

Kernel methods to handle missing responses and their application in modeling five-year glucose changes using distributional representations

Marcos Matabuena^{a,*,1}, Paulo Félix^a, Carlos García-Meixide^c and Francisco Gude^b

^a CiTIUS (Centro Singular de Investigación en Tecnoloxías Intelixentes), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, SPAIN

^b Unidade de Epidemioloxía Clínica, Complexo Hospitalario Universidade de Santiago (CHUS), Travesía da Choupana, 15706 Santiago de Compostela, SPAIN.

^c Universidade de Santiago de Compostela, 15782 Santiago de Compostela, SPAIN

ARTICLE INFO

Keywords:

Diabetes mellitus
Continuous glucose monitoring technology
Kernel methods
Missing data
Statistical independence
Variable selection
Regression modeling

Abstract

Background and objectives: Missing data is a ubiquitous problem in longitudinal studies due to the number of patients lost to follow-up. Motivated by the remarkable role of this issue in the AEGIS diabetes study, this paper aims to propose a new set of kernel methods to handle missing data in the response variable. These methods have been applied to predict long-term changes in the glycated hemoglobin (A1c), the primary biomarker used to diagnose and monitor the progression of diabetes mellitus disease.


Methods: We propose a framework of kernel methods for testing statistical independence, selecting relevant explanatory variables and quantifying the uncertainty of the resultant predictive models. As a novelty in the clinical analysis, we use a distributional representation for continuous glucose monitoring (CGM) and compare its performance against Time in Range metrics, the state-of-the-art vectorial representation.

Results: The results show that, after the incorporation of CGM information, predictive ability increases from $R^2 = 0.61$ to $R^2 = 0.64$ (Time in Range metrics), $R^2 = 0.65$ (mean amplitude of glycemic excursions) and $R^2 = 0.71$ (distributional representation). In addition, the uncertainty analysis is proven useful to characterize some subpopulations where predictivity get worse, and a more personalized 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 and have the potential to improve the results of predictive tasks by including new complex objects as explanatory variables. Moreover, by applying these methods to a longitudinal study of diabetes mellitus, we show that a distributional representation of CGM data provides greater sensitivity in predicting five-year A1c changes than classical diabetes biomarkers and traditional CGM metrics.

Highlights

- Missing data are a common threat to the validity of longitudinal studies, demanding specific approaches to correct the potential bias in data sample.
- The present paper extends several non-parametric methods to include a more complex representation of explanatory variables in the context of missing responses. As a result, new methods for testing statistical independence, selecting explanatory variables, and quantifying the uncertainty of predictive models are proposed.
- The new methods have been applied to model long-term glucose changes in a sample from general population, where high-resolution data from CGM is available.
- In order to handle CGM data, a novel distributional representation is proposed, which can be regarded as a functional extension of the standard Time in Range metrics. The results from applying the new methods show that this distributional representation provide greater sensitivity in explaining long-term glucose changes than Time in Range metrics and other classical diabetes biomarkers.

 marcos.matabuena@usc.es (M. Matabuena)
ORCID(s):

- An essential instrument for decision-making arises from the proper evaluation of the limits of the predictive models by estimating the uncertainty of the predictions. As a result, some subpopulations to whom the model provide an unreliable prediction can be phenotypically characterized, enabling new strategies for integrating personalized medicine into healthcare practice.

1. Introduction

Diabetes Mellitus is one of the most critical public health problems, being the ninth major cause of mortality worldwide Zheng, Ley and Hu (2018). At present, over 416 and 47 million patients have Type 2 and Type 1 diabetes respectively Saeedi, Petersohn, Salpea, Malanda, Karuranga, Unwin, Colagiuri, Guariguata, Motala, Ogurtsova et al. (2019). Importantly, around 50% of patients with diabetes are undiagnosed Saeedi et al. (2019). 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 Hu, Satija and Manson (2015).

