Potential predictors of type-2 diabetes risk: machine learning, synthetic data and wearable health devices

1st Paola Stolfi
Inst. Applied Computing (IAC)
National Research Council of Italy
Rome, Italy
p.stolfi@iac.cnr.it
{corresponding author}

2nd Ilaria Valentini
Inst. of Aerospace Medicine “A. Di Loreto”
Rome, Italy
ilaria.valentini@aeronautica.difesa.it

3rd Maria Concetta Palumbo
Inst. Applied Computing (IAC)
National Research Council of Italy
Rome, Italy
mariaconcetta.palumbo@cnr.it

4th Paolo Tieri
Inst. Applied Computing (IAC)
National Research Council of Italy
Rome, Italy
paolo.tieri@cnr.it

5th Andrea Grignolio
Inst. of Biomedical Technologies (ITB)
National Research Council of Italy
Rome, Italy
and
Medical Humanities - International MD Program
Vita-Salute San Raffaele University
Milan, Italy
andrea.grignolio@itb.cnr.it

6th Filippo Castiglione
Inst. Applied Computing (IAC)
National Research Council of Italy
Rome, Italy
filippo.castiglione@cnr.it

Abstract—Investigation about the mechanisms involved in the onset of type 2 diabetes in absence of familiarity is the focus of a research project which has led to the development of a computational model that recapitulates the aetiology of the disease. The model simulates the metabolic and immunological alterations related to type-2 diabetes associated to several clinical, physiological and behavioural characteristics of representative virtual patients.

In this study, the results of 46170 simulations corresponding to the same number of virtual subjects, experiencing different lifestyle conditions, are analysed for the construction of a statistical model able to recapitulate the simulated dynamics.

The resulting machine learning model adequately predicts the synthetic data and can therefore be used as a computationally-cheaper version of the detailed mathematical model, ready to be implemented on mobile devices to allow self assessment by informed and aware individuals.

Index Terms—T2D, diabetes, mathematical and computational modelling, simulation, machine learning, random forest.

I. INTRODUCTION

Type 2 diabetes (i.e., non-insulin-dependent, T2D) is a chronic multifactorial metabolic disorder typical of late adulthood characterised by less effective hormone insulin efficiency at lowering blood sugar. According to data reported by the World Health Organization, type 2 diabetes accounts for 85-90% of all cases of diabetes in the World [1]. There are many different mechanisms that contribute to the onset of type 2 diabetes [2]. For this reason research has steadily moved toward the simultaneous observation of several factors such as metabolic, immunological, genetic and nutritional aspects, with the aim of clarifying the multifactorial nature of the disease. In this regards, a recent study had pointed out a specific state of inflammation, unique for its characteristics and distinct from the classic inflammatory state, which manifests itself in the presence of a high-calorie diet and “susceptible” lifestyles [3]. The term metaflammation well describes this kind of inflammation, mainly originating in the visceral adipose tissue, caused by a high caloric and sugar-rich diet [4]. The classical response of cells to this inflammatory-eliciting insult includes several intracellular signals leading to the release of low levels of cytokines such as Tumour Necrosis Factor-α (TNF-α), and Interleukin-6 (IL-6) [5]. At the same time, it is observed the inhibition of the insulin signal by phosphorylation of a serine in its receptor Insulin Receptor Substrate-1 (IRS-1) [6]. In this way, the insulin receptor is unable to perform its activity, turning the cells as insulin-resistant, and finally resulting in hyperglycaemia. This condition further induces a pro-inflammatory response which alter the metabolic functions of the adipocytes [7]. In the long term, this state causes hyperglycaemia and eventually full blown type 2 diabetes [8].

The above described scenario exhorts a predictive approach to identify metabolic and inflammatory driving factors; a method amenable of being implemented on cheap devices, easy to use by the medical staff and patients themselves. In this perspective, the main aim of the EU-funded project

This study is partially funded by the European Commission under the 7th Framework Programme (MISSION-T2D project, contract No.600803)
“Multi-scale Immune System Simulator for the Onset of Type 2 Diabetes” (MISSION-T2D) [9] was to develop and validate a multi-level and patient-specific model, able to integrate metabolic, nutritional and lifestyle data in order to simulate and predict the metabolic and inflammatory processes underlying the development of type 2 diabetes in the absence of familiarity. The advantage of this approach is that biological data are “evaluated” in a computer in real time, that is, as soon as they are available by direct body measurements [10], hence allowing to predict the value of hormones and/or metabolites that are difficult to be measured in vivo.

