtual population. A joint parameter distribution is computed from estimated parameters (possibly weighted) of real T2D and/or healthy subjects. An initial population, larger than the desired one, is then generated by randomly extracting different realizations of model parameter vector. Then, subsets of a desired number of subjects (e.g., N=100) are randomly extracted. The final population is selected as the subset that mostly represents the real population based on the similarity of both concentration time courses and parameter distribution.

B. Database and protocols

In this work, we used the same data employed for generating the virtual population of the T2D simulator [5]. Specifically, data come from two different groups of individuals, consisting of 51 T2D subjects (16 male, Age = 54.6±8.5 years, H = 169±8 cm, W = 94.1±15.5 kg, BMI = 33.1±5.5 kg/m^2) [9]-[11] and 204 healthy subjects (117 male, Age = 55.5±21.7 years, H = 171±10 cm, W = 78±13.3 kg, BMI = 26.63±3.39 kg/m^2) [12]. In both datasets, subjects underwent a triple-tracer mixed meal tolerance test with different carbohydrates content. This particular technique [13], allowed obtaining, beside sampled measurements of glucose, insulin and C-peptide concentrations, virtually model-independent estimates of metabolic fluxes, i.e., endogenous glucose production, meal glucose rate of appearance, and glucose utilization. More details on data and protocols are described in [9]-[13]. In addition to data time courses, parameter estimates were available for both T2D and healthy subjects, obtained by the identification of T2D simulation model, as described in [5].

C. Development of Joint Parameter Distribution

In the Padova T2D Simulator, each virtual subject is represented by a vector of model parameters, **p:**

$$\boldsymbol{p} = \left[p_1, p_2, \dots, p_{Np}\right]^T \tag{1}$$

randomly extracted from a joint parameter distribution characterized by a mean vector, $\boldsymbol{\mu_p}$, and covariance matrix, $\boldsymbol{\Sigma_p}$. In particular, since model parameters follow a log-normal distribution, the *i*-th element in $\boldsymbol{\mu_p}$ actually corresponds to the average of the log-transformed *i*-th parameter, while the generic element $cov_{pi,pj}$ of $\boldsymbol{\Sigma_p}$ represents the covariance between the logarithm of the two parameters p_i and p_j. At the state of art, $\boldsymbol{\mu_p}$ was calculated using T2D parameters only [9]-[11], while a hybrid $\boldsymbol{\Sigma_p}$ was calculated using parameters of T2D [9]-[11] in combination with those of healthy subjects [12]. This was adopted in order to improve the robustness of $\boldsymbol{\Sigma_p}$, as discussed in [5].

An additional feature was considered here for possible improvement in joint parameter distribution, i.e., the precision of parameter estimates, available by model identification on T2D data [9]-[11], expressed by the coefficient of variation (CV). Specifically, a weight can be associated to a generic *i*-th parameter of subject *k* (p_i^k):

$$w_i^k = \frac{CV_{i_{min}}}{CV_i^k} \tag{2}$$

where $CV_{i_{min}}$ is the minimum CV of the *i*-th parameter among all subjects. When accounting for this information, it is possible to determine a weighted $\boldsymbol{\mu_p}$, in which the weighted average value for the a generic *i*-th parameter is:

$$\mu_i = \frac{\sum_k w_i^k \cdot p_i^k}{\sum_k w_i^k} \tag{3}$$

and the weighted covariance in $\boldsymbol{\Sigma_p}$ between two parameters, p_i and p_j of weights w_i and w_j, respectively, is calculated as follows:

$$cov_{p_i,p_j} = \sigma_{p_i}\sigma_{p_j}r_{p_ip_j} = \frac{\sum_k\left(p_i^k-\mu_i\right)\left(p_j^k-\mu_j\right)}{\sqrt{\sum_k w_i^k w_j^k}} \tag{4}$$

where σ_{p_i} and σ_{p_j} are weighted standard deviation while r_{p_i,p_j} is the weighted correlation between the two parameters.

Hence, by calculating $\boldsymbol{\Sigma_p}$ from T2D only ($\boldsymbol{\Sigma_p^{\mathrm{T2D}}}$) or hybrid (T2D and healthy) data ($\boldsymbol{\Sigma_p^{\mathrm{Hyb}}}$), and using whether weighted or unweighted information, we evaluated six different joint distributions:

- *prior 1*: unweighted $\boldsymbol{\mu_p}$ and $\boldsymbol{\Sigma_p^{\mathrm{T2D}}}$;
- *prior 2*: unweighted $\boldsymbol{\mu_p}$ and $\boldsymbol{\Sigma_p^{\mathrm{Hyb}}}$;
- *prior 3*: weighted $\boldsymbol{\mu_p}$ and $\boldsymbol{\Sigma_p^{\mathrm{T2D}}}$;
- *prior 4*: weighted $\boldsymbol{\mu_p}$ and $\boldsymbol{\Sigma_p^{\mathrm{Hyb}}}$;
- *prior 5*: unweighted $\boldsymbol{\mu_p}$ and weighted $\boldsymbol{\Sigma_p^{\mathrm{T2D}}}$;
- *prior 6*: unweighted $\boldsymbol{\mu_p}$ and weighted $\boldsymbol{\Sigma_p^{\mathrm{Hyb}}}$.

