Five-year prediction of glucose changes with missing data in a Reproducing Kernel Hilbert Space

M. Matabuena, P. Félix, C. Meijide-García and F. Gude

***Abstract*— It is estimated that approximately 50% of diabetes mellitus patients remain undiagnosed. Furthermore, adherence and effectiveness of treatments are poor across many patient groups; and disease prevalence is increasing with contemporary lifestyles. In this context, new predictive models to understand those risk factors associated with the evolution of glycemic profiles in the short and long term are vital for identifying patients at risk of disease development, improving early diagnosis, and prescribing effective treatment. Our main goal is to examine the relationship between the baseline characteristics of participants in a five-year study and a primary biomarker for diabetes diagnosis and monitoring, the glycated hemoglobin (A1c). In addition, we make use of continuous glucose monitoring (CGM) to capture individual glucose homeostasis fluctuations at a high-resolution level. As five-year A1c data are missing for approximately 40% of patients, we propose a new data-analysis framework based on Reproducing Kernel Hilbert Space (RKHS) learning. In particular, we address the statistical independence testing problem and we do several adaptions of existing model-free methods of variable selection, regression and conformal inference. By using these methods we achieve new clinical findings: i) We identify some diabetes biomarkers associated with glucose variations, both marginally and from a multivariate perspective, ii) We show that CGM provides extra information to predict long-term changes in glucose metabolism, and iii) We identify some risk phenotypes for which our predictive capacity is moderate, and therefore more personalized follow-up is needed.**

***Index Terms*— Diabetes mellitus, Reproducing Kernel Hilbert Space, missing data, statistical independence, variable selection, regression modeling.**

I. Introduction

DIABETES mellitus is one of the most critical public health problems, being the ninth major cause of mortality worldwide [1]. At present, over 416 and 47 million patients have Type 2 and Type 1 diabetes respectively [2]. Importantly, around 50% of patients with diabetes are undiagnosed [2]. Moreover, several projections forecast a significant increase in prevalence in the following decades [3]. Considering the impact of this pandemic among the general population, there is a need for new health policies and guidelines to enable early recognition of risk patients and improvement in the methodology of disease diagnosis in the standard clinical routine [4].

The availability and rapid adoption of new digital medical devices have enabled an emerging clinical paradigm based on precision medicine, which will be called to improve early diagnosis and subsequent clinical decisions through the intensive use of statistical models and machine learning techniques [5]–[8]. In the particular case of diabetes the latest advances in sensing technology allows for assessing the glucose metabolism at a high-resolution level, by capturing the individual differences in the glucose fluctuations at different time scales via continuous glucose monitoring (CGM) [9]. Recent studies have shown improved glycemic control and decreased rates of hypoglycemia in Type 1 diabetes (T1D) patients using CGM, leading both the Endocrine Society and American Diabetes Association to state that CGM use represents standard of care in T1D [10], [11].

The first aim of this paper is to provide a method for predicting glucose homeostasis in the long term, by mainly using information provided by a CGM device and some common clinical variables. Among different biomarkers, we select the glycated hemoglobin (A1c) as response variable. A1c is a measure of average blood glucose level over the past 3 months, and it is the preferred option because it provides more reproducible values in laboratory and is subject to less measurement error [12]. Furthermore, we aim to identify several variables associated with the evolution of A1c in the long term, and to assess and discuss the residuals and the predictive capacity from the clinical point of view, providing interpretable clinical phenotypes for large uncertainty cases. The validity of the present approach is tested on a five-year longitudinal population-based study, including both healthy and diabetic subjects, where a subsample of participants underwent continuous glucose monitoring procedures at the beginning of the study. As expected, an important number of participants withdrew from the study and therefore an analysis robust to missing data is demanded [13].

In this paper, we focus our attention to the case where

This work was supported by the Instituto de Salud Carlos III under Grant PI16/01395, the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-099646-B-I00, the Consellería de Educación, Universidade e Formación Profesional and the European Regional Development Fund under Grant ED431G-2019/04.

M. Matabuena and P. Félix are with the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, SPAIN (e-mail: marcos.matabuena@usc.es).

C. Meijide-García is with the Universidade de Santiago de Compostela, 15782, Santiago de Compostela, SPAIN.

F. Gude is with the Unidade de Epidemioloxía Clínica, Complexo Hospitalario Universidade de Santiago (CHUS), Travesía da Choupana, 15706 Santiago de Compostela, SPAIN.

the response variable is the only variable with missing data, a common situation in longitudinal studies. We choose the Reproducing Kernel Hilbert Space (RKHS) learning paradigm to tackle our missing data problem, due to its ability to model complex non-linear relations between study variables [14], [15]. Furthermore, RKHS paradigm is particularly suitable for dealing with heterogeneous complex data such as graphs or curves that take values on a continuum [16], as is the case with the functional representation of glucose profiles that we propose in Section II-A. Thus, we introduce new general-purpose methods for statistical independence testing (Section III-A), variable selection (Section III-B), and inference on the uncertainty of new predictions (Section III-C), by adapting previous kernel methods to the missing data setting.

The rest of this paper is outlined as follows: Section II describes the AEGIS database used for testing our proposal. Section III describes in detail those new methods for statistical independence testing, variable selection, and inference on the uncertainty of new predictions. Section IV shows the results from applying these new methods to AEGIS database. Section V discusses the advantages and drawbacks of this approach. Finally, some conclusions are provided in Section VI.

II. AEGIS DATABASE

The AEGIS population study, conducted in the Spanish town of A Estrada (Galicia), aims to analyze the steady evolution of different clinical features such as longitudinal changes in circulating glucose in 1516 patients over 10 years. In addition, non-routinary medical tests such as continuous glucose monitoring are performed every five years on a randomized subset composed of 581 patients. At the beginning of this study [17], 581 participants were randomly selected for wearing a CGM device for 3-7 days. Out of the total of 581 participants, 68 were diagnosed with diabetes before the start of the study, and 22 during the study. Table I shows the basal characteristics of these 581 patients grouped by sex. After a five-year follow-up, a significant fraction of those individuals did not agree to perform a second glucose monitoring, while some five-year relevant outcomes such as A1c could only be measured on 349 patients.

	Men ($n = 220$)	Women ($n = 361$)
Age, years	47.8 ± 14.8	48.2 ± 14.5
A1c, %	5.6 ± 0.9	5.5 ± 0.7
FPG, mg/dL	97 ± 23	91 ± 21
HOMA-IR, $mg/dL.\mu UI/m$	3.97 ± 5.56	2.74 ± 2.47
BMI, kg/m^2	28.9 ± 4.7	27.7 ± 5.3
CONGA, mg/dL	0.88 ± 0.40	0.86 ± 0.36
MAGE, mg/dL	33.6 ± 22.3	31.2 ± 14.6
MODD	0.84 ± 0.58	0.77 ± 0.33

TABLE I
CHARACTERISTICS OF AEGIS STUDY PARTICIPANTS WITH CGM MONITORING BY SEX. MEAN AND STANDARD DEVIATION ARE SHOWN. A1C: GLYCATED HEMOGLOBIN; FPG: FASTING PLASMA GLUCOSE; HOMA-IR: HOMEOSTASIS MODEL ASSESSMENT-INSULIN RESISTANCE; BMI: BODY MASS INDEX; CONGA: GLYCEMIC VARIABILITY IN TERMS OF CONTINUOUS OVERALL NET GLYCEMIC ACTION; MAGE: MEAN AMPLITUDE OF GLYCEMIC EXCURSIONS; MODD: MEAN OF DAILY DIFFERENCES.

