Computer Methods and Programs in Biomedicine 221 (2022) 106905

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Kernel machine learning methods to handle missing responses with complex predictors. Application in modelling five-year glucose changes using distributional representations

Marcos Matabuena[a,*], Paulo Félix[a], Carlos García-Meixide[b], Francisco Gude[c]

[a] CiTIUS (Centro Singular de Investigación en Tecnoloxías Intelixentes), Universidade de Santiago de Compostela, Santiago de Compostela 15782, Spain
[b] ETH Zürich
[c] Unidade de Epidemioloxía Clínica, Complexo Hospitalario Universidade de Santiago (CHUS), Travesía da Choupana, Santiago de Compostela 15706, Spain

ARTICLE INFO

Article history:
Received 14 October 2021
Revised 11 May 2022
Accepted 22 May 2022

Keywords:
Missing data
Kernel methods
Statistical independence
Variable selection
Regression modelling
Diabetes mellitus
Continuous glucose monitoring

ABSTRACT

Background and objectives: Missing data is a ubiquitous problem in longitudinal studies due to the number of patients lost to follow-up. Kernel methods have enriched the machine learning field by successfully managing non-vectorial predictors, such as graphs, strings, and probability distributions, and have emerged as a promising tool for the analysis of complex data stemming from modern healthcare. This paper proposes a new set of kernel methods to handle missing data in the response variables. These methods will be applied to predict long-term changes in glycated haemoglobin (A1c), the primary biomarker used to diagnose and monitor the progression of diabetes mellitus, making emphasis on exploring the predictive potential of continuous glucose monitoring (CGM).

Methods: We propose a new framework of non-linear kernel methods for testing statistical independence, selecting relevant predictors, and quantifying the uncertainty of the resultant predictive models. As a novelty in the clinical analysis, we used a distributional representation of CGM as a predictor and compared its performance with that of traditional diabetes biomarkers.

Results: The results show that, after the incorporation of CGM information, predictive ability increases from $R^2 = 0.61$ to $R^2 = 0.71$. In addition, uncertainty analysis is useful for characterising some subpopulations where predictivity is worsened, and a more personalised clinical follow-up is advisable according to expected patient uncertainty in glucose values.

Conclusions: The proposed methods have proven to deal effectively with missing data. They also have the potential to improve the results of predictive tasks by including new complex objects as explanatory variables and modelling arbitrary dependence relations. The application of these methods to a longitudinal study of diabetes showed that the inclusion of a distributional representation of CGM data provides greater sensitivity in predicting five-year A1c changes than classical diabetes biomarkers and traditional CGM metrics.

1. Introduction

Missing data are common in epidemiological studies. On the face of it, the extended practice of excluding participants with only partially available data on the variables of interest results in ignoring valuable information, thereby leading to biased estimates which often rely on unrealistic assumptions [1–3]. To draw reliable conclusions, principled methods are imperative by appropriate modeling of the missingness mechanism [4].

Kernel methods are a class of effective pattern recognition algorithms that are well suited to model nonlinear relations between the response and predictors. These are built on the notion of a kernel function as a similarity function between a new instance and those included in the training set [5,6]. One of the most significant achievements of kernel methods is the proposal of appropriate kernel functions for managing complex statistical objects such as graphs, strings, or probability distributions [7]. Thus, kernel methods are expanding the range of possible applications for

* Corresponding author.
E-mail address: marcos.matabuena@usc.es (M. Matabuena).

https://doi.org/10.1016/j.cmpb.2022.106905

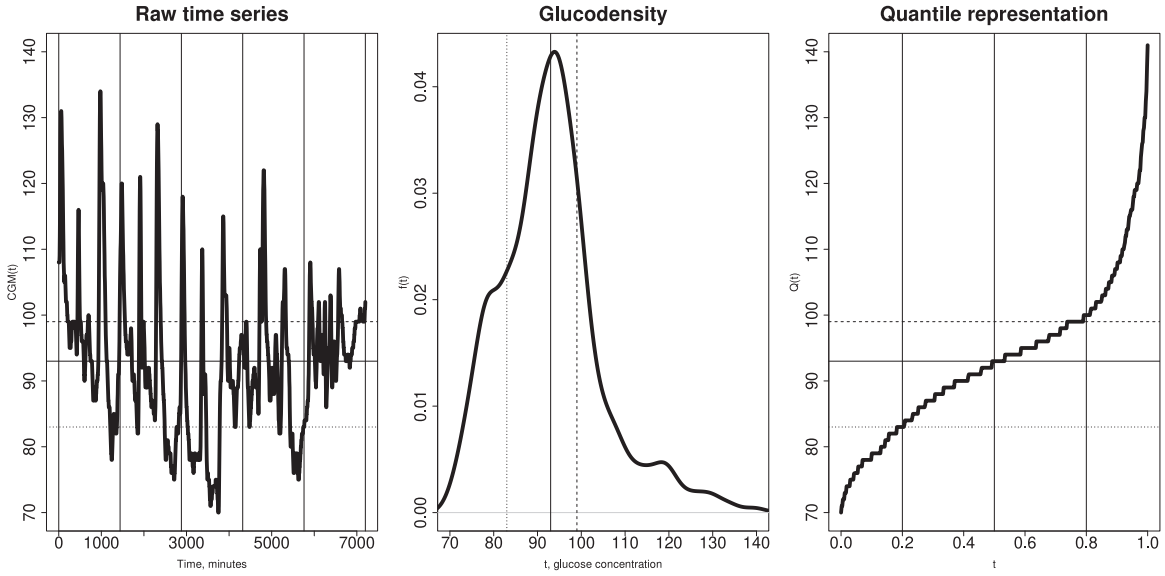

Fig. 1. Left: The 5-day CGM recording from a normoglycemic patient is shown. Centre: Glucodensity designates a distributional representation that estimates the proportion of time the patient spent at each glucose concentration. Right: Quantile representation. Dotted, solid and dashed lines represent concentrations for 20 percent, 50 percent and 80 percent quantiles, respectively.

machine learning in the health domain, challenged with the rapid increase in new complex medical data.

The main purpose of this study is to propose a set of kernel methods (Section 2) to handle missing responses for statistical independence testing (Section 2.2), variable selection (Section 2.3), and conformal inference (Section 2.4). One major advantage of these methods is their ability to operate as a sequence of predictive stages which increasingly filter out irrelevant information, while also providing an evaluation of the limits of the ensuing predictions. In particular, the present proposal is based on the reproducing kernel Hilbert space (RKHS)[1] framework, providing a Hilbert space of functions that is fully characterized by a reproducing kernel. Importantly, every function in an RKHS that minimizes an empirical risk function can be written as a linear combination of the kernel function evaluated at the training data, and it is ensured that a solution for a machine learning problem that is close to the true solution and also generalizes well to the test data can be obtained. An essential property of the RKHS framework is that it overcomes the limitations of previous proposals focused on Euclidean functional representations [8].

The proposed methods are motivated by the need to explore the limits of predicting long-term glucose changes in a five-year longitudinal population-based study, including both healthy and diabetic individuals, where a subsample of participants underwent continuous glucose monitoring procedures at the beginning of the study. As expected, a substantial number of participants withdrew from the study, and therefore, an analysis robust to missing values in the response variable is required in order to maintain the validity of the statistical inferences [9]. We include a novel distributional representation for CGM data as a predictor (see Fig. 1) [10]. Among the different biomarkers, we select the glycated haemoglobin (A1c) as the response variable. A1c is a measure of the average blood glucose level over the past three months, and it is the preferred option because it provides more reproducible values in the laboratory and is subject to less measurement error [11]. Furthermore, we aimed to assess and discuss the residuals and predictive capacity of several variables associated with the evolution of A1c in the long term, providing interpretable clinical phenotypes for large uncertainty cases.

[1] Appendix contains a guide to acronyms used in this document.

The rest of this paper is outlined as follows. Section 2 describes in detail the methods for statistical independence testing, variable selection, and inference on the uncertainty of new predictions. Section 3 presents the application of these methods for modeling long-term changes in glucose. Section 3.1 describes the AEGIS database used to test the proposed method. Section 4 presents the results of the application of these methods to the AEGIS database. Section 5 discusses the advantages and drawbacks of the proposed approach. Finally, conclusions are presented in Section 6.