The definition of the risk of T2D is a complex task accounting for the level of insulin resistance, the level of inflammatory cytokines and the pro-inflammatory cell counts. These observables are, among other, used in the introduced mathematical description of the complex interdependencies among metabolites and pancreatic control as well as among adipose tissue components and inflammation. This model, upon setting anthropometric parameters such as age, sex, body weight, height, and providing nutritional habits, fitness status and physical activity patterns by the user, predicts the risk of progressing toward a T2D-related state in a predefined time horizon.

Machine learning (ML) is becoming a popular and efficient approach to evaluate multidimensional longitudinal health data in different fields of medical research such as diagnostics of asymptomatic liver disease [11], predicting opioid dependence [12], evaluating sociodemographic determinants of health status in aging [13], predicting the mobility of medical rescue vehicles [14], forecasting adverse perioperative outcomes [15], measuring caloric intake at the population level [16], personalizing oncological treatment in radiogenomics [17], determining features of systolic blood pressure variability [18], revealing clinical variables in bipolar disorder [19], with a specific interest in uncovering potential predictors of diabetes (type 1 and 2) using large set of data [20]–[27]. ML may offer accurate results with fewer requirements if compared with traditional mathematical modelling, that is why its methods are being used to extract harder-to-detect knowledge. ML models are particularly useful in research where the input is represented by an enormous amount of diagnostic data and the output by predictive therapeutic data.

Due to the high level of sophistication, the computational model mentioned above (M-T2D) is quite computationally expensive and is therefore not a viable solution for self-assessment when executed on mobile devices. In view of this limitation of the preventive potential of this approach, we present here the results from a large number of virtual (i.e., simulated) subjects experiencing different lifestyle conditions. Such results have been analysed with the purpose of setting up a ML derived-model able to recapitulate the simulated dynamics, thus computing the risk index, with a reduced computational effort thus allowing to work in real time. Therefore, we applied statistical learning theory, that has already led to successful results in many fields such as speech recognition, computer visions and sports just to name a few, to find a simplified predictive model based on data [28], [29].

II. METHODS

In this section we briefly describe i) the computational model used to generate the data that is analysed, ii) the set up for the simulations generating the data and iii) the procedure for their analysis.

A. The computational model

The whole-body multi-scale computational for fuel homeostasis (herein referred to as M-T2D) describes the metabolic, hormonal and inflammatory changes due to exercise sessions [30] and food ingestion [31]. In few words M-T2D consists in the combination of differential equations and an agent-based model merged together into a unified multi-scale simulation tool of the metabolic/inflammatory response to physical activity and food intake.

M-T2D is an extended formulation of the model proposed by Kim et al. [32], describing fuel homeostasis in response to a session of exercise and that incorporates the hormonal model inspired by the work of Saunders [33], in which both glucagon and insulin are produced and glucose regulation is achieved by altering the balance between the two. With respect to the original work of Kim and colleagues, in [30] we provide a better description of the physical exercise as provided by Roy et al. [34] and Kildegaard and colleagues [35] with the aim of achieving greater generalisation and user-customization. In particular, we used a “relative” (rather than fixed) exercise intensity as well as the estimation of functional capacity in relation to age, sex, anthropometric characteristics and current fitness status. Moreover, M-T2D includes oxygen consumption and the dynamics of epinephrine as directly dependent on the relative exercise intensity to modulate hormones and metabolites responses to different exercise modalities (e.g., cycling, walking, running, stepping).