To note, *prior 2* is the configuration currently implemented in the T2D simulator [5].

D. Virtual population generation and assessment

From each joint parameter distribution (*prior 1* to *9*), 1500 subjects, more than the expected population size, were generated by randomly extracting different realizations of T2D model parameters (to note, the desired number of subjects composing the final virtual population (N=100) is obtained as detailed in the next section II-E). Subjects with Mahalanobis distance from the distribution higher than the 95% were discarded, as done in [4],[5]. By adopting this strategy, six different virtual populations have been generated, each one consisting of a different number of subjects.

E. Subjects selection and optimal population assessment

Each subject of [9]-[11] was univocally associated to a certain class, based on the fact that its glucose (G), insulin (I), C-peptide (Cp) was lower (l) or greater (h) than the respective population average. As such, eight classes were possible, and their probability was calculated as P(Class i)=N_i/N_{tot}, with N_i the number of subjects belonging to the i-th Class and N_{tot} the total number of subjects:

- Class 1: G_h-I_h-Cp_h → P(Class 1) = 0.10 (*5 subjects*)
- Class 2: G_l-I_h-Cp_h → P(Class 2) = 0.04 (*2 subjects*)
- Class 3: G_h-I_l-Cp_h → P(Class 3) = 0.25 (*13 subjects*)
- Class 4: G_h-I_h-Cp_l → P(Class 4) = 0 (*none*)
- Class 5: G_h-I_l-Cp_l → P(Class 5) = 0.02 (*1 subject*)
- Class 6: G_l-I_l-Cp_h → P(Class 6) = 0.06 (*3 subjects*)
- Class 7: G_l-I_h-Cp_l → P(Class 7) = 0.25 (*13 subjects*)
- Class 8: G_l-I_l-Cp_l → P(Class 8) = 0.28 (*14 subjects*)

For each of the six virtual cohorts, 10000 subsets were generated, each of them consisting of 100 subjects sampled from the large cohort assuming the same distribution among the eight classes described above (e.g., 6 subjects were sampled from those having glucose, insulin lower and C-peptide greater than the respective average data).

Among the 10000 subset generated from each prior, the one better matching the real population was determined in terms of both glucose, insulin, C-peptide concentrations and model parameters, as follows.

For each subset, the 100 subjects underwent a 420-min scenario with 83.5 g of carbohydrates at time t = 0 min, i.e., the same experiment performed in real T2D subjects [9]-[11]. The similarity between virtual subjects and real data was then evaluated based on plasma glucose, insulin, and C-peptide time courses and the distribution of subjects' model parameters. Simulated glucose, insulin and C-peptide were compared to real data in terms of *FIT* index:

$$FIT = 1 - \sqrt{\frac{\sum_{k=1}^{N}\left(y^{meas}(t_k)-y^{sim}(t_k)\right)^2}{\sum_{k=1}^{N}(y^{meas}(t_k)-\bar{y})^2}} \qquad (5)$$

where y^{sim}, y^{meas} and $\bar{y}$ are the simulated plasma concentration, the measured plasma concentration, and its mean value, respectively. For each subset, *FIT* was calculated with respect to three characteristic curves, i.e., a central one (mean) and lower and upper bound (mean ± standard deviation), for both glucose, insulin and C-peptide.

For what concerns similarity of model parameters, for each subset, the *a posteriori* covariance matrix ($\boldsymbol{\Sigma'_p}$) was calculated, and compared to that *a priori* ($\boldsymbol{\Sigma_p}$ - i.e., that used for subject generation) by calculating the Euclidian distance:

$$\Delta = |\boldsymbol{\Sigma_p} - \boldsymbol{\Sigma'_p}| \qquad (6)$$

For sake of comparison, a vector of reference distances Δ^{ref} has been considered, where the i-th element Δ_i is the Euclidean distance between the *a priori* covariance matrix and its "biased" version, calculated from the original parameter distribution affected by a random percent variation of i = 5, 10, 15, …, 100%.