1.1. Data analysis outline

Finally, this paper presents a data analysis framework designed as a pipeline of kernel methods for predictive problems with missing data. Their subsequent application to diabetes mellitus will allow us to examine the relationship between the baseline characteristics of participants in a five-year study and A1c as the response variable. The proposed framework comprises the following steps:

1. To measure the statistical association between each predictor and the response variable with an efficient statistical independence test. If the response is proven to be independent of a predictor, it can be screened out from further consideration. To this end, we adapted a previous kernel independence test and designed a new bootstrap method to perform test calibration. The test was applied to check the association between some diabetes biomarkers and five-year changes in the A1c variable, $A1c_{5years} - A1c_{initial}$.
2. To identify the best subset of predictors revealing higher-order interactions with the response variable in order to improve the prediction. To this end, we adapted a previous kernel variable selection method and applied it to find the best subset of diabetes biomarkers most strongly associated with $A1c_{5years}$.
3. To explore the prediction ability of a set of explanatory variables through a non-linear regression method. To this end, we adapted a previous kernel ridge regression method and applied it to predict $A1c_{5years}$.
4. To estimate the uncertainty of the predictions. To this end, we designed a new method to provide a prediction interval for the response variable, based on conformal inference. Using this method, we can measure the limits of the regression models previously obtained and, significantly, identify specific patient

subpopulations that do not fit the expected behaviour, which is a key issue for clinical decision-making.

2. Methods

2.1. Preliminaries

We first pose the problem in general terms. Let $(\mathbf{X}, Y, R) \in \mathcal{X} \times \mathbb{R} \times \{0,1\}$ be a random vector such that $\mathbf{X} = (X^1, \ldots, X^p)$ denotes the predictor variables, Y is the response variable, and R is a binary random variable that indicates whether the response is missing. $\mathcal{X}$ denotes a general topological space, meaning that it can be arbitrary, discrete, continuous, or structured.

Let $D = \{(\mathbf{x}_i, y_i, r_i)\}_{i=1}^n$ be a dataset of independent, identically distributed observations, where y_i is missing if $r_i = 0$. We assume R is distributed according to $R|X \sim Ber(\pi(X))$, with $\pi(\cdot) = P(R = 1|\mathbf{X} = \cdot)$; hence, some of the predictors can have an impact on the mechanism of missing data $\pi(\cdot) = P(R = 1|\mathbf{X} = \cdot)$. For instance, in our example, older patients are more reluctant to perform a second CGM monitoring, so the probability of not observing a patient increases with age. We also assume a missing at random (MAR) mechanism in the response Y, namely, R and Y are conditionally independent given $\mathbf{X}$ or, in short, $R \perp Y|\mathbf{X}$.

Consider the following relation between $\mathbf{X}$ and Y:

$$Y = f(\mathbf{X}) + \epsilon, \tag{1}$$

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

These methods are based on the reproducing kernel Hilbert space (RKHS) learning paradigm. The core element of this paradigm is a positive definite kernel function $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which allows us to measure the similarity between any x and y, with $x, y \in \mathcal{X}$. The positive definiteness of the kernel function guarantees the existence of a dot product space $\mathcal{H}$ and feature mapping $\phi : \mathcal{X} \to \mathcal{H}$, such that $k_{\mathcal{X}}(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. Thus, we can express a broad spectrum of statistical modeling problems as linear [12], allowing computational algorithms to easily determine optimal solutions.

2.2. Testing statistical independence

We wish to test whether two random variables $X \sim P_X$ and $Y \sim P_Y$ are not independent, that is, if we can reject the null hypotheses $H_0 : X \perp\!\!\!\perp Y$, from n samples $\{(x_i, y_i)\}_{i=1}^n$. To do this, we must calibrate the test under the null hypothesis to determine the results that are expected to occur with a certain probability if the null hypothesis holds. In our specific case, we must consider the effects of the mechanism of missing data in the response variable Y. We propose a methodology to address 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 the test calibration adapted to kernel mean embeddings.

Hilbert space embeddings of distributions or, in short, kernel mean embeddings [7], allow us to map distributions into a 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), which in turn 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 [13]. We extended this test to a missing data setting.

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 an RKHS, $\mathcal{H}_{k_{\mathcal{X}}}$. Kernel mean embedding results from extending the mapping ϕ to the space of probability distributions by representing each distribution as a mean function $\phi(P) = \int_{\mathcal{X}} k(\cdot, x) dP(x)$, resulting in the transformation of a distribution P into an element of the RKHS $\mathcal{H}_{k_{\mathcal{X}}}$. Given two probability measures P and Q, the RKHS distance between their embeddings can be defined as the MMD [14]:

$$\mathrm{MMD}_{k_{\mathcal{X}}}(P, Q) = \|\phi(P) - \phi(Q)\|_{\mathcal{H}_{k_{\mathcal{X}}}}. \tag{2}$$

For the class of *characteristic* kernels, the embeddings are injective, that is, $\mathrm{MMD}_k(P, Q) = 0$, if and only if $P = Q$. MMD can then be applied to measure 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 $P_{X,Y}$. 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 $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}$ (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 [14], defined as

$$\mathrm{HSIC}(P_{X,Y}, P_X P_Y) = ||\phi_{X,Y} - \phi_X \otimes \phi_Y||^2_{\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}} \tag{3}$$

It can be shown that when $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are characteristic kernels, $\mathrm{HSIC}(P_{X,Y}, P_X P_Y) = 0$ if and only if $X \perp\!\!\!\perp Y$. Expanding Eq. (3), we have

$$\begin{aligned} \mathrm{HSIC}(P_{X,Y}, P_X P_Y) &= \langle \phi_{X,Y} - \phi_X \otimes \phi_Y, \phi_{X,Y} - \phi_X \otimes \phi_Y \rangle_{\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}} \\ &= \langle \phi_{X,Y}, \phi_{X,Y} \rangle + \langle \phi_X \otimes \phi_Y, \phi_X \otimes \phi_Y \rangle \\ &\quad -2 \langle \phi_{X,Y}, \phi_X \otimes \phi_Y \rangle, \end{aligned} \tag{4}$$

where $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}$ is dropped in the subscript for brevity. From 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 obtain

$$\begin{aligned} \mathrm{HSIC}(P_{X,Y}, P_X P_Y) &= \mathbb{E}_{X,Y,X',Y'}[k_{\mathcal{X}}(X, X') k_{\mathcal{Y}}(Y, Y')] \\ &\quad + \mathbb{E}_{X,X'}[k_{\mathcal{X}}(X, X')] \mathbb{E}_{Y,Y'}[k_{\mathcal{Y}}(Y, Y')] - \\ &\quad - 2\mathbb{E}_{X,Y}[\mathbb{E}_{X'}(k_{\mathcal{X}}(X, X')) \mathbb{E}_{Y'}[k_{\mathcal{Y}}(Y, Y')]], \end{aligned} \tag{5}$$

where X' and Y' are independent copies of random variables X and Y, respectively. Ultimately, testing independence involves calculating the squared distance between two mean functions in the appropriate RKHS space, resulting from transforming the original data 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 with the sample mean, defined through its empirical distribution. Then, the Hilbert-Schmidt independence criterion can be estimated as

$$\begin{aligned} \widehat{\mathrm{HSIC}}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_{\mathcal{X}}(x_i, x_j) k_{\mathcal{Y}}(y_i, y_j) \\ &\quad + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_{\mathcal{X}}(x_i, x_j) \sum_{i=1}^n \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_{\mathcal{Y}}(y_i, y_j) \\ &\quad - \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n k_{\mathcal{X}}(x_i, x_j) k_{\mathcal{Y}}(y_i, y_k). \end{aligned} \tag{6}$$

Under the MAR assumption, we observe $\{(x_i, y_i, r_i)\}_{i=1}^n$, and we must 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 purpose, such as logistic regression, lasso, random forest, and ensemble methods. Subsequently, we re-weighted the dataset, taking into account the difficulty of observing the response of the i^{th} datum. In particular, we associate weight w_i with the i^{th} datum via an inverse probability weighting (IPW) estimator [4] given by

$$w_i = \frac{r_i}{n\pi(x_i)}, \quad i = 1, \dots, n, \tag{7}$$

which results in assigning large w_i values as the probability of observing a response decreases. Using this procedure, we obtain an asymptotic unbiased estimator that balances the sampling mechanism and allows us to make a proper inference according to the target population examined.