For what concern the description of the physiological changes due to food ingestion, stomach emptying and absorption of macronutrients monomers in the gut, in [31] M-T2D took inspiration from a former model by Dalla Man [36] and Elashoff [37]. The description of the dynamics of alanine and triglycerides from proteins and fats ingestion, respectively, needed the settings of proper parameters, since the model in [36] is limited to the description of glucose dynamics. Since insulin resistance or insulin-deficient states lead to a reduced response of tissues, such as the skeletal muscle, liver, and adipose tissue, to insulin, M-T2D implements the effects of insulin resistance on the glucose uptake by peripheral organs [38]. Besides that, while modelling the fasting plasma glucose concentration we took into consideration factors depending on dietary habits, physical activity and inflammation. These factors contribute in different ways to increase or diminish the blood sugar level. The glycaemia rises due to unhealthy eating habits, leading to inflammation. Also, it decreases if the patients do physical exercises.

All together, M-T2D includes several model compartments: i) a model of energy balance and weight gain/loss is added
ANTHROPOMETRIC MEASURES
- Sex \( S \in \{ \text{female, male} \} \)
- Age \( A \in \{ 28, 38, 48, 58, 68 \} \)
- Weight \( W \in \{ \text{underweight, normal, overweight} \} \)
- Height \( H \in \{ \text{short, average, tall} \} \)

PHYSICAL ACTIVITY
- Number of sessions per week \( N_{PA} \in \{ 0, 1, 2, 3 \} \)
- Duration (mins) \( D_{PA} \in \{ \text{low} = 30, \text{medium} = 60, \text{high} = 90 \} \)
- Intensity (% of VO\(_{2}\)max) \( I_{PA} \in \{ \text{low} = 40, \text{high} = 60 \} \)

FOOD INTAKE (3 meals per day, breakfast, lunch, dinner)
- Carbohydrates (grams) \( C_{ME} \in \{ \text{low, med, high} \} \)
- Proteins (grams) \( P_{ME} \in \{ \text{low, med, high} \} \)
- Fats (grams) \( F_{ME} \in \{ \text{low, med, high} \} \)

TABLE I: The different virtual subjects have been generated by varying the parameters in this table and corresponding to 46170 different initial conditions.

In [38], based on the equations provided by Mifflin [39] and Westerterp [40]; ii) the emergence of the inflammation is described as the result of adipose mass increase which, in turn, is a direct consequence of a prolonged excess of high calorie intake [41]; iii) to counteract the inflammatory scenario, the presence of anti-inflammatory mechanisms promoted during exercise by skeletal muscle have been considered, on the basis of a previous published study [42]; iv) finally, to describe the inflammatory status of the subject, M-T2D merges the metabolic model with a general purpose simulator of the immune system [43], a modelling framework used to study different human pathologies [44]–[46], specific aspects of the immune response [47], [48] and also aspects of non-human immunity [49].

B. The generation of synthetic data

Simulated trajectories of the dynamical model M-T2D starting from different initial conditions corresponding to different virtual subjects have been generated by varying the parameters in TABLE I. The total number of \( m = 46170 \) is thus the product of the following terms (\( | \cdot | \) indicates the cardinality of the set),

\[
m = |S| \cdot |A| \cdot |W| \cdot |H| \cdot (1 + N_{PA} \cdot |D_{PA}| \cdot |I_{PA}|) \\
|C_{ME}| \cdot |P_{ME}| \cdot |F_{ME}|
\]

The low/medium/high quantities of carbohydrates, proteins and fats are computed taking into account the balance of calories resulting from the meal with the total daily energy expenditure (TDEE) as explained in [38]. In more details, TDEE is the result of the sum of Resting Energy Expenditure (REE), Activity Energy Expenditure (AEE) [50] and Thermic Effect of Food (TEF) [51]. We implemented the equations by Mifflin and coworkers in [39] to estimate the AEE considering weight, height, age and sex. We determine the AEE on the basis of the intensity, duration, volume of oxygen consumed and the number of sessions of the exercise as in [38]. Finally, the TEF is the amount of energy expenditure that occurs after eating, due to the cost of digesting and processing food and represents about 10% of the calories due to meal ingestion [51]. The resulting TDEE represents the quantity of calories that have to be ingested in order to have a balance among energy intake and expenditure. In our calculation these calories are somehow arbitrarily yet realistically split between breakfast (25%TDEE), lunch (45%TDEE) and dinner (30%TDEE). Furthermore, for each meal we divided the caloric content of the meal in quantity of calories coming from amount of carbohydrates, proteins and fats equal to 50%, 20% and 30%, respectively. Finally, to convert grams from calories we used the Atwater general factor system [52]. These “standard” or average values of grams for carbohydrates, proteins and fats are used as reference values (‘med’) and simple multiplications to the constants 0.8 and 1.5 are meant to correspond to ‘low’ and ‘high’ quantities of the food intake description given above.