A. Glucodensity

We adopt a novel functional representation for CGM data, termed glucodensity, to assess glucose homeostasis in diabetes patients [18]. Glucodensity is a natural extension of Time in Range (TIR) metrics, the gold standard measure for representing CGM data. TIR measures the proportion of time that a person spends with their blood glucose levels in a target range [19], [20]. TIR is an intuitive metric, but has the disadvantage that the range will vary depending on the individual and there is a loss of information caused by the discretization of the recorded data in different intervals. Instead, glucodensity effectively measures the proportion of time each individual spends at each specific glucose concentration, thus providing better predictions of A1c, HOMA-IR, and the CONGA, MAGE, and MODD variability measures than TIR does.

Given a series of CGM data $\{x_j\}_{j=1}^m$ the glucodensity can be modeled as a probability density function $f(\cdot)$ that can be approached by a kernel density estimation,

$$\hat{f}(x) = \frac{1}{m}\sum_{j=1}^{m} k_h(x - x_j), \tag{1}$$

where $h > 0$ is the smoothing parameter and $k_h(s) = \frac{1}{h}k(\frac{s}{h})$ is a non-negative real-valued integrable function (Figure 1).

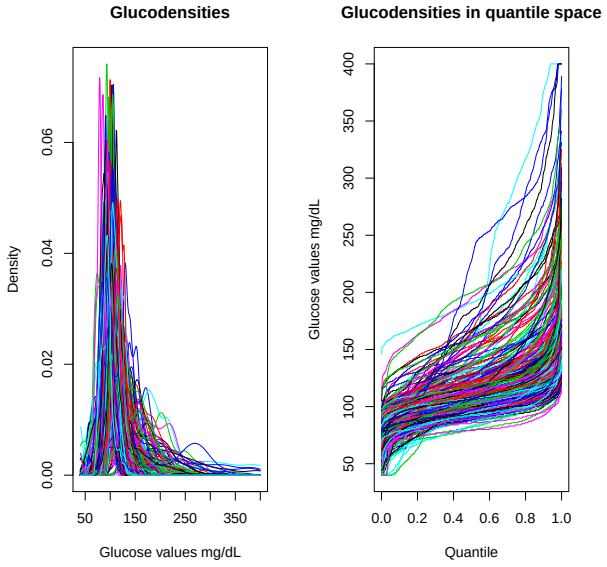

Fig. 1. Glucodensities estimated from a random sample of the AEGIS study on diabetic and normoglycemic patients are shown. Left: glucose representation estimates the proportion of time spent by a patient at each glucose concentration over a continuum. Right: the representation of the glucodensities in the space of quantile functions is shown.

III. METHODS

We shall first pose the problem in general terms. Let $(\mathbf{X}, Y, R)$ be a random vector such that $\mathbf{X} = (X^1, \ldots, X^p) \in \mathcal{X}$ denote the covariates, $Y \in \mathbb{R}$ the response variable, and $R \in \{0, 1\}$ a binary random variable that indicates whether the response is missing or not. $\mathcal{X}$ denotes a topological space, meaning that can be arbitrary, either discrete, continuous or

structured. Let $D = \{(\mathbf{x}_i, y_i, r_i)\}_{i=1}^n$ be a dataset of independently and identically distributed observations. We assume R to be distributed according to the probability law $\pi(\cdot) = P(R = 1|\mathbf{X} = \cdot)$, which only depends on the covariates $\mathbf{X}$. We may also assume that $\mathbf{X}$ is sufficient to explain the possible dependence between R and Y, or equivalently, R and Y are conditionally independent given $\mathbf{X}$, $R \perp\!\!\!\perp Y|\mathbf{X}$. Let us also assume the following relation between $\mathbf{X}$ and Y:

$$Y = f(\mathbf{X}) + \epsilon, \tag{2}$$

where ϵ denotes a random noise with $\mathbb{E}(\epsilon|\mathbf{X}) = 0$, and f is the true regression function that is assumed to be smooth. Our goal is to predict Y by proposing a new data analysis framework that is robust to those datasets where some values for Y are not observed; specifically, y_i is missing if $r_i = 0$. To this aim we provide: 1) a method for univariate analysis based on testing statistical independence between each covariate and the response variable; 2) a method for selecting the subset of covariates that best predict the response variable; and 3) methods for predicting the response variable and for inferring the uncertainty in the predictions.

A. Testing statistical independence

We study if there exists a statistical association between each covariate in the AEGIS study and the response variable A1c. To keep the notation uncluttered, we remove the subscript from the covariate unless necessary. In general, we wish to test whether random variables X and Y are not independent, i.e., if we can reject the null hypothesis, $H_0 : X \perp\!\!\!\perp Y$, from n samples $\{(x_i, y_i)\}_{i=1}^n$. To do this, we must calibrate the test under the null hypothesis to determine what results are expected to happen with a certain probability if the null hypothesis holds. In our specific case, we have to take into account the effects of the mechanism of missing data in the response variable Y. We propose a methodology to deal with this problem based on kernel mean embeddings, which is valid when both covariate and response variables live in a separable Hilbert space. In addition, we introduce a new bootstrap procedure to perform test calibration, adapted to kernel mean embeddings.

Hilbert space embeddings of distributions or, in short, kernel mean embeddings [16], allows us to map distributions into a Reproducing Kernel Hilbert Space (RKHS) in which kernel methods can be extended to probability measures. Kernel mean embeddings can be used to define a metric for distributions, the maximum mean discrepancy (MMD), that can be applied to define an independence test, the Hilbert-Schmidt Independence Criterion (HSIC), a non-parametric test of independence with the important property that it does not make any assumption as to the nature of the possible dependence among the two variables [21]. We shall extend this test to the missing data setting.

The kernel mean embedding is built upon a positive definite function known as kernel function $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The positive definiteness of $k_{\mathcal{X}}$ guarantees the existence of a dot product space $\mathcal{H}$, and a mapping $\phi : \mathcal{X} \to \mathcal{H}$, such that $k_{\mathcal{X}}(x, y) = \langle \phi(x), \phi(y) \rangle$. $\mathcal{H}$ is a Hilbert space of real-valued functions defined on $\mathcal{X}$. A reproducing kernel of $\mathcal{H}$ is a kernel function that satisfies: 1) $\forall x \in \mathcal{X}$, $k_{\mathcal{X}}(\cdot, x) \in \mathcal{H}$, and 2) $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k_{\mathcal{X}}(\cdot, x) \rangle_{\mathcal{H}} = f(x)$. If $\mathcal{H}$ has a reproducing kernel, it is said to be a RKHS $\mathcal{H}_{k_{\mathcal{X}}}$. A kernel mean embedding results from extending the mapping ϕ to the space of probability distributions by representing each distribution as a mean function $\phi(P) = \int_{\mathcal{X}} k(\cdot, \mathbf{x}) dP(\mathbf{x})$, resulting in transforming a distribution P into an element of the RKHS $\mathcal{H}_{k_{\mathcal{X}}}$. Given two probability measures, P and Q, a RKHS distance between their embeddings can be defined as the MMD [22]:

$$\mathrm{MMD}_{k_{\mathcal{X}}}(P, Q) = \|\phi(P) - \phi(Q)\|_{\mathcal{H}_{k_{\mathcal{X}}}}. \tag{3}$$