We define the normalized weight of w_i as

$$w_i^* = \frac{w_i}{\sum_{i=1}^n w_i}, \quad i = 1, \dots, n. \tag{8}$$

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

To obtain an estimator of HSIC with missing data, it is sufficient to replace the uniform weight $1/n$ of the empirical distribution with normalised weights $\hat{W}^* = (\hat{w}_1^*, \dots, \hat{w}_n^*)$ in Eq. (6). Thus, we obtain

$$\begin{aligned}\widehat{\mathrm{HSIC}}(\hat{P}_{X,Y}, \hat{P}_X\hat{P}_Y) = &\sum_{i=1}^n\sum_{j=1}^n \hat{w}_i^*\hat{w}_j^* k_{\mathcal{X}}(x_i, x_j)k_{\mathcal{Y}}(y_i, y_j) \\ &+ \sum_{i=1}^n\sum_{j=1}^n \hat{w}_i^*\hat{w}_j^* k_{\mathcal{X}}(x_i, x_j) \sum_{i=1}^n\sum_{j=1}^n \hat{w}_i^*\hat{w}_j^* k_{\mathcal{Y}}(y_i, y_j) \\ &- \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \hat{w}_i^*\hat{w}_j^*\hat{w}_k^* k_{\mathcal{X}}(x_i, x_j)k_{\mathcal{Y}}(y_i, y_k). \end{aligned} \tag{9}$$

Calibration under the null hypothesis with the precedent statistic is not trivial, and the permutation approach is generally not valid because the response Y is not exchangeable due to the non-homogeneous missing data mechanism. To overcome this difficulty, we propose a novel bootstrap approach that properly deals with non-vectorial predictors [15].

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

$$\begin{aligned}\widehat{\mathrm{HSIC}}(\hat{P}_{X,Y}, \hat{P}_X\hat{P}_Y) &= \langle \hat{\phi}_{X,Y} - \hat{\phi}_X \otimes \hat{\phi}_Y, \hat{\phi}_{X,Y} - \hat{\phi}_X \otimes \hat{\phi}_Y \rangle_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} \\ &= \langle \hat{\phi}_{X,Y} - \phi_{X,Y} + \phi_X \otimes \phi_Y - \hat{\phi}_X \otimes \hat{\phi}_Y, \hat{\phi}_{X,Y} \\ &\quad - \phi_{X,Y} + \phi_X \otimes \phi_Y - \hat{\phi}_X \otimes \hat{\phi}_Y \rangle. \end{aligned} \tag{10}$$

Then, a natural bootstrap procedure that allows us to estimate the p-value for the independence test is 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

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

where $j = 1, \dots, m$, $\hat{\phi}_{X,Y}^{j^*}(\cdot)$, $\hat{\phi}_X^{j^*}(\cdot)$, 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

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

Bootstrap consistency with missing data can be proved by using standard tools of empirical process theory [16], and it is provided elsewhere.

2.3. Variable selection

Independence screening methods select predictor variables based on individual prediction ability; hence, they are ineffective in selecting a subset of variables that are individually weak but strong in combination. 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 to directly optimize 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 [17].

Subset selection has recently been approached using the RKHS paradigm with satisfactory results. Two strategies stand out: first, minimizing the trace of the conditional covariance operator [18] and second, identifying those variables with a non-zero gradient function [19]. The first strategy scales poorly with the number of variables used. The second strategy can be formulated in a more compact manner. Here, it is extended to missing data.

Following [19], we identify the relevant predictors by learning the gradient of regression function f. Thus, it is assumed that if 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 $\mathbf{g}(\mathbf{X}) = \nabla f(\mathbf{X}) = (g_1(\mathbf{X}), \dots, g_p(\mathbf{X}))^T$ the gradient function. In a small neighborhood of $\mathbf{x}_i$ we can use the Taylor expansion to approximate $f(\mathbf{X})$, so when $\mathbf{x}_j$ is sufficiently close to $\mathbf{x}_i$, $f(\mathbf{x}_j) \approx y_i + g(\mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)$. We then 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 $\mathbf{X}', Y'$ denote independent and random variables distributed as $\mathbf{X}$ and Y, respectively. Function $\omega(\mathbf{x}_i, \mathbf{x}_j)$ is an appropriate weight function that decreases as $\|\mathbf{x}_i - \mathbf{x}_j\|$ increases and ensures that the local neighbourhood of $\mathbf{x}_i$ contributes more to estimating the gradient $\mathbf{g}(\mathbf{x}_i)$. Typically, $\omega(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\tau_n^2}$, where τ_n^2 is a positive parameter which must be adjusted to ensure asymptotic estimation consistency.

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

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

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 the partial derivatives g_r towards zero 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 $\mathcal{H}_k^p$, and thus $\mathbf{g} \in \mathcal{H}_k^p$ 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. Thus, we propose the following optimization formula to learn the gradient vector:

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

Under the MAR assumption, we propose substituting ω_{ij} weights with $\hat{\omega}^*_{ij} = \hat{\omega}^*_i \hat{\omega}^*_j \omega_{ij}$, where $\hat{\omega}^*_i$ and $\hat{\omega}^*_j$ denote the estimated normalized weights associated with data i^{th} and j^{th}, respectively, according to (8). The variable selection expression can be rewritten as follows:

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

The representer theorem states that the minimizer of (15) can be represented as a finite linear combination of kernel products evaluated on the dataset samples [20]:

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

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

Several regularization terms have been considered in previous studies. We adopted the group lasso penalty [19,21]:

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

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

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

where $f^*(\mathbf{x}_i, \mathbf{x}_j) = y_j - \sum_{r=1}^{p} \mathbf{k}_i^T \alpha^r (x_i^r - x_j^r)$, $\mathbf{k}_i = (k(\mathbf{x}_i, \mathbf{x}_1), \ldots, k(\mathbf{x}_i, \mathbf{x}_n))^T$ is the i^{th} row of $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$, and λ_n are tuning parameters. This last expression simplifies the original optimization framework (14) from a functional space to a vector space, and it can be solved in $O(|U|^2 p^2)$ using a block coordinate descent algorithm [19].

2.4. Prediction and uncertainty analysis

Let us recall that the ultimate goal is to predict Y using the information provided by predictor variables $\mathbf{X}$. To achieve this aim, we adopt the kernel ridge regression approach proposed by Liu and Goldberg [22]. However, we draw on the linear regression theory to efficiently compute the leave-one-out cross-validation regularization parameter. This class of regularization parameters has been proven to largely shape the model performance [23]. Furthermore, estimating the uncertainty of predictions by providing robust confidence intervals is a valuable tool for subsequent decisions. Thus, we compute intervals with good finite sample coverage using advances in conformal inference recently exploited in causal theory [24].

Let us assume a linear regression model:

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

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:

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

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 the regularization term.

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

In [22], the authors proposed two estimators for the missing data. In both cases, the solution has the same closed-form expression, given by the representer theorem. The first is $\hat{\alpha} = (\lambda\mathbf{I} + \mathbf{W})^{-1}\mathbf{W}\mathbf{y}$, where the missing data mechanism is handled using the IPW estimator. The second is obtained through doubly robust estimation, combining a preliminary imputation of the missing response with the 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 Eq. (7)) 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 $w_1, \ldots, w_n$ and 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 a finite sample [25,26], thereby failing to provide real advantages with respect to the IPW estimator.

The impact of the smoothing parameter on model generalization is an essential issue for the ensuing performance and is strongly related to the minimum-norm interpolation problem in the context of RKHS. Therefore, we propose the selection of the smoothing parameter through *leave-one-out* cross-validation by adapting estimators to missing data [23].

To supply a prediction interval for the response with a confidence level of $1 - \alpha$, we provide a novel algorithm for performing conformal inference [24,27], which is valid for handling missing responses and heteroscedastic noise.

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 $D^{test} = \{(\mathbf{x}_i^{test}, y_i^{test}, r_i^{test})\}_{i=1}^{n_2}$, where $n = n_1 + n_2$.

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 Eq. (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 $r_i^{test} = 1$. The value $\hat{\sigma}(x_i^{test})$ is estimated by a regression function that predicts the absolute deviation of the residuals fitted with the training sample.
3. Estimate the empirical distribution as follows:

$$\hat{F}^{\epsilon}_{n_2+1}(x) = \frac{1}{\sum_{i=1}^{n_{2+1}} \hat{w}_i^{test}} \left(\sum_{i=1}^{n_2} 1\{\hat{\epsilon}_i \le x\} \hat{w}_i^{test} + \hat{w}^{test}_{n_{2+1}} \right), \tag{22}$$

   where we also incorporate the weights of $\mathbf{x}_{n+1}$ and $\hat{w}^{test}_{n_{2+1}}$ into the estimation.
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.

3. An application in modelling long-term changes in glucose levels

Diabetes mellitus is one of the most critical public health problems and the ninth major cause of mortality worldwide [28]. At present, over 416 million and 47 million patients have type 2 (T2D) and type 1 (T1D) diabetes, respectively, [29]. Significantly, around 50% of patients with diabetes are undiagnosed [29]. 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 at-risk patients and improvement in the methodology of disease diagnosis in the standard clinical routine [30].

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 within ten years using logistic regression [31]. In addition, the German GDRS provides a different score to predict the time to becoming a diabetic person using a survival model based on Cox regression [32]. In contrast, some authors argue against using thresholds and categorising patients into different ranges of glucose levels, and hence, against defining diabetes as a homogeneous disease [33].