The complete patient specification of the initial condition of the simulation is thus given as a string vector. For example the initial condition specified by the string female 28 obese tall 2 60/40 low/high/low corresponds to a 28 years old female subject, tall and obese, which exercises twice a week (sixty minutes each time with an intensity of 40%VO\(_{2}\)max) and which follows a diet made of low amount of carbohydrates and fats but rich in proteins. So in general we indicate the vector corresponding to the initial condition as follows:

\[
x = [S, A, W, H, (N_{PA}, D_{PA}, I_{PA}), (C_{ME}, P_{ME}, F_{ME})].
\]

Simulations’ outputs were analysed on the basis of the following variables which are deemed the most significant to calculate the risk of developing T2D: Glucose Base Level (GBL, namely the fasting glucose concentration), Body Mass Index (BMI), and Tumor Necrosis factor-\(\alpha\) (TNF) as measured in the adipose tissue compartment. The execution of M-T2D starting from the initial condition \( x \) generated a complete trajectory of these variables with a time resolution of ten seconds. However, since we are interested in analysing the condition of the virtual subject only at the end of a specified period of six months, these measures are taken after six months of routinely and uninterrupted physical activity and diet patterns as specified (among the other things) in \( x \). Formally,

\[
y = [BMI(t), GBL(t), TNF(t)]
\]

where \( t \) is six months. The couples \((x^{(k)}, y^{(k)})\) for \( k = 1, \ldots, m \), are used as a training set for the development of a statistical model able to recapitulate, given \( x \), the dynamics shown by the computational model and to predict the risk of developing T2D over a time horizon of six months. In other words, our goal was to find a statistical/ML model (which should not be confused with the computational model M-T2D) able to predict the dependent variables, namely \( y \) given a set of regressors/predictors \( x \), that is, the initial conditions defining the virtual subject and her/his lifestyle.

The main reason for finding such ML model is that the complexity of M-T2D requires a significant computational effort to run, whereas a statistical model, once trained, provides a real time solution of computing \( y^{(i)} \) given \( x^{(i)} \) allowing a
fast generalisation to cases other than those in the training set\
\[ \{x^{(k)}, y^{(k)}\}_{k=1,...,m}. \]

III. Results

The ML model presented here provides a simplified way to predict the expected behaviour at time \( t \) of the output variables given the input variables at time \( t_0 < t \), where \( t - t_0 = 6 \) months. To this end, we adopted a data driven approach over the simulated patterns. In particular, the ML model has been constructed and validated by using the initial conditions \( x \) of the regressors as input variables and the dependent variables \( y \) as output variables, therefore the training set consisting in couples \( \{(\text{regressors}, \text{dependent variables})\} \), has been used to train a model able to generalise to unseen couples.

A. Preliminary analysis

In Fig.1 we report i) the correlations, ii) the histograms and iii) the scatter plots of both the regressors and the dependent variables organised in a comprehensive matrix-like figure. In particular, the top off-diagonal boxes report the sample Pearson Correlation Coefficients \( \rho_{ij} \) between couple of variables \( x_i \) and \( x_j \), that is, \( \rho_{ij} = \frac{(n-1)s_is_j^{-1}}{\sum_{k=1}^{m} \left[ (x_i^{(k)} - \mu_i)(x_j^{(k)} - \mu_j) \right]} \), where \( s_i, s_j \) are the standard deviations and \( \mu_i, \mu_j \) are the mean of variables \( x_i \) and \( x_j \) respectively. Their significance is highlighted with a number of red stars from a maximum of three (highly significant correlation) to zero (absence of correlation). On the diagonal we find the histograms of each variable and the bottom off-diagonal boxes report the scatter plots of each couple of variables together with the fit shown in red (a poor fit, that is, a lack of dependence between the two variables, appears as a horizontal or vertical red line).