For the class of *characteristic* kernels the embeddings are injective, i.e., $\mathrm{MMD}_k(P, Q) = 0$ if and only if $P = Q$. MMD can then be applied to measuring the degree of dependence between the random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with marginal distributions P_X and P_Y, and jointly distributed as $P_{X,Y}$. Let us note that testing the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ is equivalent to testing $H_0 : P_{X,Y} = P_X P_Y$. We denote by $\phi_X(\cdot)$, $\phi_Y(\cdot)$ and $\phi_{X,Y}(\cdot)$ the kernel mean embeddings of P_X, P_Y and $P_{X,Y}$, respectively. Assuming $\mathcal{H}_{k_{\mathcal{Z}}}$ is a RKHS over $\mathcal{X} \times \mathcal{Y}$ with kernel $k_{\mathcal{Z}}((x, y), (x', y')) = k_{\mathcal{X}}(x, x') k_{\mathcal{Y}}(y, y')$, so that $\mathcal{H}_{k_{\mathcal{Z}}}$ is a direct product $\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}$ (with $\otimes$ being the tensor product), then a natural way of testing independence is measuring the MMD distance between the functions $\phi_{X,Y}(\cdot)$ and $\phi_Y(\cdot) \otimes \phi_X(\cdot)$, which can be written as the Hilbert-Schmidt Independence Criterion (HSIC) between X and Y [22], defined as

$$\mathrm{HSIC}(P_{X,Y}, P_X P_Y) = ||\phi_{X,Y} - \phi_X \otimes \phi_Y||^2_{\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}} \tag{4}$$

and it can be shown that when $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are characteristic kernels then $\mathrm{HSIC}(P_{X,Y}, P_X P_Y) = 0$ if and only if $X \perp\!\!\!\perp Y$. Expanding Equation 4 we have

$$\begin{aligned}
&\mathrm{HSIC}(P_{X,Y}, P_X P_Y) = \\
&\quad = \langle \phi_{X,Y} - \phi_X \otimes \phi_Y, \phi_{X,Y} - \phi_X \otimes \phi_Y \rangle_{\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}} \\
&\quad = \langle \phi_{X,Y}, \phi_{X,Y} \rangle + \langle \phi_X \otimes \phi_Y, \phi_X \otimes \phi_Y \rangle - \\
&\qquad\qquad - 2\langle \phi_{X,Y}, \phi_X \otimes \phi_Y \rangle.
\end{aligned} \tag{5}$$

where we drop $\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}$ in subscript for brevity. By the reproducing property, $\mathbb{E}_P[f(x)] = \langle f, \phi(P) \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$ and Fubini's theorem, we get

$$\begin{aligned}
&\mathrm{HSIC}(P_{X,Y}, P_X P_Y) = \\
&\quad = \mathbb{E}_{X,Y,X',Y'}[k_{\mathcal{X}}(X, X') k_{\mathcal{Y}}(Y, Y')] + \\
&\qquad + \mathbb{E}_{X,X'}[k_{\mathcal{X}}(X, X')] \mathbb{E}_{Y,Y'}[k_{\mathcal{Y}}(Y, Y')] - \\
&\qquad - 2\mathbb{E}_{X,Y}[\mathbb{E}_{X'}(k_{\mathcal{X}}(X, X')] \mathbb{E}_{Y'}[k_{\mathcal{Y}}(Y, Y')]],
\end{aligned} \tag{6}$$

where X' and Y' are independent copies of random variables X and Y. Ultimately, testing independence involves calculating the squared distance between two mean functions in the appropriate RKHS space, resulting from transforming original data in order to capture all distributional differences between both random variables.

In practice, a limited number of samples $\{(x_i, y_i, r_i)\}_{i=1}^n$ are observed. Therefore, we must replace the population mean by sample mean defined through its empirical distribution.

Then, the Hilbert-Schmidt independence criterion can be estimated as

$$\begin{aligned}\widehat{\mathrm{HSIC}}(\hat{P}_{X,Y},\hat{P}_X\hat{P}_Y) &= \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k_{\mathcal{X}}(x_i,x_j)k_{\mathcal{Y}}(y_i,y_j)+ \\ &+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k_{\mathcal{X}}(x_i,x_j)\sum_{i=1}^{n}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k_{\mathcal{Y}}(y_i,y_j)- \\ &-\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}k_{\mathcal{X}}(x_i,x_j)k_{\mathcal{Y}}(y_i,y_k). \end{aligned} \tag{7}$$

We assume data are missing not at random (MNAR), and hence some of the predictor variables can have a certain impact on the mechanism of missing data; for instance, in our example older patients are less susceptible to perform a second CGM monitoring, so that the probability of not observing a patient increase with age. Under MNAR assumption, we observe $\{(x_i,y_i,r_i)\}_{i=1}^n$ and we have to estimate the missing data mechanism, given by the function $\pi(\cdot)=\mathbb{P}(R=1|X=\cdot)$. Several procedures have been proposed in the literature for this aim such as logistic regression, lasso, random forest, or ensemble modeling among others. Afterwards, we re-weight the dataset, taking into account how difficult it is to observe the response of the i^{th} datum. In particular, we associate a weight w_i with the i^{th} datum via inverse probability weighting (IPW) estimator [23], given by

$$w_i=\frac{r_i}{n\pi(x_i)},\quad i=1,\dots,n. \tag{8}$$

We define the normalized weight of w_i as

$$w_i^*=\frac{w_i}{\sum_{i=1}^n w_i},\quad i=1,\dots,n. \tag{9}$$

We denote the estimated i^{th} weight and normalized i^{th} weight as $\hat{w}_i$ and $\hat{w}_i^*$, respectively, after estimate $\hat{\pi}(\cdot)$.

To get an estimator of HSIC with missing data, it is enough to replace the uniform weight $\frac{1}{n}$ of the empirical distribution with the normalized weights $\hat{W}^*=(\hat{w}_1^*,\dots,\hat{w}_n^*)$ in the Equation 7. Thus, we obtain

$$\begin{aligned}\widehat{\mathrm{HSIC}}(\hat{P}_{X,Y},\hat{P}_X\hat{P}_Y) &= \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}_i^*\hat{w}_j^*k_{\mathcal{X}}(x_i,x_j)k_{\mathcal{Y}}(y_i,y_j)+ \\ &+\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}_i^*\hat{w}_j^*k_{\mathcal{X}}(x_i,x_j)\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}_i^*\hat{w}_j^*k_{\mathcal{Y}}(y_i,y_j)- \\ &-\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\hat{w}_i^*\hat{w}_j^*\hat{w}_k^*k_{\mathcal{X}}(x_i,x_j)k_{\mathcal{Y}}(y_i,y_k). \end{aligned} \tag{10}$$

Calibration under the null hypothesis with the precedent statistic is not trivial, and the permutation approach is generally not valid. We propose a novel bootstrap approach, which properly deals with glucodensities and, in general, with complex constrained distributional objects that do not live in vector spaces [24].

Under the null hypothesis $H_0: P_{X,Y}=P_XP_Y$, it can be assumed that $\phi_{X,Y}(\cdot)-\phi_X(\cdot)\otimes\phi_Y(\cdot)=0(\cdot)$. Therefore,

$$\begin{aligned}\widehat{\mathrm{HSIC}}(\hat{P}_{X,Y},\hat{P}_X\hat{P}_Y) &= \\ &= \langle\hat{\phi}_{X,Y}-\hat{\phi}_X\otimes\hat{\phi}_Y,\hat{\phi}_{X,Y}-\hat{\phi}_X\otimes\hat{\phi}_Y\rangle_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}} \\ &= \langle\hat{\phi}_{X,Y}-\phi_{X,Y}+\phi_X\otimes\phi_Y-\hat{\phi}_X\otimes\hat{\phi}_Y, \\ &\hat{\phi}_{X,Y}-\phi_{X,Y}+\phi_X\otimes\phi_Y-\hat{\phi}_X\otimes\hat{\phi}_Y\rangle_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}}. \end{aligned} \tag{11}$$

Then, a natural bootstrap procedure that allows to estimate the p-value for the independence test can be developed as follows:

1) To randomly sample with replacement n elements from the original dataset D, repeating m times. We denote by $D^{j^*}=\{(x_i^{j^*},y_i^{j^*},r_i^{j^*})\}_{i=1}^n$, $j=1,\dots,m$, the j^{th} random sample obtained.
2) To calculate $\widehat{\mathrm{HSIC}}^{j^*}(\hat{P}_{X,Y},\hat{P}_X\hat{P}_Y)$ as

$$\begin{aligned}\widehat{\mathrm{HSIC}}^{j^*}(\hat{P}_{X,Y},\hat{P}_X\hat{P}_Y) &= \\ &= \langle\hat{\phi}_{X,Y}-\hat{\phi}_{X,Y}^{j^*}+\hat{\phi}_X^{j^*}\otimes\hat{\phi}_Y^{j^*}-\hat{\phi}_X\otimes\hat{\phi}_Y, \\ &\hat{\phi}_{X,Y}-\hat{\phi}_{X,Y}^{j^*}+\hat{\phi}_X^{j^*}\otimes\hat{\phi}_Y^{j^*}-\hat{\phi}_X\otimes\hat{\phi}_Y\rangle_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}}, \end{aligned}$$

where $j=1,\dots,m$ and $\hat{\phi}_{X,Y}^{j^*}(\cdot)$, $\hat{\phi}_X^{j^*}(\cdot)$ and $\hat{\phi}_X^{j^*}(\cdot)$ are the kernel mean embeddings estimated from the j^{th} bootstrap sample $D^{j^*}=\{(x_i^{j^*},y_i^{j^*},r_i^{j^*})\}_{i=1}^n$.
3) To estimate the p-value as

$$p\text{-value}=\frac{1}{m}\sum_{j=1}^{m}I\Big(\widehat{\mathrm{HSIC}}^{j^*}(\hat{P}_{X,Y},\hat{P}_X\hat{P}_Y)\geq \widehat{\mathrm{HSIC}}(\hat{P}_{X,Y},\hat{P}_X\hat{P}_Y)\Big). \tag{12}$$

The bootstrap consistency with missing data can be proved by using some standard tools of empirical process theory [25], and it is provided as supplementary material [26].

B. Variable selection

Independence screening methods select predictor variables based on individual prediction ability, and hence they are ineffective in selecting a subset of variables that are individually weak but in combination strong predictors. Subset selection aims to overcome this drawback by considering and evaluating the prediction ability of a subset of variables as a whole. One popular approach to subset selection is based on directly optimizing an objective function consisting of two terms: a data fitting term to attain prediction accuracy, and a regularization term to penalize a large number of variables [27].

Subset selection has been recently approached from the RKHS paradigm with satisfactory results. Two strategies stand out: first, minimizing the trace of the conditional covariance operator [28]; and second, identifying those variables with non-zero gradient function [29]. The first strategy scales badly with the number of variables. The second strategy can be formulated in a more compact way, and here it will be extended to missing data.

Following [29], we propose to identify the relevant predictors by learning the gradient of the true regression function f directly from samples. Thus, it is assumed that if a variable X^r

is not relevant for predicting Y then $g_r = \partial f(\mathbf{X})/\partial X^r = 0$ for any value of $\mathbf{X}$. Let us denote by $\mathbf{g}(\mathbf{X}) = \nabla f(\mathbf{X}) = (g_1(\mathbf{X}), \ldots, g_p(\mathbf{X}))^T$ the true gradient function. We adopt the mean squared error as data fitting term. In a small neighbourhood of $\mathbf{x}_i$ we can use the Taylor expansion to approximate $f(\mathbf{X})$ so when $\mathbf{x}_j$ is close enough to $\mathbf{x}_i$ then $f(\mathbf{x}_j) \approx y_i + g(\mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)$. Then, we can define the estimation error as a function of $\mathbf{g}(\cdot)$:

$$\mathcal{E}(\mathbf{g}) = \mathbb{E}_{\mathbf{X},Y,\mathbf{X}',Y'}[\omega(\mathbf{X},\mathbf{X}')(Y - Y' - \mathbf{g}(\mathbf{X})^T(\mathbf{X} - \mathbf{X}'))]^2,$$

where $\omega(\mathbf{X},\mathbf{X}')$ is an appropriate weight function that decreases as $\|\mathbf{X} - \mathbf{X}'\|$ increases and ensures that the local neighborhood of $\mathbf{X}$ contributes more to estimating the gradient $\mathbf{g}(\mathbf{X})$. Typically, $\omega(\mathbf{X},\mathbf{X}') = e^{-\|\mathbf{X}-\mathbf{X}'\|^2/\tau_n^2}$, where τ_n^2 is a positive parameter which should be adjusted to warrant asymptotic estimation consistency. In addition, with $\mathbf{X}', Y'$, we denote independent and random variables distributed as $\mathbf{X}$ and Y, respectively.

Since only a limited number of samples $\{(\mathbf{x}_i, y_i, r_i)\}_{i=1}^n$ are observed we approximate $\mathcal{E}(\mathbf{g})$ by its empirical version

$$\hat{\mathcal{E}}(\mathbf{g}) = \frac{1}{n^2}\sum_{i,j=1}^{n} \omega_{ij}(y_j - y_i - \mathbf{g}(\mathbf{x}_i)^T(\mathbf{x}_j - \mathbf{x}_i))^2, \quad (13)$$

where $\omega_{ij} = \omega(\mathbf{x}_i, \mathbf{x}_j)$.

We can add a regularization term for enforcing a sparsity constraint on the gradient vector, with the aim of shrinking towards zero the partial derivatives g_r with respect to irrelevant variables. We then add the term $J(\mathbf{g}) = \lambda_n \sum_{r=1}^p \eta_r J(g_r)$ where η_r are adaptive tuning parameters. On the other hand, we can define the estimation error in (13) as a functional in the RKHS $\mathcal{H}_k^p$, so $\mathbf{g} \in \mathcal{H}_k^p$ and $\mathcal{E} : \mathcal{H}_k \times \overset{p}{\cdots} \times \mathcal{H}_k \to \mathbb{R}^+$, induced by a pre-specified positive kernel k, which is assumed to be universal so that, on every compact subset of the input space, every continuous function can be uniformly approximated by sections of the kernel. Thus, we propose the following optimization formula to learn the gradient vector:

$$\underset{\mathbf{g}\in\mathcal{H}_k^p}{\arg\min} \frac{1}{n^2}\sum_{i,j=1}^{n} \omega_{ij}(y_i - y_j - \mathbf{g}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j))^2 + J(\mathbf{g}). \quad (14)$$

When analyzing missing data we propose to substitute ω_{ij} weights by $\hat{\omega}_{ij}^* = \hat{\omega}_i^*\hat{\omega}_j^*\omega_{ij}$, where $\hat{\omega}_i^*$ and $\hat{\omega}_j^*$ denote the estimated normalized weights associated with data i^{th} and j^{th} according to (9). The variable selection expression can be rewritten as

$$\underset{\mathbf{g}\in\mathcal{H}_k^p}{\arg\min} \frac{1}{n^2}\sum_{i,j=1}^{n} \hat{\omega}_{ij}^*(y_i - y_j - \mathbf{g}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j))^2 + J(\mathbf{g}). \quad (15)$$

The representer theorem states that the minimizer of (15) can be represented as a finite linear combination of kernel products evaluated on the samples of the data set [30]:

$$g_r(\cdot) = \sum_{i=1}^{n} \alpha_i^r k_{\mathcal{X}}(\cdot, \mathbf{x}_i), \quad r = 1, ..., p, \quad (16)$$

where $\alpha^r \in \mathbb{R}^n$. Given this representation, $g_r(\cdot) = 0$ iff $\alpha^r = (\alpha_1^r, ..., \alpha_n^r)^T = (0, ..., 0)^T$, or more concisely, $\|\alpha^r\|_2 = 0$.