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 raise early diagnosis and guide subsequent clinical decision-making through the intensive use of statistical models and machine learning techniques [34–37]. In the case of diabetes, the latest advances in sensing technology allow for the assessment of glucose metabolism at a high-resolution level by capturing the individual differences in glucose fluctuations at different time scales via continuous glucose monitoring (CGM) [38]. In this sense, although T1D cannot be prevented at present, monitoring is of utmost importance. Recent studies have shown improved glycemic control and decreased rates of hypoglycaemia in T1D patients using CGM, leading both the Endocrine Society and the American Diabetes Association to state that CGM use represents the standard of care for T1D [39,40]. With respect to T2D, strong scientific evidence shows that it can be prevented by regular exercise, healthy eating, and the control of blood pressure and lipids [41], spurring innovation in wearable technology to enable its prediction and prevention in the general population [42].

Still, few studies have explored the use of CGM data from healthy populations to draw new conclusions regarding glucose homeostasis. It is worth mentioning [43], which provides some remarkable insights into the heterogeneity of glucose dysregulation, highlighting the inadequacy of a common designation as T2D for categorising and subsequently managing predictably different conditions. Importantly, this study refutes the assumption of similar glucose excursions for the same amount and composition of food. Ultimately, the specific glucose profiles observed for each patient depend on the complex interplay between the pathophysiological mechanisms of insulin resistance and insulin secretion [38,43], thus enabling the treatment of the glycemic profile of an individual as a personal signature of glucose homeostasis. Accordingly, an appropriate interpretation of CGM data could help identify early stages of glucose dysregulation in apparently healthy individuals, with the possibility of providing early and tailored interventions. In this sense, there is a need for further research on the predictive value of CGM data [38].

This study aimed to examine the predictability of long-term changes in glucose levels by using a random sample of the general population. The exploration of the predictive value of the information provided by CGM data is of particular interest. For this purpose, we consider a new distributional representation and compare the corresponding model performance with well-established biomarkers for the diagnosis and management of diabetes.

3.1. The AEGIS diabetes study

The AEGIS diabetes population study, conducted in the Spanish town of A Estrada (Galicia), aimed to analyze the longitudinal changes in some clinical features related to circulating glucose in 1516 patients over 5 years. In addition, non-routine medical tests, such as CGM, are performed every five years on a randomized subset composed of 581 patients. At the beginning of this study [44], 581 participants were randomly selected to wear a CGM device for $3-7$ days. Of the 581 participants, 68 were diagnosed with diabetes before the start of the study and 22 were diagnosed during the study. Table 1 lists the baseline characteristics of the 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 in 339 patients. Complete details of the study design and measurement methodology protocol can be found in [44].

Table 1
Characteristics of AEGIS study participants with CGM monitoring by sex. Means and standard deviations are shown. A1c: glycated haemoglobin; 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 difference.

	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.μIU/mL	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

3.2. A distributional representation for CGM data: Glucodensity

We adopted a novel functional representation for CGM data, termed glucodensity, which allows us to obtain a personalized functional profile of patient glucose homeostasis [10]. Glucodensity is a natural extension of those metrics based on the time spent in certain ranges. Time in range (TIR) measures the proportion of time a person spends with their blood glucose levels within the target range of $70-180$ mg/dL. Accordingly, the time below range (TBR) and time above range (TAR) are defined. These are the current gold standard for representing CGM data [45,46]. Although very intuitive, they have two main disadvantages: first, the range fits poorly depending on the characteristics of the population examined; second, there is a loss of information caused by the discretization of the recorded data into intervals. Instead, glucodensity effectively measures the proportion of time each individual spends at a specific glucose concentration.

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 kernel density estimation:

$$\hat{f}(x) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{h}k\Big(\frac{x-x_j}{h}\Big), \tag{23}$$

where $h > 0$ is the smoothing parameter and $k(\cdot)$ denotes a non-negative real-valued integrable function (Fig. 2).

A relevant issue in kernel analysis is to measure the difference between two density functions. In this study, we use the 2-Wasserstein distance. Given two glucodensities, $\hat{f}_1$ and $\hat{f}_2$, the 2-Wasserstein distance between them is given by

$$d_{\mathcal{W}_2}(\hat{f}_1, \hat{f}_2) = \sqrt{\int_0^1 \left|\hat{Q}_{f_1}(t) - \hat{Q}_{f_2}(t)\right|^2 dt}, \tag{24}$$

where $\hat{Q}_{f_1}$, and $\hat{Q}_{f_2}$ are the corresponding quantile functions. Because the 2-Wassertein distance between two densities depends only on their quantile functions, it is not necessary to resort to

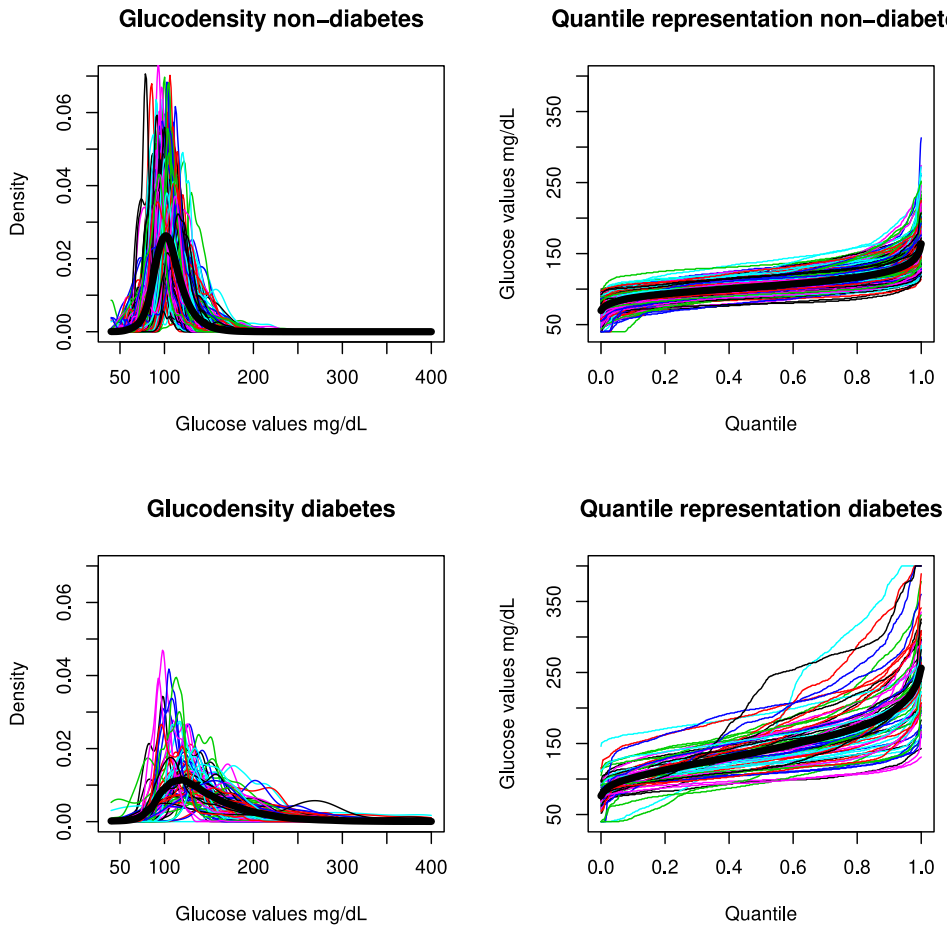

Fig. 2. Glucodensities estimated from a random sample of the AEGIS study on diabetic and normoglycemic patients. Left: Representation of the proportion of time spent by a patient at each glucose concentration over a continuum is shown. Right: Representation of the glucodensities in the space of quantile functions is shown. A meaningful difference between the average profiles for both groups of patients can be noticed.

density estimation methods, and we can approximate this distance using quantile-function estimations through empirical distributions.

Intuitively, glucodensity is more sensitive than the previous CGM summary metrics. We then explored its use in modeling long-term glucose changes and compared it with TBR, TIR, and TAR measures. In addition, we also considered different summary metrics derived from CGM data [47,48]: CONGA (continuous overall net glycemic action), MAGE (mean amplitude of glycemic excursions), and MODD (mean of daily differences).