Fig.1 provides an overview to identify critical key features of the dataset. We noticed that there are non-linear dependencies between the output variables and the regressors (see for instance the scatter plot of \( BMI(0) = W(0)/H^2 \)). This suggested that a non-linear ML model should be considered rather than a linear one. This is not surprising given the high level of complexity of the process generating the data, M-T2D. Moreover, the variables related to the diet (\( C_{\text{ME}}, P_{\text{ME}}, F_{\text{ME}} \), cfr. scatter plots in the bottom off-diagonal panels) do appear strongly correlated. However, these correlations are “spurious” because the corresponding variables depend linearly on another variable indicating the amount of calories intake. The origin of this correlation, which lays in the use of the model described in [38], has already described above in section II. Lastly, the scatter plots show that the output variables \( BMI, GBL \) and \( TNF \) are correlated.

Of all above strongly suggested that a multivariate model is the appropriate choice in the attempt to construct a ML model recapitulating the data. In formula, we are looking for a statistical model defined as

\[ y = \psi(x) + \epsilon, \quad \epsilon \sim \mathcal{N}_3(0, \Sigma), \]

where \( x \) and \( y \) are the vectors of regressors and dependent variables respectively, \( \mathcal{N}_3 \) is a Gaussian in \( \mathbb{R}^3 \) with zero mean and covariance matrix \( \Sigma \), and \( \psi(\cdot) \) is a function to be determined.

We tested several statistical models starting from the simplest, namely, the linear regression model. Even though preliminary results already prove it unfit, it is interesting to quantify the error made by a linear model. Furthermore, we tested non-linear models, specifically, polynomial regression models of several orders, from 2 to 4. Finally we tested the random forest method [28]. To evaluate each of the above models we divided the dataset into train set consisting of \( 2m/3 \) data points used to estimate the parameters of the models and the remaining \( m/3 \) data points in the test set used to test the predictive performance of the model. The obtained results are reported in Fig.2. Each row, corresponding to one of the ML models tried, shows the out-of-sample (i.e., computed on the test set) scatter plot of the true and the predicted values. The linear regression model, obtained by defining \( \psi \) in eq(1) as a linear combination of the regressors, is not able to describe the behaviour of none of the dependent variables. Indeed, all the scatter plots in Fig.2a are far from the \( y = x \) line. It is interesting to observe that the scatter plots of \( BMI \) (leftmost panel) and the \( TNF \) (rightmost panel) suggests that the linear model does partially capture these variables’ dynamics as the dots are aligned along the \( y = x \) line, meaning that there is positive correlation between predicted and true values even though an unwanted very large variability in the predicted.
value. Conversely, the scatter plot of GBL (middle panel) suggests that the ML model completely fails in describing its behaviour because there is no evident correlation between the true and the fitted values. This result confirms the non-linear structure among the input $x$ and the output $y$ variables and justifies the use of non-linear model. Figures 2b, 2c and 2d are related to the polynomial regression models of degree 2, 3 and 4 respectively, obtained by defining $\psi$ in eq(1) as a polynomial of order $d$. From the plots it is clear that BMI (leftmost panel) and TNF (rightmost panel) are only partially described by these models because the scatter plots again show large variation in the predicted value hence the use of polynomial models does not improve significantly even by increasing its degree. As for the linear model, the middle panels relating to the scatter plots of true vs. predicted GBL fail to show a clear correlation hence leading to the conclusion that the polynomial structure is not appropriate.

B. Decision trees and random forest

In statistics/ML one of the most well-known regression/classification ensemble method algorithms is the random forest introduced in [28] as a special case of decision trees [53]. The general idea of this algorithm is to construct a forest of decision trees and to define the output to be either the mean of all the outputs in the case of regression trees or the result of a majority rule on the output in the case of classification trees.

C. Learning the parameters of the random forest from synthetic data to predict the risk of T2D

In applying the random forest algorithm to predict $y$ from $x$ with respect to eq(1) we can write

$$\psi(x) = \frac{1}{N} \sum_{i=1}^{N} T_i(x),$$

where $N$ is the number of trees and $T_i$ is the “structure” of the $i$-th tree or, more generally, the regression value of the $i$-th decision tree constituting the forest.