Several regularization terms have been considered in the bibliography. We adopt the Group Lasso penalty [29], [31]:

$$J(g_r) = \inf\Big\{\|\alpha^r\|_2 : g_r(\cdot) = \sum_{i=1}^{n} \alpha_i^r k_{\mathcal{X}}(\cdot, \mathbf{x}_i)\Big\}, \quad (17)$$

which encourages the entire α_i^r, $i = 1, ..., n$ to be selected or shrunk to zero together, achieving the purpose of variable selection. Thus, our optimization problem can be rewritten as

$$\underset{\alpha^1,\ldots,\alpha^p}{\arg\min} \sum_{i,j=1}^{n} \hat{\omega}_{ij}^*(y_i - f^*(\mathbf{x}_i, \mathbf{x}_j))^2 + \lambda_n \sum_{r=1}^{p} \eta_r \|\alpha^r\|_2, \quad (18)$$

where $f^*(\mathbf{x}_i, \mathbf{x}_j) = y_j - \sum_{r=1}^p \mathbf{k}_i^T \alpha^r (x_i^r - x_j^r)$, being $\mathbf{k}_i = (k(\mathbf{x}_i, \mathbf{x}_1), \ldots, k(\mathbf{x}_i, \mathbf{x}_n))^T$ the i^{th} row of $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$, and λ_n a tuning parameter. This last expression simplifies the original optimization framework (14) from a functional space to a vector space, and it can be solved in $O(|U|^2 p^2)$ by a block coordinate descent algorithm [29].

C. Prediction and uncertainty inference

Let us recall that the ultimate goal is to predict Y by explaining its relationship with covariates $\mathbf{X}$. To this aim we adopt the kernel ridge regression approach proposed by Liu and Goldberg [32]. However, we draw on linear regression theory to compute the leave-one-out cross-validation regularization parameter efficiently. This class of regularization parameters has proven to largely shape the model performance [33]. Furthermore, estimating the uncertainty of the predictions, by providing robust confidence intervals, is considered a valuable tool for subsequent decision. Thus, we compute intervals with good finite sample coverage by using advances in conformal inference recently exploited in causal theory [34].

Let us assume a linear regression model:

$$y_i = f(\mathbf{x}_i) + \epsilon = \mathbf{x}_i^T\beta + \epsilon \quad i = 1, \ldots, n, \quad (19)$$

where β is the vector of coefficients of the linear model. Given the original dataset $D = \{(\mathbf{x}_i, y_i, r_i)\}_{i=1}^n$, kernel ridge regression is based on solving the following optimization problem:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T\beta)^2 + \lambda||\beta||_2^2, \quad (20)$$

which is solved by $\hat{\beta} = (\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I})^{-1}\mathbf{x}^T\mathbf{y}$, where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, and $\lambda > 0$ is the smoothing parameter of regularization term.

Let $\mathcal{H}_k$ be a RKHS with kernel $k_{\mathcal{X}}$. Then, by replacing every $\mathbf{x}_i$ by $\phi(\mathbf{x}_i)$, and further assuming that $\beta = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)$, we obtain an analogue solution to that of Equation (20), but changing the usual dot product by the inner product of the selected RKHS. Particularly, we have $\hat{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$, where $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$.

Authors propose two different estimators for missing data. In both cases, the solution has the same closed-form expression given by Representer Theorem [30]. The first one is given by $\hat{\alpha} = (\lambda\mathbf{I} + \mathbf{W})^{-1}\mathbf{W}\mathbf{y}$, where they handle missing data mechanism via IPW estimator. The second is obtained through

doubly robust estimation, combining a preliminary imputation of the missing response with IPW estimator:

$$\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1}(\mathbf{W}\mathbf{y} + (\mathbf{I} - \mathbf{W})\mu(\mathbf{x})), \tag{21}$$

where $\mathbf{W} = diag(w_1, \ldots, w_n)$ denotes a diagonal matrix containing the weights (see Equation 8) and $\mu(\mathbf{x}) = (\mu(\mathbf{x}_i), \ldots, \mu(\mathbf{x}_n))^T$ denotes the imputation function.

Doubly robust estimators achieve optimal asymptotic variance when their weights $w_1, ..., w_n$ and their imputation function $\mu(\cdot)$ are correctly specified, and only one of them needs to be correctly specified to achieve consistency. However, when any of them fails the regression model performance can deteriorate dramatically with finite sample [35], [36].

The impact of the smoothing parameter on model generalization is an essential issue for the ensuing performance, and is strongly connected with the interpolation problem in RKHS with minimum norm. We propose to select the smoothing parameter through *leave-one-out* cross-validation, by adapting the estimators to missing data [33].

In order to provide a prediction interval for the response with a confidence level of $1 - \alpha$, we introduce a specific algorithm to perform conformal inference [34], [37], valid to handle missing responses and heteroscedastic noisy.

We randomly split the dataset $D = \{(\mathbf{x}_i, y_i, r_i)\}_{i=1}^{n}$ into training and test sets $D^{train} = \{(\mathbf{x}_i^{train}, y_i^{train}, r_i^{train})\}_{i=1}^{n_1}$, and $D^{test} = \{(\mathbf{x}_i^{test}, y_i^{test}, r_i^{test})\}_{i=1}^{n_2}$, with $n = n_1 + n_2$.

For a given new observation $\mathbf{x}_{n+1}$ we go through the following steps:

1) Fit the mean regression function $\hat{f}(\cdot)$ from the set D^{train}, according to Equation 21.
2) Compute the residuals $\hat{\epsilon}_i = |y_i^{test} - \hat{f}(\mathbf{x}_i^{test})|/\hat{\sigma}(x_i^{test})$, for every $i = 1, \ldots, n_2$ with $r_i^{test} = 1$. $\hat{\sigma}(x_i^{test})$ is estimated by a regression function that predicts the absolute deviation of the residuals, fitted with the training sample.
3) Estimate the empirical distribution $\hat{F}^{\epsilon}_{n_2+1}$ by using the previous residuals and assuming an infinite value for the theoretical residual of observation $\mathbf{x}_{n+1}$. For this task, we use the weights defined in Equations (8) and (9) and the function $\hat{\pi}^{train}(\cdot)$, where we must incorporate also the weight of $\mathbf{x}_{n+1}$, $\hat{w}_{n_{2+1}}$.
4) Compute the $1 - \alpha$ quantile, $\hat{q}_{1-\alpha}$, from $\hat{F}^{\epsilon}_{n_2+1}$.
5) Finally, return $[\hat{f}(\mathbf{x}_{n+1}) - \hat{q}_{1-\alpha}\hat{\sigma}(x_{n+1}), \hat{f}(\mathbf{x}_{n+1}) + \hat{q}_{1-\alpha}\hat{\sigma}(x_{n+1})]$ as the required prediction interval.

D. Handling multiple sources with a kernel

RKHS offers a powerful data analysis paradigm that is able to cope with data of different nature [38]. A crucial issue is to select a suitable kernel that accurately captures the differences and specific characteristics of each of the information sources examined. In our particular case, we take into account a continuous probability distribution, and certain real-valued and categorical data, $\mathbf{x} = (x^{gluco}, \mathbf{x}^{real}, \mathbf{x}^{categ})$. A reasonable choice commonly used in the literature is the Laplacian kernel, $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{n \times n}$, where $k(\mathbf{x}_i, \mathbf{x}_j) = exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{\sigma})$. The Laplacian kernel with the standard Euclidean distance is a characteristic and universal kernel in a real vector space. Moreover, it can be shown that the Laplacian kernel retains these properties considering the set of continuous density functions endowed with $L2-$Wasserstein geometry, providing theoretical guarantees that we can approximate a large variety of regression functions. Additionally, we want to detect possible interactions between different data sources. Based on the connection between positive kernels and negative type metrics [39], [40], we propose a simple global Laplacian kernel that integrates the three sources:

$$k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) = \\ = e^{-(a\frac{||x_i^{gluco} - x_j^{gluco}||}{\sigma_{gluco}} + b\frac{||\mathbf{x}_i^{real} - \mathbf{x}_j^{real}||}{\sigma_{real}} + c\frac{||\mathbf{x}_i^{categ} - \mathbf{x}_j^{categ}||}{\sigma_{categ}})}, \tag{22}$$

where $a, b, c, \sigma_{gluco}, \sigma_{real}, \sigma_{categ} > 0$ and we assume for the sake of simplicity that $(a, b, c) \in \mathbb{R}^3$ such that $a + b + c = 1$ and $0 \leq a \leq 1, 0 \leq b \leq 1, 0 \leq c \leq 1$.