3.3. Integrating multiple data sources

RKHS offers a powerful and natural data analysis paradigm that can cope with data of different natures [49]. A crucial issue is to select a suitable kernel that accurately captures the differences and specific characteristics of each information source 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 $k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||}{\sigma}}$. Here, we propose using the Laplacian kernel with the standard Euclidean distance as a universal and characteristic 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 2−Wasserstein distance, providing theoretical guarantees that we can approximate a large variety of regression functions. Based on the connection between positive kernels and negative-type metrics [50,51], we propose using a simple and global Laplacian kernel that integrates these three sources:

$$k_{\mathcal{X}}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\left(a\frac{||x_i^{gluco} - x_j^{gluco}||}{\sigma_{gluco}} + b\frac{||\mathbf{x}_i^{real} - \mathbf{x}_j^{real}||}{\sigma_{real}} + c\frac{||\mathbf{x}_i^{categ} - \mathbf{x}_j^{categ}||}{\sigma_{categ}}\right)}, \quad (25)$$

where $a, b, c, \sigma_{gluco}, \sigma_{real}, \sigma_{categ} > 0$ and we assume for the sake of simplicity that $(a, b, c) \in \mathcal{S}^2$, where $\mathcal{S}^2$ is a 2-simplex, $\mathcal{S}^2 = \{(a, b, c) \in \mathbb{R}^3 : a + b + c = 1; 0 \le a \le 1, 0 \le b \le 1, 0 \le c \le 1\}$.

4. Results

The present framework of predictive tools allows us to answer some open clinical questions concerning long-term glucose changes from the analysis of data in the AEGIS study.

1. Glycated haemoglobin A1c is a haemoglobin-glucose combination formed within the cell; it is a useful indicator of long-term blood glucose control and is considered the standard biomarker for diabetes diagnosis and management. *Is there a prognostic variable that can be used to predict future A1c changes in healthy individuals?*
2. Current medical literature assigns a considerable relevance to all of the predictor variables listed in Table 1 for characterizing the evolution and impact of glucose homeostasis on health. However, from a biological perspective, these variables are well known to be highly correlated. *Can we identify a reduced subset of relevant explanatory variables to predict five-year A1c changes?*

Table 2
Estimated raw p-values of A1c total variation vs each biomarker using the method proposed in Section 2.2 with normoglycemic patients.

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

3. CGM technology may provide a more suitable tool for assessing glucose homeostasis than traditional diabetes biomarkers. *How does CGM data improve our ability to predict future A1c changes?*
4. An increased uncertainty in predictions for a specific region of the feature space may suggest a subpopulation that has not been properly modeled. *Can we provide a characterization of individuals for whom the model yields a less accurate prediction?*

4.1. Is there a prognostic variable that can be used to predict future A1c changes in healthy individuals?

To answer this question, we studied whether there was any evidence of a univariate statistical association for normoglycemic patients (A1c<5.7% and FPG<100 mg/dL) between glucose variation measured by $A1c_{5years}-A1c_{initial}$ and the predictor variables shown in Table 1.

For this purpose, we use the Hilbert-Schmidt independence criterion proposed in the context of missing data (Section 2.2) together with a specific bootstrap approach designed for this task. The underlying mechanism of missing data was estimated using univariate logistic regression.

The results in Table 2 show that the only statistically significant variables with a p-value of less than 5% are glucodensity and basal A1c. Fig. 3 illustrates that the marginal relationships with other variables, if any, are weak.

4.2. Can we identify a reduced subset of relevant variables to predict five-year A1c changes?

Multivariate models can exploit higher-order interactions between the predictors and the response to improve predictions. However, a key point in increasing the interpretability and generalization ability of the model is to identify a subset of the variables that capture the essential information in the dataset, thus removing redundancy. We adjusted the method proposed in Section 2.3 to find the subset of variables most strongly associated with $A1c_{5years}$. For this purpose, both diabetic and non-diabetic patients were analysed, and we considered all the variables in Table 1 except for sex. We also included the TBR, TIR, and TAR measures specified in Section 3.2. To avoid overfitting and improve the reproducibility of the results, we selected model parameters by cross-validation. We estimated the underlying missing data mechanism using lasso logistic regression.

Finally, the explanatory variables selected by the algorithm were age, $A1c_{initial}$, FPG, BMI, and MAGE. Notably, the CGM contribution is made through the specific MAGE index, leaving aside time in ranges.

4.3. How does CGM data improve our ability to predict future A1c changes?

To answer this question, we fit several kernel ridge regression models (Section 2.4) for predicting $A1c_{5years}$: 1) excluding CGM data as a predictor; 2) including CGM data through the MAGE index; 3) including CGM data through the above-mentioned time in ranges; and 4) including CGM data through glucodensity representation. Both share age, $A1c_{initial}$, FPG, and BMI as covariates. The kernel selection and parameter tuning were calibrated as described in Section 3.3.

To compare the performance of these regression models, we used R^2 after including the specific missing data mechanism:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \frac{\sum_{j=1,j\neq i}^{n} w_j y_j}{\sum_{j=1,j\neq i}^{n} w_j}\right)^2}{\sum_{i=1}^{n}\left(y_i - \hat{f}_{-i}(x_i)\right)^2}, \tag{26}$$

where $\hat{f}_{-i}(\cdot)$, is the regression function fitted to $\{(x_j, y_j, r_j)\}_{j\neq i}^{n}$, i.e. excluding the i^{th}-datum.

The performance results, obtained using leave-one-out cross-validation, are: 1) R^2_{noCGM}=0.61; 2) R^2_{MAGE}=0.65, 3) R^2_{TIR}=0.64; and 4) R^2_{gluco}=0.71. For comparison purposes, partial least squares regression has also been applied, by using leave-one-out cross-validation, obtaining the following results: 1) R^2_{noCGM}=0.52; 2) R^2_{MAGE}=0.57; and 3) R^2_{TIR}=0.56. Fig. 4 depicts the residuals versus the $A1c_{initial}$ values for the model including glucodensities. As can be seen, the highest residuals are found in diabetic patients; besides that, their distribution is heterogeneous. As expected, the prediabetic range is the most challenging to model. From a total of 60 prediabetic (according to A1c criteria: 5.7% $\geq$ A1c $\geq$ 6.4%) individuals at baseline, 22 of them were diagnosed with diabetes (A1c $\geq$ 6.5% or FPG $\geq$ 126 mg/dL)[2] five years later, 14 of them with A1c $\geq$ 6.5%; 5 of the latter were correctly predicted by our models and 9 were wrongly predicted. Because of the limited number of individuals in this range any extrapolation needs to be restrained. Ultimately, CGM data represented by glucodensities provide valuable information for predicting long-term A1c changes.

4.4. Can we provide a characterization of individuals for whom the model yields a less accurate prediction?

Fig. 5 depicts prediction intervals at a confidence level of 90%, after applying conformal inference (Section 2.4) to measure the uncertainty of the predictions performed by the above regression model (CGM data included as a covariate).

We regard an $A1c_{5year}$ prediction as significantly affected by uncertainty if the length of the interval is greater than 0.7 because 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 certain clinical features that allow us to assign each patient to high-or low-variability groups based on the uncertainty of future glucose values. This can be useful to phenotypically characterize some subpopulations for which the model provides an unreliable prediction, and therefore, a more personalised follow-up is advisable. In particular, Fig. 6 shows that long-term changes cannot be adequately predicted for individuals with elevated FPG levels. The same holds true for individuals with FPG levels in the normoglycemic range and overweight. More refined decision rules can be established at higher measurement costs.

[2] Diagnostic criteria according to American Diabetes Association.