Results obtained by applying the random forest algorithm are shown in Fig.2e. As clearly shown by the three panels, the multivariate random forest outperforms the previous ones in predicting $y$. Indeed the scatter plots of all three variables are aligned on the $y = x$ line indicating a fairly good correlation. Just, a bit of variability is still observed for small values of GBL and TNF.

In table II we quantify the error produced by each ML model as the Mean Square Error (MSE) in measuring the goodness of the fit, in particular, MSE$_\text{In}$ has been computed over the train set (i.e., in sample) while the MSE$_\text{Out}$ has been computed over the test set (i.e., out of sample). As expected, the highest error corresponds to the linear model and it slightly decreases in polynomial regression models when using higher degree polynomials. Finally, the multivariate random forest regression shows to outperform all other regression methods bringing down the MSE to more than one order of magnitude compared to the polynomial regression. Also to note that the small increase of the MSE$_\text{Out}$ compared to the MSE$_\text{In}$ denotes the absence of overfitting of data.

**Fig. 2:** Each row shows the out-of-sample (i.e., in the test set) scatter plots of the true and fitted (i.e., predicted) values of the variables specified in each panel’s caption (from left to right, BMI, GBL and TNF). Inset plots show the histogram of the out-of-sample residues’ (i.e., the prediction error). The last row shows that multivariate random forest performs better predictions when compared to the linear or polynomial regression.

IV. DISCUSSION

The random forest statistical model derived in the previous section, which has shown a good predictive ability, is the fittest to describe the data used to train it. In other words, it can be used to perform a dedicated analysis of the data set with the scope of finding interesting or non-trivial bio-clinical readouts.
As first results, we looked at the variables’ importance using a method already described in [54]. In few words, we measured the impact of each variable on the predictive power of the model, as the difference between the prediction error computed when some noise is added to the predictors and the prediction error computed on the original predictors. Such impact is shown in Fig.3, where the variables’ importance for each of the elements of \( y \) are plotted. The impact of some variables appears to be the same for the three variables \( BMI, GBL \) and \( TGF \). Indeed, we observed that the variables related to the physical activity (i.e., \( N_{PA}, D_{PA}, \) and \( I_{PA} \)) appear as the less important. This fact points out that a better accounting for the physical activity on anti-inflammatory factors as well as on the reduction of glucose base level already on time horizons smaller than six months is required in M-T2D. This is a task which is already ongoing and will be reported in due time [38].

The most important variable for both the \( BMI \) (grey bar in Fig.3) and the \( GBL \) (black bar) is the initial value of the body mass index (\( BMI_0 \)). This means that the weight plays an important role in determining the glucose base level thus in the determination of the risk of T2D. As for the remaining variables, we observed that they have comparable impact on the \( BMI \). This is not the same for the glucose base level or \( GBL \) index, for which the second most important dependence is with the amount of carbohydrates in the diet (\( C_{ME} \)). For what concerns the inflammation represented by the level of TNF-\( \alpha \) (i.e., \( TNF \) index, white bar in Fig.3) the most important dependence is, as expected, the age (\( A \)) followed in order of importance by \( C_{ME} \), the body mass index. This is interesting as it goes along the recently defined concept of inflammaging [55] which joins immune-metabolic processes with age-related diseases in a single, integrated, clinical framework.

To carry on with the analysis of the relative importance of each input variable, we calculated their influence when taken in pairs. Again, we measure the impact of the couple \((x_i, x_j)\) as the difference between prediction error when to \((x_i, x_j)\) is added some noise versus the prediction error calculated in the unmodified case [54]. Looking at the pairwise co-influence on \( y \) in Fig.4, we noted that the most common of them involve \( BMI_0 \). This is somehow expected since \( BMI_0 \) is the most important variable for all the output variables and both importance analyses are computed using the same methodology.

To overcome any bias coming from this procedure, we considered another method to investigate the variables' co-

---

**TABLE II:** In-sample and out-of-sample MSE of all tested models.

| ML model                           | MSEIn          | MSEOut         |
|-----------------------------------|----------------|----------------|
| linear regression                 | 0.6220638      | 0.6913094      |
| polynomial degree 2               | 0.5798217      | 0.6456507      |
| polynomial degree 3               | 0.5016218      | 0.5675261      |
| polynomial degree 4               | 0.4233801      | 0.4937277      |
| multivariate random forest        | 0.01991242     | 0.0276875      |

---

**Fig. 3:** Impact of each input variable \( x \) on the output \( y \). Inset plot shows the same data in y-log scale to increase readability (y-scale is in arbitrary units). This plot offers a one-sight readout of the impact of subjects anthropometric measures and lifestyle patterns on the likelihood to progress toward a state of higher risk of development of diabetes.