Finally, the optimal subset (SUB_{opt}) was selected among the 10000s as the one maximizing the sum of *FITs*, subtracted Δ:

$$SUB_{opt} = \mathrm{argmax}[\textstyle\sum_{n=1}^{9}(FIT_n) - \Delta] \qquad (7)$$

Each of the six optimal subset was assessed by comparing simulated data against the available clinical data, in terms of concentration time courses, area under the concentration curve (AUC), maximum concentration (C_{max}) and its time (T_{max}). Statistical tests were performed to assess the statistical significance between outcomes calculated on real and simulated data. Normality of outcomes has been assessed with Lilliefors test. Statistical comparisons have been performed by unpaired two-sample t-test or Wilcoxon rank sum test, for normal or non-normal distributions, respectively, with significance level p = 0.05.

III. Results

Among all the possible combinations, the 100 *in silico* subjects which better reflect the real population variability are obtained from T2D weighted covariance matrix and non-weighted mean vector (*prior 5*). As shown, in Fig. 1, simulated glucose, insulin and C-peptide reproduce both average and variability observed in real subjects [9]-[11]. Distributions of the outcome metrics is reported in Table 1. Specifically, no statistically significant differences are reported in all the comparison except for glucose T_{max} (p < 0.01). Regarding the distribution of subjects' model parameters, the *a posteriori* covariance matrix showed good similarity with that *a priori*: Δ = 0.4023, meaning a difference between 5-10% (Δ_5 = 0.2574, Δ_{10} = 0.6896). Similarity was also higher, if compared to state-of-art $\boldsymbol{\Sigma_p^{\mathrm{Hyb}}}$ configuration (*prior 2*, Δ = 0.4534) and its weighted version (*prior 6*, Δ = 1.5172).

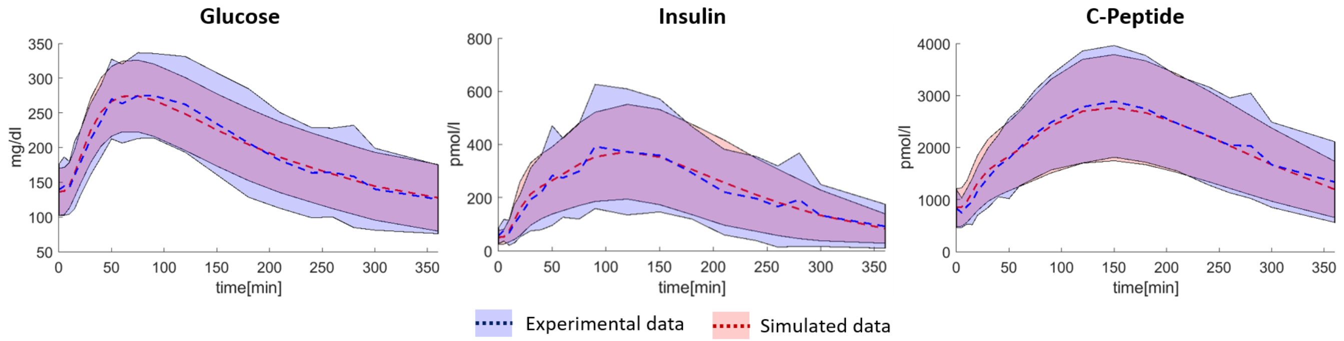

Fig. 2. Glucose, insulin and C-peptide concentrations of 100 T2D subjects generated from weighted covariance matrix and non-weighted mean vector and triple variable classification. Time courses are reported as mean (dotted lines) ± standard deviation (shaded areas).

IV. Discussion

We developed a novel method for generating *in silico* subjects to be included in the T2D simulator. In particular, two novelties are introduced with respect to the state-of-art method for subject generation [5]. First, the use of a refined *a priori* information, consisting of a mean vector and, especially, a covariance matrix calculated from weighted parameters of T2D data only. In particular, the use of only T2D parameters with their weights (calculated as function of estimated parameter coefficient of variation) provides a more robust covariance matrix than the hybrid one (i.e., that derived from both T2D and healthy data), even in its weighted form. Second, the selection of virtual subjects, which are randomly sampled from different classes related to their glucose insulin and C-peptide concentrations allow to properly cover the population variability observed in real subjects. In this regard, the *in silico* vs. *in vivo* comparison done by evaluating *FIT* index on average and boundary curves reduces post-processing to obtain the final population, like the manual discard of outliers.

As shown in the previous section, the resulting optimal *in silico* population well agrees the T2D data. To note, glucose T_{max} was significantly lower *in silico* than *in vivo*. However, it is worth noting that experimental glucose variations are quite contained in the [60-100] min time interval, so that small errors in glucose measurement may have a potential impact on evaluating the outcome.