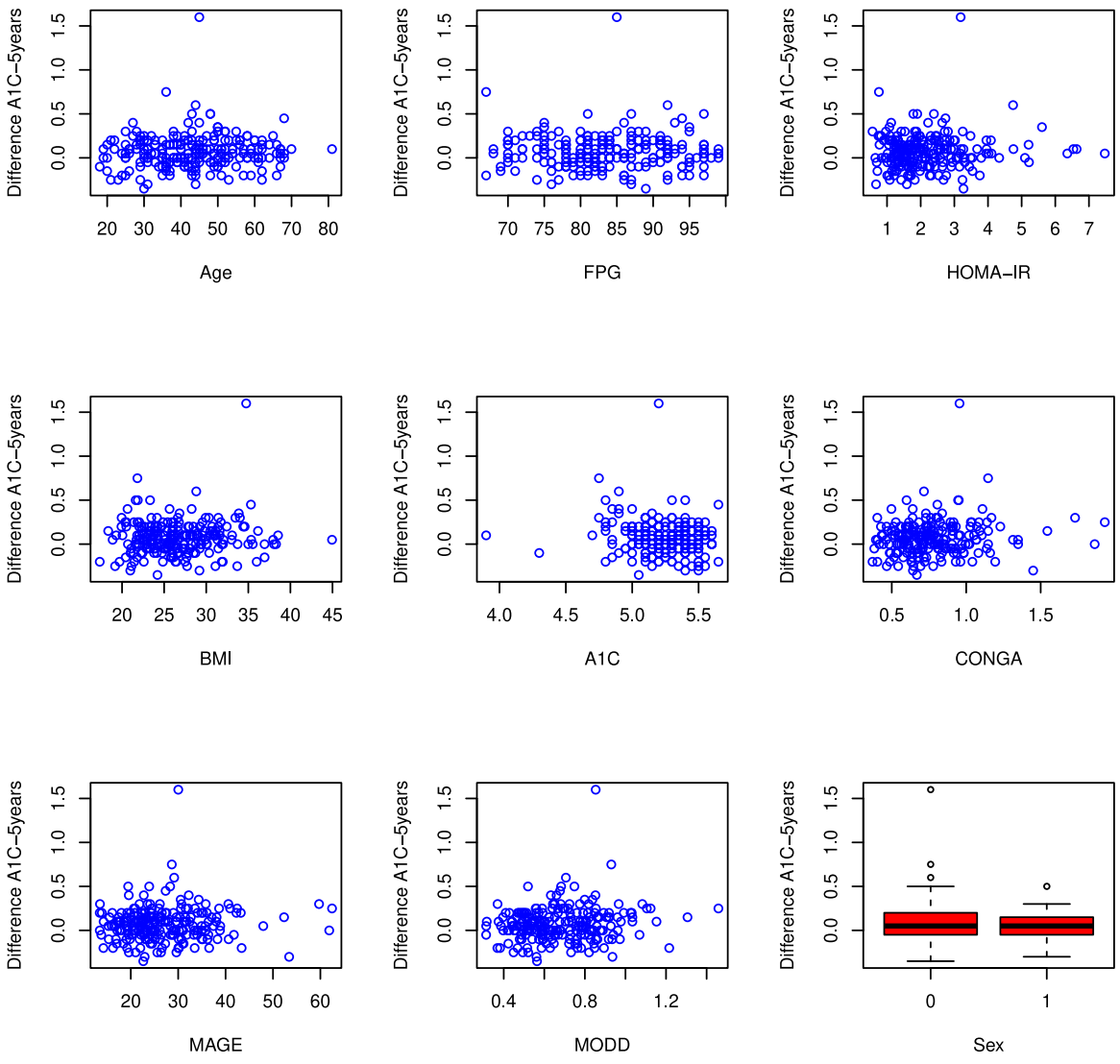

Fig. 3. Marginal dependence relation between examined variables in the AEGIS database.

5. Discussion

The above analysis aimed to explore the predictability of glucose regulation in the general population by studying the relationship between patient basal characteristics at the start of a longitudinal study and A1c values obtained five years later. Specifically, we intend to exploit the ability of CGM data to effectively capture a personal signature of glucose homeostasis through the inclusion of glucodensity, a novel distributional representation of glucose excursions, as a predictor.

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, providing some interesting findings. First, glucodensity shows a significant association with A1c changes, using statistical dependence measures in normoglycemic subjects. Nevertheless, 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 a subset of relevant biomarkers (age, $A1c_{initial}$, FPG, BMI, and MAGE) resulting from the detection of higher-order interactions with $A1c_{5years}$. Then, the ability to predict $A1c_{5years}$ from this subset of biomarkers is analysed using several regression models that differ in terms of including CGM data as a predictor. As a result, the R^2_{gluco} value, corresponding to the model which adopts a glucodensity-based representation for CGM data, shows a good proportion of variance explained by the model and is similar to that reported by other authors for short-term predictions [52,53]. Moreover, glucodensity has a positive impact on improving accuracy in predicting $A1c_{5years}$ by expanding the model expressiveness along the continuous spectrum of glucose concentrations.

Some recent studies have proposed different machine learning methods for predicting the progression to diabetes from a healthy or prediabetic state with relatively good performance [54,55]. The strength of both studies lies in their inclusion of a large number of subjects. Both of them proposed a classification strategy relying on a threshold-based categorization upon different ranges of glucose or A1c. This allows them to obtain robust results but at the expense of making an inappropriate interpretation in physiological terms [33,38]. Furthermore, both studies are of observational nature and subjects with only partially available data were excluded from the analysis. This suggests that caution should be exercised when utilising these results for decision-making.

The present line of research assigns a key role to the analysis of glucose excursions from CGM data in search of better phenotyping and corresponding progress towards the implementation

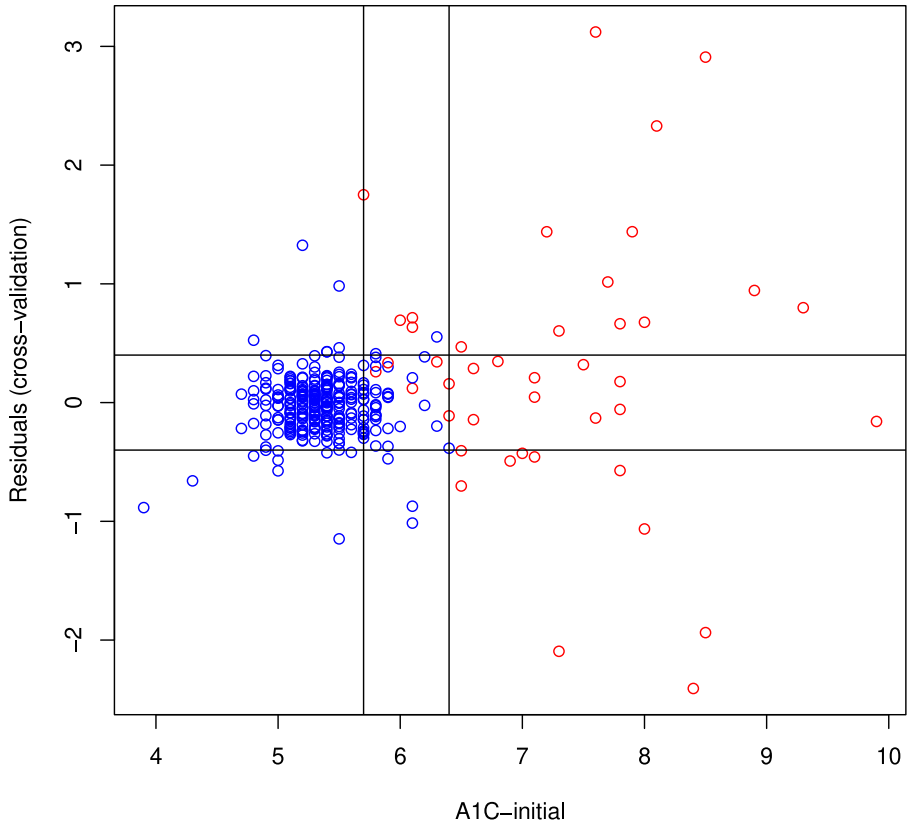

Fig. 4. Residuals vs. $A1c_{initial}$ for the model that includes glucodensity as a covariate in the AEGIS database. Red circles correspond to diabetic patients.

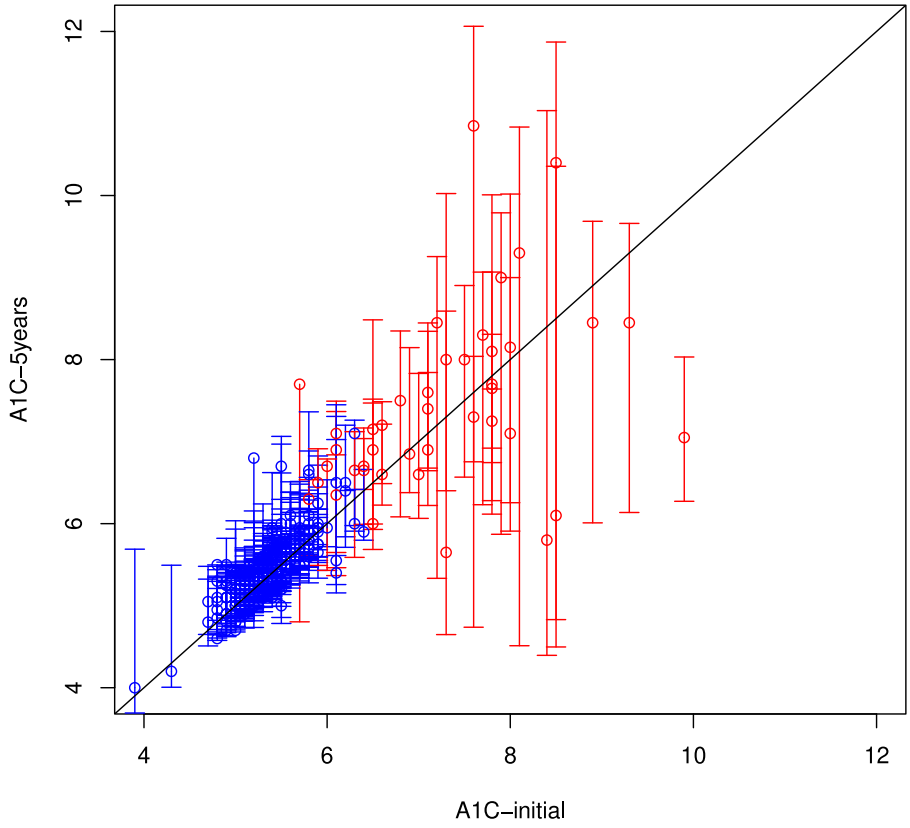

Fig. 5. Prediction intervals for each response observed in the AEGIS database (90% confidence level). The red circles correspond to patients with diabetes.