**Fig. 4:** Top twelve pairwise co-influence on \( y \) calculated by method in [54].
reach better efficiency in diabetes care and ameliorate patient involvement in diabetes self-management, which can decrease the surge of diabetes-related healthcare expenditures, paving the way to the future scenario of a patient-driven diabetes care in the technology era [58]. Also, this new approach has great potential as a low-cost monitoring tool for nutritional habits and physical activity of different segments of the population, permitting their users to achieve knowledge hardly comprehensible by even the best expert.

In this work we have shown how a computational model running very complex simulations of realistic multivariate scenarios can be used to feed a machine learning method which demonstrated to perform satisfactorily to predict the risk of T2D using notably less time and computational resources, making it compliant for mobile devices use and for customised and immediate responses to the users. In perspective, the ability to link the subject’s parameters with measuring devices such as those in portable communication systems (smartphones and wrist watches) enables the development of health care systems linked in real time to issue alerts, warnings or simple recommendations to the patient [29]. In the near future, the “real time” execution of the model, with completely customisable input parameters, can be envisaged as a dedicated bioinformatics service, able to provide increasingly personalised healthcare and facilitating self monitoring.

We conclude by looking at the near future, where we envision at least two avenues of research. A new era of medicine is opening up by combining traditional data from randomised clinical trials with new real-world data, collected from registries, electronic health records, social media, and wearable devices which produce real-world evidence, which can both uncover potential predictors of diabetes or challenge several RCTs data so far collected [27].

REFERENCES

[1] WHO Media centre. http://who.int/mediacentre/factsheets/fs312/en/ [accessed 2016-09-27].
[2] M. Y. Donath, D. M. Schumann, M. Faulenbach, H. Ellingsgaard, A. Perren and J. A. Ehses, “Islet Inflammation in Type 2 Diabetes. From metabolic stress to therapy,” Diabetes Care, vol. 31 (Suppl.1), pp. S161-S164, Feb 2008.
[3] M. Y. Donath, and S. E. Shoelson, “Type 2 diabetes as an inflammatory disease,” Nat. Rev. Immunol., vol. 11, pp. 98-107, Feb 2011.
[4] M. F. Gregor and G. S. Hotamisligil, “Inflammatory Mechanisms in Obesity,” Annu. Rev. Immunol., vol. 29, pp. 415-445, Apr 2011.
[5] M. S. Akash, K. Rehman and S. Chen, “Role of inflammatory mechanisms in pathogenesis of type 2 diabetes mellitus,” J. Cell Biochem., vol. 114, pp. 525-531, Mar 2013.
[6] G. S. Hotamisligil, “Inflammation and metabolic disorders,” Nature, vol. 444, pp. 860-867, Dec 2006.
[7] G. S. Hotamisligil, and E. Erbay, “Nutrient sensing and inflammation in metabolic diseases,” Nat. Rev. Immunol. vol. 8, pp. 923-934, Dec 2008.
[8] M. Y. Donath, E. Dalmas, N. S. Sauer and M. B’onni-Schnetzler, “Inflammation in Obesity and Diabetes: islet Dysfunction and Therapeutic Opportunity,” Cell Metab., vol. 17, pp. 860-872, Gia 2013.
[9] F. Castiglione, et al., “The onset of type 2 diabetes: proposal for a multi-scale model,” JMRIR Res. Protoc., vol. 2, p. e44, Oct 2013.
[10] Y. Vodovotz, M. Csete, J. Bartels, S. Chang and G. An, “Translational systems biology of inflammation,” PLoS Comput. Biol., vol. 4, p. e1000014, Apr 2008.
[11] A. Babic, G. Bodemar, et al. “Machine learning to support diagnostics in the domain of asymptomatic liver disease,” Medinfo 8 Pt 1: pp. 809-813, 1995.