E. Selection parameter in Laplacian kernels

The kernel and the initial values for its parameters should be selected to model the adequate behavior of the present approach. Choosing a proper kernel among the family of characteristic and universal kernels is usually not as crucial as adequately tuning its parameters [14].

In order to select the bandwidth parameter $\sigma > 0$ for the Laplacian kernel, the median heuristic has been widely used in kernel methods [41]:

$$\sigma = \sqrt{median\{||\mathbf{x}_i - \mathbf{x}_j||^2 : 1 \leq i < j \leq n\}} \tag{23}$$

Following [42], we search for the optimal kernel bandwidth parameter σ^* in a grid of points of the form σ^{γ} with $\gamma \in (0, 3]$. In the setting of our global Laplacian kernel (Equation 22), we use the median heuristic for each data type to select σ_{gluco}, σ_{mult} and σ_{categ}. Values for a, b, and c are also searched for in a grid, but in this case in a 3-dimensional simplex.

Finally, to incorporate the missing data mechanism in the kernel bandwidth selection, we calculate the median through IPW estimator.

IV. Results

The present framework of predictive tools allows us to provide an answer to some clinical open questions:

1) Glycated hemoglobin A1c is a hemoglobin-glucose combination formed within the cell, which is a useful indicator of long-term blood glucose control and is considered the standard biomarker for diabetes diagnosis and management. *Is there a diagnostic variable that can be used to predict the future glucose changes in individuals by predicting A1c levels?*
2) Current medical literature assigns a prominent role to all of the predictor variables listed in Table I for diagnosing and managing diabetes. It is common knowledge that glucose metabolism is complex and a multivariate model is required to capture glucose changes, *but is there a reduced subset of the predictor variables that better balances complexity and generalization ability?*
3) CGM technology may be able to provide a more suitable assessment of glucose homeostasis by means of an

appropriate representation in terms of glucodensities. *How do CGM data impact on improving our ability to predict future A1c changes?*

4) An increased uncertainty in the predictive power of the model focused in a region of the feature space may suggest a subpopulation that has not been properly modeled. *Can we provide a characterization of those individuals whose future glucose behavior cannot be precisely predicted for a more personalized follow-up?*

A. Is there a diagnostic variable that can be used to predict the future glucose changes in individuals by predicting A1c levels?

To answer this question we study whether there is any evidence of univariate statistical association for normoglycemic patients (A1c<5.7% and FPG<100 mg/dL) between glucose variation measured by $\text{A1c}_{5years}-\text{A1c}_{initial}$ and those predictor variables shown in Table I. Let us note that previous literature have proven a significant statistical association in diabetes patients, particularly when the outcome is poor.

For this purpose, we use the Hilbert-Schmidt independence criterion that we propose in the context of missing data (Section III-A), together with a specific bootstrap approach designed for this task. The underlying missing data mechanism is estimated using univariate logistic regression.

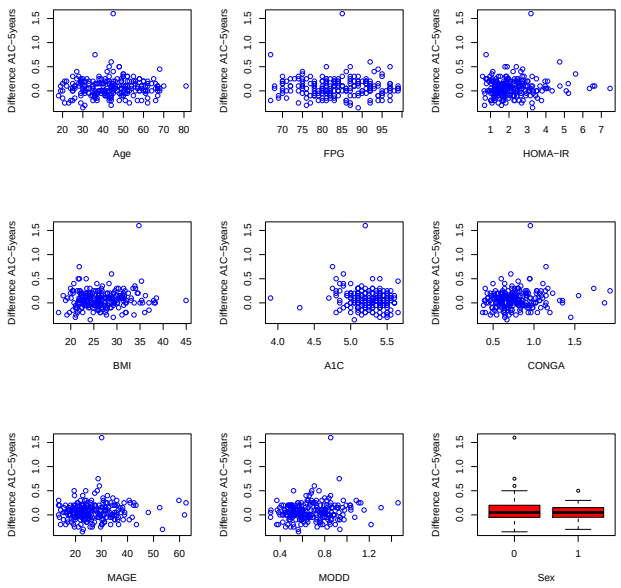

Fig. 2. Marginal dependence relation between examined variables in the AEGIS database.

Results in Table II show that the only statistically significant variables with a p-value less than 5% are glucodensity and basal A1c. Figure 2 illustrates that marginal relations with other variables, if any, are weak.

B. Is there a reduced subset of the predictors that better balances complexity and generalization ability?

Multivariate models can exploit higher-order interactions between the covariates and the response to improve the prediction of changes in A1c levels. Assuming glucodensity has already proven relevance, we adjust the method proposed in Section III-B for finding the subset of variables most strongly associated with A1c_{5years}. For this purpose, both diabetic and non-diabetic patients are analyzed, and we consider all the variables on Table I except sex. In order to avoid overfitting and to improve reproducibility of results, we select model parameters by cross-validation. We estimate the underlying missing data mechanism via lasso logistic regression.

Finally, the predictor variables selected by the algorithm are: Age, $\text{A1c}_{initial}$, FPG, BMI, and MAGE.

Variable	$p-value$
Age	0.32
Sex	0.16
FPG	0.50
HOMA-IR	0.52
BMI	0.42
A1c	0.03
CONGA	0.24
MAGE	0.68
MODD	0.16
Glucodensity	< 0.001

TABLE II
ESTIMATED RAW P-VALUES OF A1C TOTAL VARIATION VS EACH BIOMARKER USING THE ME PROPOSED IN SECTION III-A WITH NORMOGLYCEMIC PATIENTS.

C. How do CGM data impact on improving our ability to predict future A1c changes?

To answer this question, we fit two kernel ridge regression models (Section III-C) for predicting A1c_{5years}: one that includes CGM data as a covariate and another which includes MAGE measure instead. MAGE measure can be considered embedded in glucodensity. Both of them share Age, $\text{A1c}_{initial}$, FPG and BMI as covariates. Information provided by CGM is represented by glucodensity. Kernel selection and parameter tuning have been calibrated following Sections III-D and III-E. The R^2 of the first model, according to missing data mechanism and by using leave-one cross-validation, is 0.70, whereas in the second case it is 0.65. Figure 3 depicts the residuals versus $\text{A1c}_{initial}$ values. As can be seen, the highest residuals are found in diabetic patients, otherwise the distribution of residuals is heterogeneous. Ultimately, CGM data represented by glucodensity provides a piece of valuable extra knowledge in predicting long-term A1c changes.

D. Can we provide a characterization of those individuals whose future glucose behavior cannot be precisely predicted for a more personalized follow-up?

Figure 4 depicts prediction intervals at a confidence level of 90%, after applying conformal inference (Section III-C) to measure the uncertainty of the predictions performed by the above regression model (CGM data included as a covariate).