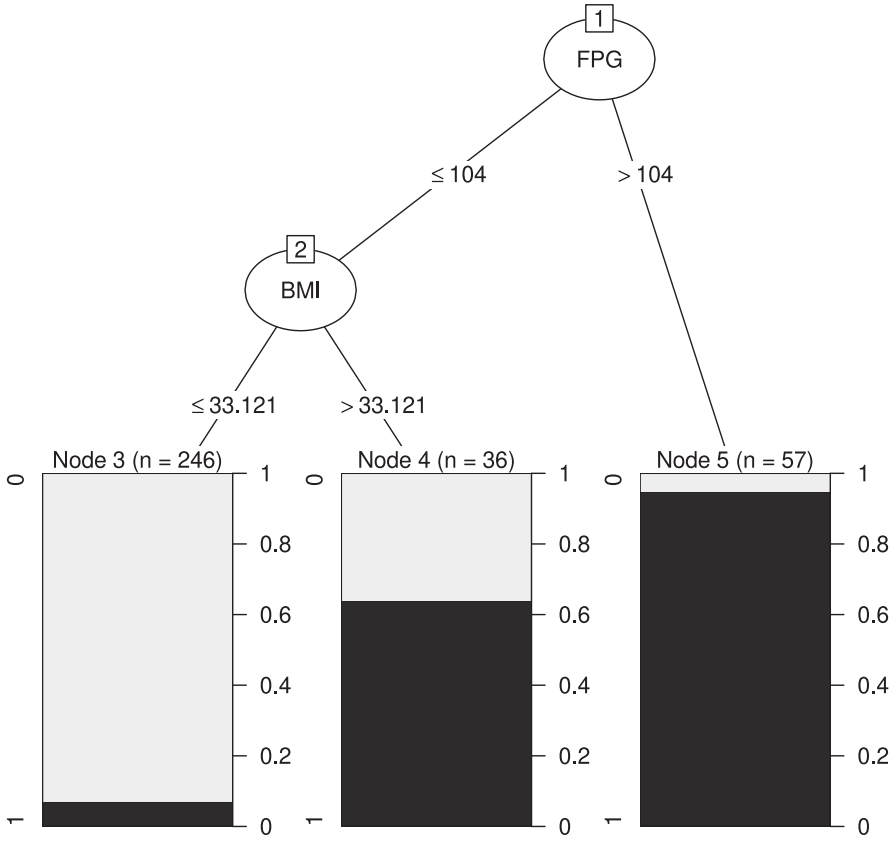

Fig. 6. Clinical decision rules that allow us to identify those patients with a significant uncertainty in their $A1c_{5year}$ predictions.

of a personalized intervention [43,56]. An interesting asset of the present proposal is the proper evaluation of the limits of predictive models by estimating the uncertainty of the predictions for each new subject. Thus, a careful analysis of the results that exhibit significant discrepancies with the model predictions provides the opportunity to identify certain patient phenotypes that need to be followed up more closely. These discrepancies can be explained by many different factors (lifestyle, diet, disease, pharmacological treatments, etc.) over time. The present study shows that these discrepancies can be promptly recognized using routine clinical practice biomarkers.

An inherent limitation of the AEGIS study was its modest sample size. In this respect, kernel methods have proven effective in coping with a distributional representation, but at the cost of a substantial amount of data to show a significant advantage in high uncertainty settings. A larger sample size would refine the predictive model and enable the inclusion of stratification effects in future studies [57,58]. Another limitation can be found in the 3-7-days CGM recording period of this study. An extension of this period to 14 days would probably limit possible intraday variations in glycemic profile representation; however, the discomfort from wearing a CGM device for such a long period is not a minor issue.

Ultimately, our findings enforce the prominent role of CGM data in providing a comprehensive picture of glucose metabolism [59] and allow us to envision new research on further characterizing glucose dynamics by devising new methods for (1) measuring the variability of glucose excursion, (2) clustering different glucose profiles, or (3) unveiling patterns of glucose excursions related to specific pathophysiological mechanisms. In particular, the inclusion of both CGM-based information and longitudinal multi-omics information in the analysis may provide deep insight into the underlying mechanisms involved in the onset and progression of the disease [60]. Lastly, further research is needed on new glycemic outcomes, beyond average measures like A1c, in order to capture a more accurate picture of glycemic dynamics, and glucodensity might be exploited as a new source of information for more robust predictions.

6. Conclusions

The present work proposes a data analysis framework that is 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 appropriate 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 explanatory variables and allowing capturing higher-order interactions. Furthermore, RKHS provides a natural integration of complex statistical objects into the same predictive task, offering a powerful tool for simultaneously coping with multiple sources of information.

We illustrated the usefulness of this approach for predicting long-term changes in a standard biomarker for glycemic control. Importantly, our analysis included glucose density, a novel representation of CGM data, as a predictor. The results show that glucodensity provides more predictive information than previous widely used diabetes biomarkers, enhancing the role of CGM data as a personal signature of glucose homeostasis. Furthermore, our approach estimates the uncertainty of predictions, enabling the characterization of the phenotypes of those subjects for whom this uncertainty is significant, thus guiding a personalized follow-up.

Implementation

To support reproducible research, the source code of the methods presented in this paper has been published under an open-source license[3].

Acknowledgment

This study was supported by ISCIII (PI20/01069, RD21/0016/0022; Cofinanciado por la Unión Europea/FEDER, "A way to make Europe"); and the Ministry of Science, Innovation and Universities of Spain (RTI2018-099646-B-I00).

Appendix. Guide to acronyms

Acronyms	Meaning
A1c	Glycated hemoglobin
BMI	Body mass index
CGM	Continuous glucose monitoring
CONGA	Glycemic variability in terms of continuous overall net glycemic action
FPG	Fasting plasma glucose
HOMA-IR	Homeostasis model assessment-insulin resistance
HSIC	Hilbert-Schmidt independence criterion
IPW	Inverse probability weighting
MAGE	Mean amplitude of glycemic excursions
MAR	Missing at random
MMD	Maximum mean discrepancy
MODD	Mean of daily difference
RKHS	Reproducing kernel Hilbert space
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TAR	Time above range
TBR	Time below range
TIR	Time in range