TABLE I. OUTCOMESMETRICS

	Clinical	*Simulated*	*p-value*
Glucose			
AUC [10^4 mg/dl·min]	6.98 ± 2.08	6.58 ± 1.64	N.S.
C_{max} [mg/dl]	286 ± 59	279 ± 51	N.S.
T_{max} [min]	91 ± 29	73 ± 15	<0.001
Insulin			
AUC [10^4 pmol/l·min]	7.96 ± 4.55	7.78 ± 3.79	N.S.
C_{max} [pmol/l]	432 ± 249	402 ± 191	N.S.
T_{max} [min]	128 ± 48	128 34	N.S.
C-peptide			
AUC [10^5 pmol/l·min]	7.42 ± 2.61	7.19 ± 2.46	N.S.
C_{max} [pmol/l]	3155 ± 1083	2871 ± 1047	N.S.
T_{max} [min]	158 (49)	152 (35)	N.S.

Distribution of glucose, insulin and C-peptide AUC, C_{max} and T_{max} calculated in clinical and simulated data. Values are reported as mean ± standard deviation (SD). Statistical *p*-vale is calculated by unpaired two-sample t-test or Wilcoxon rank sum test, for normal or non-normal distributions, respectively.

V. Conclusion

The proposed method can effectively replace the current procedure for virtual subjects generation, simplifying the post-processing required to obtain a reliable population (i.e., representative of real subjects). This methodology can be also used for tuning the simulator to a desired target population (e.g. advanced-stage T2D, or T1D), thus enabling *in silico* testing for different forms of diabetes.

References

[1] M. Viceconti, *et al.*, “In silico assessment of biomedical products: The conundrum of rare but not so rare events in two case studies.,” *Proc. Inst. Mech. Eng. H.*, vol. 231, no. 5, pp. 455–466, 2017.

[2] C. Cobelli *et al.*, “Diabetes: Models, Signals, and Control”, *IEEE Rev. Biomed. Eng.,* vol. 2, pp. 54-96, 2009.

[3] B. Kovatchev, *et al.,* “In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes, *J. Diabetes Sci. Technol.*, vol. 3, no. 1, pp. 44-55, 2009.

[4] R. Visentin *et al.*, “The UVA/Padova Type 1 Diabetes Simulator Goes from Single Meal to Single Day,” *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 273–281, 2018.

[5] R. Visentin, C. Cobelli, C. Dalla Man, “The Padova Type 2 Diabetes Simulator from Triple-Tracer Single Meal Studies: *In Silico* Trials also possible in Rare but Not-So-Rare Individuals,” *Diabetes Technol. Ther.*, vol. 22, no. 11, pp. 892-903, 2020.

[6] C. Dalla Man, R. A. Rizza, and C. Cobelli, “Meal simulation model of the glucose-insulin system,” *IEEE Trans Biomed Eng*, vol. 54, no. 10, pp. 1740–1749, 2007.

[7] R.S. Sherwin, *et al.*, “A Model of the Kinetics of Insulin in Man”, *J Clin Invest.,* vol. 53, no. 5, pp. 1481-1492, 1974.

[8] R. P. Eaton, *et al.*, “Prehepatic insulin production in man: kinetic analysis using peripheral connecting peptide behavior.,” *J. Clin. Endocrinol. Metab.*, vol. 51, no. 3, pp. 520–528, 1980.

[9] A. Vella *et al.*, “Effects of Dipeptidyl Peptidase-4 Inhibition on Gastro intestinal Function, Meal Appearance, and Glucose Metabolism in Type 2 Diabetes,” *Diabetes*, vol. 56, no. 5, pp. 1475–1480, 2007.

[10] A. Basu, *et al.*, “Effects of type 2 diabetes on insulin secretion, insulin action, glucose effectiveness, and postprandial glucose metabolism”, *Diabetes Care*, vol. 32, no. 5, pp. 866–872, 2009.

[11] G Bock, *et al.* “Mechanisms of fasting and postprandial hyperglycemia in people with impaired fasting glucose and/or impaired glucose tolerance”, *Diabetes*, vol. 55, pp. 3536–3549, 2006.

[12] R. Basu, *et al.*, “Effects of age and sex on postprandial glucose metabolism: differences in glucose turnover, insulin secretion, insulin action, and hepatic insulin extraction”, *Diabetes*, vol. 55, no. 7, pp. 2001-14, 2006.

[13] R. Basu *et al.*, “Use of a novel triple-tracer approach to assess postprandial glucose metabolism”, *Am J Endocrinol Metab.,* vol. 284, no. 1, pp. E55-69, 2003.