We regard a A1c_{5year} prediction as significantly affected by uncertainty if the length of the interval is greater than 0.7, since a deviation greater than this threshold can entail a change in the glycemic state of the patient, for example, from normoglycemic to diabetes. Hence, we can identify

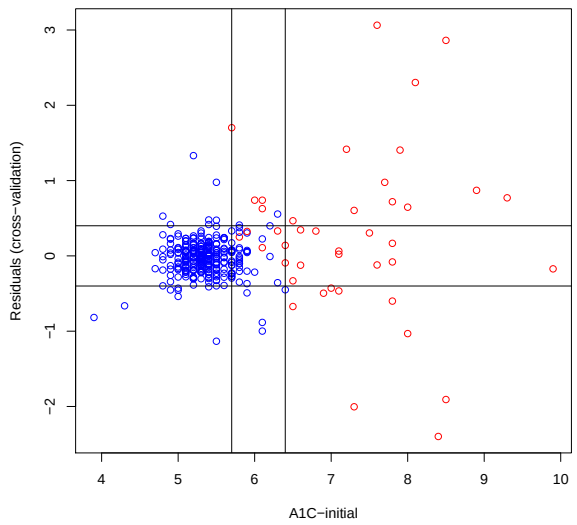

Fig. 3. Residuals vs. A1c$_{initial}$ for the model that includes glucodensity as a covariate in the AEGIS database. Red circles correspond to diabetic patients

certain clinical features that allow us to assign each patient to high or low variability groups, based on the uncertainty of future glucose values, and this can be useful to phenotipically characterize some subpopulations to whom the model provide an unreliable prediction, and therefore, a more personalized follow-up is advisable. Particularly, Figure 5 shows that, in an individual with an elevated FPG, changes in the long term cannot be adequately predicted. The same holds for individuals with FPG in the normoglycemic range and overweight. More refined decision rules can be established but at a higher measurement cost.

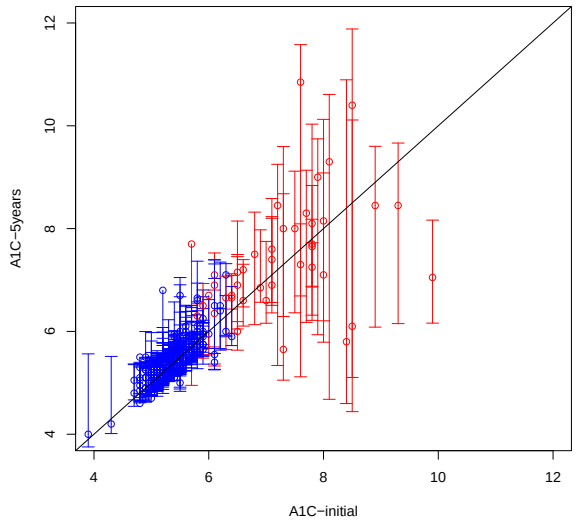

Fig. 4. Prediction intervals of regression model for each observed response of A1c$_{5year}$ in the AEGIS database (90% confidence level). Red circles correspond to diabetic patients.

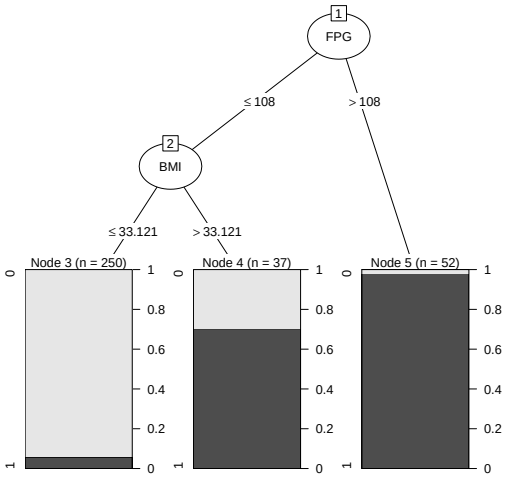

Fig. 5. Clinical decision rules that allow us to identify those patients with a significant uncertainty in their A1c$_{5year}$ predictions.

V. Discussion

The incidence and proliferation of diabetes is one of the major public health problems in the world. The present work aims at gaining new insights into the glucose metabolism and hence at supporting more informed decision-making, by studying the relationship between patient basal characteristics at the start of a longitudinal study and A1c values obtained five years later.

Some previous studies have focused on developing predictive models for patient stratification. Thus, the Finnish FINDRISC provides a diabetes score to predict the probability of developing diabetes in ten years time with a logistic regression [43]. Also, the German GDRS provides a different score to predict the time to becoming a diabetic person with a survival model based on Cox regression [44]. In contrast, some authors argue against using thresholds and categorizing patients into different ranges of levels of glucose, and hence against defining diabetes as a homogeneous disease, resulting in an oversimplistic approximation for a heterogeneous metabolic disorder [9], [45]. In this sense, some recent contributions has been made to modeling blood glucose dynamics as a function of time, with an application in predicting A1c in the short term [46], [47]. This line of research assigns a key role to the analysis of glucose excursions from CGM data, in search of a better phenotyping and a corresponding progress towards the implementation of a personalized intervention [48], [49].

The present work tries to exploit the potential of CGM data by using glucodensity as a novel representation of glucose excursions. The AEGIS study makes it possible to assess the predictive capacity of glucodensity in the context of well-known biomarkers for diabetes diagnosis and control. First, glucodensity is the only one showing a significant association with A1c changes, by using statistical dependence measures with normoglycemic patients. Still, the weak marginal association of biomarkers with A1c$_{5years}$ suggests the need for a multivariate approach to capture the complexity of long-term glucose changes. The application of a variable selection

procedure supplies us with a subset of relevant biomarkers (Age, $\mathrm{A1c}_{initial}$, FPG, BMI, and MAGE) resulting from the detection of higher-order interactions with $\mathrm{A1c}_{5years}$. We then analyze the ability to predict $\mathrm{A1c}_{5years}$ from this subset of biomarkers, with two nonparametric regression models differing on the inclusion or not of glucodensity as covariate. The R^2 value for a model including glucodensity shows a good proportion of variance explained by the model, and is similar to the one reported by other authors for short-term predictions [46], [47]. Furthermore, glucodensity demonstrates a positive impact on improving accuracy in predicting $\mathrm{A1c}_{5years}$. Ultimately, these results enforce the prominent role of CGM data to provide a comprehensive picture of the glucose metabolism [50], and allows us to envisage new research on further featuring glucose dynamics, by devising new methods for (1) measuring the variability of glucose excursion, (2) clustering different glucose profiles, or (3) discovering temporal patterns associated to pathophysiological mechanisms, among others. In this sense, further research is also needed on new glycemic outcomes, beyond average measures like A1c, in order to capture a more accurate picture of glycemic dynamics; and glucodensity can be exploited as a new source of information for more robust predictions [50].

A careful analysis of those results that exhibit significant discrepancies with the model predictions gives us the opportunity to identify certain patient phenotypes that need to be followed-up more closely. These discrepancies can be explained by many different causes (lifestyle, diet, disease, pharmacological treatments, etc.) along these five years. The present work shows that these discrepancies can be promptly recognized by using routine biomarkers of standard clinical practice. Further research is needed from the interdisciplinary cooperation between sensor technology, statistical learning, biology, pharmacology and medicine to provide a better insight into the complexity of this disorder.

VI. Conclusions

The present work proposes a data analysis framework well suited to datasets affected by missing outcome data, which are particularly common in longitudinal studies. Our approach is based on the RKHS paradigm, providing proper tools for testing statistical independence, selecting relevant variables, predicting, and making inferences about the uncertainty of predictions. The RKHS paradigm enables a nonparametric approach to these tasks, thus making few model assumptions on the relation between the response and the covariates, and allowing to capture higher-order interactions. Furthermore, RKHS provides a natural integration of multiple data modalities (functional, real-valued or categorical) into the same predictive task, supplying a powerful tool for simultaneously coping with multiple sources of information.

We have illustrated the usefulness of this approach for predicting long-term changes in the standard biomarker for glycemic control. Importantly, our analysis includes glucodensity, a novel representation of CGM data, as a predictor. Results show that CGM data provide more predictive information than previous, widely used diabetes biomarkers. Our predictive model can support clinical decision-making from the identification of patients at risk for developing diabetes or complications, when model uncertainty is low, and provide a characterization of the phenotype of patients for whom this uncertainty is significant.

Implementation

With the aim of supporting reproducible research, the source code of the methods presented in this paper has been published under an open source license[1].