References

[1] N.J. Perkins, S.R. Cole, O. Harel, E.J. Tchetgen Tchetgen, B. Sun, E.M. Mitchell, E.F. Schisterman, Principled approaches to missing data in epidemiologic studies, Am. J. Epidemiol. 187 (3) (2017) 568–575, doi:10.1093/aje/kwx348.
[2] R.A. Hughes, J. Heron, J.A.C. Sterne, K. Tilling, Accounting for missing data in statistical analyses: multiple imputation is not always the answer, Int. J. Epidemiol. 48 (4) (2019) 1294–1304, doi:10.1093/ije/dyz032.
[3] R.J. Little, R. D'Agostino, M.L. Cohen, K. Dickersin, S.S. Emerson, J.T. Farrar, C. Frangakis, J.W. Hogan, G. Molenberghs, S.A. Murphy, J.D. Neaton, A. Rotnitzky, D. Scharfstein, W.J. Shih, J.P. Siegel, H. Stern, The prevention and treatment of missing data in clinical trials, N Top N. Engl. J. Med. 367 (14) (2012) 1355–1360, doi:10.1056/NEJMsr1203730. PMID: 23034025.
[4] A. Tsiatis, Semiparametric theory and missing data, Springer Science & Business Media, 2007.
[5] B. Schölkopf, A.J. Smola, Learning with kernels: Support vector machines, regularization, optimization, and beyond, MIT press, 2001.
[6] T. Hofmann, B. Schölkopf, A.J. Smola, Kernel methods in machine learning, Ann. Stat. 36 (3) (2008) 1171–1220.
[7] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, Kernel mean embedding of distributions: a review and beyond, Found. Trends Mach. Learn. 10 (1–2) (2017) 1–141.
[8] M. Febrero-Bande, P. Galeano, W. González-Manteiga, Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random, Comput. Stat. Data Anal. 131 (2019) 91–103, doi:10.1016/j.csda.2018.07.006.
[9] R.J.A. Little, D.B. Rubin, Statistical analysis with missing data, volume 793, John Wiley & Sons, 2019.
[10] M. Matabuena, A. Petersen, J.C. Vidal, F. Gude, Glucodensities: a new representation of glucose profiles using distributional data analysis, Stat. Methods Med. Res. 30 (6) (2021) 1445–1464.
[11] E. Selvin, C.M. Crainiceanu, F.L. Brancati, J. Coresh, Short-term variability in measures of glycemia and implications for the classification of diabetes, Arch. Intern. Med. 167 (14) (2007) 1545–1551.
[12] B. Li, Linear operator-based statistical analysis: a useful paradigm for big data, Can. J. Stat. 46 (1) (2018) 79–103.
[13] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, A. Smola, A kernel statistical test of independence, Adv. Neural Inf. Process Syst. 20 (2007) 585–592.
[14] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (1) (2012) 723–773.
[15] B. Efron, R.J. Tibshirani, An introduction to the bootstrap, CRC press, 1994.
[16] S.A. Van de Geer, Applications of empirical process theory, volume 91, Cambridge University Press Cambridge, 2000.
[17] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[18] J. Chen, M. Stern, M.J. Wainwright, M.I. Jordan, Kernel feature selection via conditional covariance minimization, Adv. Neural Inf. Process. Syst. (NIPS 2017) 30 (2017) 6946–6955.
[19] L. Yang, S. Lv, J. Wang, Model-free variable selection in reproducing kernel hilbert space, J. Mach. Learn. Res. 17 (1) (2016) 2885–2908.
[20] B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, in: International conference on computational learning theory, Springer, 2001, pp. 416–426.
[21] K. Fukumizu, C. Leng, Gradient-based kernel method for feature extraction and variable selection, in: Advances in Neural Information Processing Systems, in: NIPS'12, 2012, pp. 2114–2122.
[22] T. Liu, Y. Goldberg, et al., Kernel machines with missing responses, Electron. J. Stat. 14 (2) (2020) 3766–3820.
[23] T. Liang, A. Rakhlin, et al., Just interpolate: kernel ridgeless regression can generalize, Ann. Stat. 48 (3) (2020) 1329–1347.
[24] L. Lei, E.J. Candès, Conformal inference of counterfactuals and individual treatment effects, J. R. Stat. Soc. Ser. B 83 (5) (2021) 911–938.
[25] J.D.Y. Kang, J.L. Schafer, et al., Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data, Statist. Sci. 22 (4) (2007) 523–539.
[26] K. Vermeulen, S. Vansteelandt, Bias-reduced doubly robust estimation, J. Am. Stat. Assoc. 110 (511) (2015) 1024–1036.
[27] J. Lei, M. G'Sell, A. Rinaldo, R.J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, J. Am. Stat. Assoc. 113 (523) (2018) 1094–1111.
[28] Y. Zheng, S.H. Ley, F.B. Hu, Global aetiology and epidemiology of type 2 diabetes mellitus and its complications, Nat. Rev. Endocrinol. 14 (2) (2018) 88–98.
[29] 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 Res. Clin. Pract. 157 (2019) 107843.
[30] F.B. Hu, A. Satija, J.E. Manson, Curbing the diabetes pandemic: the need for global policy solutions, JAMA 313 (23) (2015) 2319–2320.
[31] K. Makrilakis, S. Liatis, S. Grammatikou, D. Perrea, C. Stathi, P. Tsiligros, 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 Metab. 37 (2) (2011) 144–151.
[32] K. Mühlenbruch, R. Paprott, H.-G. Joost, H. Boeing, C. Heidemann, 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 Res. Care 6 (1) (2018) e000524.
[33] E.A.M. Gale, Is type 2 diabetes a category error? Lancet 381 (2013) 1956–1957.
[34] E.J. Topol, Transforming medicine via digital innovation, Sci. Transl. Med. 2 (16) (2010) 16cm4.
[35] N.J. Schork, Personalized medicine: time for one-person trials, Nature 520 (7549) (2015) 609–611.
[36] M.R. Kosorok, E.B. Laber, Precision medicine, Annu. Rev. Stat. Appl. 6 (2019) 263–286.
[37] D. Cirillo, A. Valencia, Big data analytics for personalized medicine, Curr. Opin. Biotechnol. 58 (2019) 161–167.
[38] F. Zaccardi, K. Khunti, Glucose dysregulation phenotypes - time to improve outcomes, Nat. Rev. Endocrinol. 14 (11) (2018) 632–633.
[39] A.L. Peters, A.J. Ahmann, T. Battelino, A. Evert, I.B. Hirsch, M.H. Murad, W.E. Winter, 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. 101 (11) (2016) 3922–3937.
[40] A.D. Association, 7. Diabetes technology: standards of medical care in diabetes-2019, Diabetes Care 42 (2019) S71–S80.
[41] W.H. Organization, Global report on diabetes, World Health Organization, 2016.
[42] L. Johnston, G. Wang, K. Hu, C. Qian, G. Liu, Advances in biosensors for continuous glucose monitoring towards wearables, Front. Bioeng. Biotechnol. 9 (2021).
[43] H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin, M. Snyder, Glucotypes reveal new patterns of glucose dysregulation, PLoS Biol. 16 (7) (2018) e2005143.
[44] 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, A. González-Quintela, Glycemic variability and its association with demographics and lifestyles in a general adult population, J. Diabetes Sci. Technol. 11 (4) (2017) 780–790.
[45] 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 42 (8) (2019) 1593–1603.
[46] R.W. Beck, R.M. Bergenstal, T.D. Riddlesworth, C. Kollman, Z. Li, A.S. Brown, K.L. Close, Validation of time in range as an outcome measure for diabetes clinical trials, Diabetes Care 42 (3) (2019) 400–405.

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

[47] A.M. Gómez, D.C. Henao, A. Imitola Madero, L.B. Taboada, V. Cruz, M.A. Robledo Gomez, M. Rondon, O. Munoz-Velandia, M. Garcia-Jaramillo, F.M. Leon Vargas, Defining high glycemic variability in type 1 diabetes: comparison of multiple indexes to identify patients at risk of hypoglycemia, Diabetes Technol. Therapeut. 21 (8) (2019) 430–439.
[48] D. Rodbard, Glucose variability: a review of clinical applications and research developments, Diabetes Technol. Therapeut. 20 (S2) (2018) S2–5.
[49] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.-P. Kriegel, B. Schölkopf, A.J. Smola, Integrating structured biological data by kernel maximum mean discrepancy, Bioinformatics 22 (14) (2006) e49–e57.
[50] C. Berg, J.P.R. Christensen, P. Ressel, Harmonic analysis on semigroups: Theory of positive definite and related functions, volume 100, Springer, 1984.
[51] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu, Equivalence of distance-based and RKHS-based statistics in hypothesis testing, Ann. Stat. (2013) 2263–2291.
[52] I. Gaynanova, N. Punjabi, C. Crainiceanu, Modeling continuous glucose monitoring (CGM) data during sleep, Biostatistics (2020).
[53] A. Zaitcev, M.R. Eissa, Z. Hui, T. Good, J. Elliott, M. Benaissa, A deep neural network application for improved prediction of HbA1c in Type 1 diabetes, IEEE J. Biomed. Health Inform. 24 (10) (2020) 2932–2941.
[54] Y. Wu, H. Hu, J. Cai, R. Chen, X. Zuo, H. Cheng, D. Yan, Machine learning for predicting the 3-year risk of incident diabetes in chinese adults, Front. Public Health 9 (2021).
[55] A. Cahn, A. Shoshan, T. Sagiv, R. Yesharim, R. Goshen, S. Varda, I. Raz, Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model, Diabetes Metab. Res. Rev. 36 (2) (2020) e3252.
[56] A.A. Tsiatis, Dynamic treatment regimes: Statistical methods for precision medicine, CRC Press, 2019.
[57] E. Ahlqvist, P. Storm, A. Käräjämäki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. Prasad B, D. Mansour Aly, P. Almgren, Y. Wessman, N. Shaat, P. Spégel, H. Mulder, E. Lindholm, O. Melander, O. Hansson, U. Malmqvist, A. Lernmark, L. Groop, Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables, Lancet Diabetes Endocrinol. 6 (2018), doi:10.1016/S2213-8587(18)30051-2.
[58] E. Ahlqvist, T. Tuomi, L. Groop, Clusters provide a better holistic view of type 2 diabetes than simple clinical features, Lancet Diabetes Endocrinol. 7 (9) (2019) 668–669.
[59] B.A.W. Group, Need for regulatory change to incorporate beyond A1c glycemic metrics, Diabetes Care 41 (6) (2018) e92–e94.
[60] W. Zhou, M.R. Sailani, K. Contrepois, Y. Zhou, S. Ahadi, S.R. Leopold, M.J. Zhang, V. Rao, M. Avina, T. Mishra, et al., Longitudinal multi-omics of host–microbe dynamics in prediabetes, Nature 569 (7758) (2019) 663–671.