References

[1] Y. Zheng, S. H. Ley, and F. B. Hu, "Global aetiology and epidemiology of type 2 diabetes mellitus and its complications," *Nature Reviews Endocrinology*, vol. 14, no. 2, pp. 88–98, 2018.

[2] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas," *Diabetes Research and Clinical Practice*, vol. 157, p. 107843, 2019.

[3] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, and B. Malanda, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 138, pp. 271–281, 2018.

[4] F. B. Hu, A. Satija, and J. E. Manson, "Curbing the diabetes pandemic: the need for global policy solutions," *JAMA*, vol. 313, no. 23, pp. 2319–2320, 2015.

[5] E. J. Topol, "Transforming medicine via digital innovation," *Science Translational Medicine*, vol. 2, no. 16, p. 16cm4, 2010.

[6] N. J. Schork, "Personalized medicine: time for one-person trials," *Nature*, vol. 520, no. 7549, pp. 609–611, 2015.

[7] M. R. Kosorok and E. B. Laber, "Precision medicine," *Annual Review of Statistics and its Application*, vol. 6, pp. 263–286, 2019.

[8] D. Cirillo and A. Valencia, "Big data analytics for personalized medicine," *Current opinion in biotechnology*, vol. 58, pp. 161–167, 2019.

[9] F. Zaccardi and K. Khunti, "Glucose dysregulation phenotypes - time to improve outcomes," *Nature Reviews Endocrinology*, vol. 14, no. 11, pp. 632–633, 2018.

[10] A. Peters, A. Ahmann, T. Battelino, A. Evert, I. Hirsch, M. Murad, W. Winter, and H. Wolpert, "Diabetes technology-continuous subcutaneous insulin infusion therapy and continuous glucose monitoring in adults: An endocrine society clinical practice guideline," *J Clin Endocrinol Metab.*, vol. 101, no. 11, pp. 3922–3937, 2016.

[11] American Diabetes Association, "7. Diabetes technology: Standards of medical care in diabetes-2019," *Diabetes Care*, vol. 42, pp. S71–S80, 2019.

[12] E. Selvin, C. M. Crainiceanu, F. L. Brancati, and J. Coresh, "Short-term variability in measures of glycemia and implications for the classification of diabetes," *Archives of internal medicine*, vol. 167, no. 14, pp. 1545–1551, 2007.

[13] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[14] B. Schölkopf and A. J. Smola, *Learning with kernels*. MIT press, 2002.

[15] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[16] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[17] F. Gude, P. Díaz-Vidal, C. Rúa-Pérez, M. Alonso-Sampedro, C. Fernández-Merino, J. Rey-García, C. Cadarso-Suárez, M. Pazos-Couselo, J. M. García-López, and A. Gonzalez-Quintela, "Glycemic variability and its association with demographics and lifestyles in a general adult population," *Journal of diabetes science and technology*, vol. 11, no. 4, pp. 780–790, 2017.

[18] M. Matabuena, A. Petersen, J. C. Vidal, and F. Gude, "Glucodensities: a new representation of glucose profiles using distributional data analysis," *Statistical Methods in Medical Research (In press)*, 2021.

[1] https://gitlab.citius.usc.es/marcos.matabuena/RKHSmissing

[19] T. Battelino, T. Danne, R. M. Bergenstal, S. A. Amiel, R. Beck, T. Biester, E. Bosi, B. A. Buckingham, W. T. Cefalu, K. L. Close *et al.*, "Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range," *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019.

[20] R. W. Beck, R. M. Bergenstal, T. D. Riddlesworth, C. Kollman, Z. Li, A. S. Brown, and K. L. Close, "Validation of time in range as an outcome measure for diabetes clinical trials," *Diabetes Care*, vol. 42, no. 3, pp. 400–405, 2019.

[21] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola, "A kernel statistical test of independence," *Advances in neural information processing systems*, vol. 20, pp. 585–592, 2007.

[22] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[23] A. Tsiatis, *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[24] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[25] S. A. Van de Geer, *Applications of empirical process theory*. Cambridge University Press Cambridge, 2000, vol. 91.

[26] M. Matabuena, P. Félix, C. Meijide-García, and F. Gude, "Five-year prediction of glucose homeostasis with missing data in a reproducing kernel hilbert space - supplementary material," Universidade de Santiago de Compostela, Tech. Rep., 2021.

[27] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[28] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," *Advances in Neural Information Processing Systems (NIPS 2017)*, vol. 30, pp. 6946–6955, 2017.

[29] L. Yang, S. Lv, and J. Wang, "Model-free variable selection in reproducing kernel hilbert space," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2885–2908, 2016.

[30] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *International conference on computational learning theory*. Springer, 2001, pp. 416–426.

[31] K. Fukumizu and C. Leng, "Gradient-based kernel method for feature extraction and variable selection," in *Advances in Neural Information Processing Systems*, ser. NIPS'12, 2012, pp. 2114–2122.

[32] T. Liu, Y. Goldberg *et al.*, "Kernel machines with missing responses," *Electronic Journal of Statistics*, vol. 14, no. 2, pp. 3766–3820, 2020.

[33] T. Liang, A. Rakhlin *et al.*, "Just interpolate: Kernel "ridgeless" regression can generalize," *Annals of Statistics*, vol. 48, no. 3, pp. 1329–1347, 2020.

[34] L. Lei and E. J. Candès, "Conformal inference of counterfactuals and individual treatment effects," *arXiv preprint arXiv:2006.06138*, 2020.

[35] J. D. Kang, J. L. Schafer *et al.*, "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical science*, vol. 22, no. 4, pp. 523–539, 2007.

[36] K. Vermeulen and S. Vansteelandt, "Bias-reduced doubly robust estimation," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 1024–1036, 2015.

[37] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[38] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[39] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semigroups: theory of positive definite and related functions*. Springer, 1984, vol. 100.

[40] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *The Annals of Statistics*, pp. 2263–2291, 2013.

[41] D. Garreau, W. Jitkrittum, and M. Kanagawa, "Large sample analysis of the median heuristic," *arXiv preprint:1707.07269*, 2017.

[42] S. J. Reddi, A. Ramdas, B. Póczos, A. Singh, and L. Wasserman, "On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions," *arXiv preprint:1406.2083*, 2014.

[43] K. Makrilakis, S. Liatis, S. Grammatikou, D. Perrea, C. Stathi, P. Tsiligros, and N. Katsilambros, "Validation of the finnish diabetes risk score (FINDRISC) questionnaire for screening for undiagnosed type 2 diabetes, dysglycaemia and the metabolic syndrome in greece," *Diabetes & Metabolism*, vol. 37, no. 2, pp. 144–151, 2011.

[44] K. Mühlenbruch, R. Paprott, H.-G. Joost, H. Boeing, C. Heidemann, and M. B. Schulze, "Derivation and external validation of a clinical version of the german diabetes risk score (GDRS) including measures of HbA1c," *BMJ Open Diabetes Research and Care*, vol. 6, no. 1, p. e000524, 2018.

[45] E. Gale, "Is type 2 diabetes a category error?" *The Lancet*, vol. 381, pp. 1956–1957, 2013.

[46] I. Gaynanova, N. Punjabi, and C. Crainiceanu, "Modeling continuous glucose monitoring (CGM) data during sleep," *Biostatistics*, 2020.

[47] A. Zaitcev, M. R. Eissa, Z. Hui, T. Good, J. Elliott, and M. Benaissa, "A deep neural network application for improved prediction of HbA1c in Type 1 diabetes," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[48] H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin, and M. Snyder, "Glucotypes reveal new patterns of glucose dysregulation," *Plos Biology*, vol. 16, no. 7, p. e2005143, 2018.

[49] A. A. Tsiatis, *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC Press, 2019.

[50] Beyond A1C Writing Group, "Need for regulatory change to incorporate beyond A1c glycemic metrics," *Diabetes Care*, vol. 41, no. 6, pp. e92–e94, 2018.