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 Topol (2010); Schork (2015); Kosorok and Laber (2019); Cirillo and Valencia (2019). In the particular case of diabetes, the latest advances in sensing technology allow 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) Zaccardi and Khunti (2018). 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 Peters, Ahmann, Battelino, Evert, Hirsch, Murad, Winter and Wolpert (2016); American Diabetes Association (2019). Still, few works explore the use of CGM technology with healthy populations in order to draw new conclusions on long-term glucose changes. Worthy of mentioning is the work by Hall, Perelman, Breschi, Limcaoco, Kellogg, McLaughlin and Snyder (2018), which shows some advantages of using CGM data to develop new screening tests in prediabetes population.

This paper aims at predicting long-term changes in glucose homeostasis, by using information provided by a CGM device and some common diabetes biomarkers. We include a novel distributional representation for CGM data and compare its performance against Time in Range (TIR) metrics, considered as the key CGM-derived metric for assessing short-time glycemic control. Among different biomarkers, we select the glycated hemoglobin (A1c) as response variable. A1c is a measure of average blood glucose level over the past three months, and it is the preferred option because it provides more reproducible values in laboratory and is subject to less measurement error Selvin, Crainiceanu, Brancati and Coresh (2007). Furthermore, we aim to assess and discuss the residuals and the predictive capacity of several variables associated with the evolution of A1c in the long term, providing interpretable clinical phenotypes for large uncertainty cases.

The effectiveness of the present approach is tested on the AEGIS study, 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 demanded in order to maintain the validity of the statistical inferences Little and Rubin (2019).

In order to deal with the aforementioned prediction problem, we adapt previous general-purpose methods for statistical independence testing (Section 3.2), variable selection (Section 3.3), and conformal inference (Section 3.4) to the missing data setting. We build the present proposal under Reproducing Kernel Hilbert Space (RKHS) learning paradigm due to their ability to model general and complex dependence relations between study variables Schölkopf, Smola, Bach et al. (2002); Hofmann, Schölkopf and Smola (2008). Furthermore, the RKHS paradigm is particularly suitable for dealing with heterogeneous complex data such as graphs or curves that take values on a continuum Muandet, Fukumizu, Sriperumbudur and Schölkopf (2017), as is the case with the functional distributional representation of glucose profiles that we introduce in Section 2.2.

The rest of this paper is outlined as follows: Section 2 describes the AEGIS database used for testing our proposal. Subsequently, Section 3 describes in detail the methods for statistical independence testing, variable selection, and inference on the uncertainty of new predictions. Section 4 shows the results from applying these methods to the AEGIS

	Men (<i>n</i> = 220)	Women (<i>n</i> = 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.μUI/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 1

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.

database. Section 5 discusses the advantages and drawbacks of this approach. Finally, some conclusions are provided in Section 6.

1.1. Data analysis outline

Ultimately, this paper depicts a data analysis framework designed as a pipeline of model-free methods for predictive problems with missing data. Their application to diabetes mellitus allows us to examine the relationship between the baseline characteristics of participants in a five-year study and A1c as response variable. Among the set of explanatory variables, a new distributional representation for CGM data poses a major challenge for modeling purposes. What is more, the question arises as to whether this new information is relevant to the prediction task. Basically, the proposed framework comprises the following steps:

1. To measure the statistical association between each explanatory variable and the response variable with a statistical independence test. To this end, we adapt a previous kernel independence test to check association between some diabetes biomarkers and five-year changes in A1c variable, $A1c_{5years} - A1c_{initial}$, and we design a new bootstrap method to perform test callibration. Moreover, in order to improve the clinical interpretation of the results we evaluate the association strength by graphical representation.
2. To identify the best subset of explanatory variables revealing higher-order interactions with the response variable in order to improve the prediction. To this end, we adapt a previous kernel variable selection method and apply it for finding 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 adapt a previous kernel ridge regression method and we apply it to predict $A1c_{5years}$.
4. To estimate the uncertainty of the predictions. To this end, we design a new method to provide a prediction interval for the response variable based on conformal inference. By using this method, we can measure the limits of the regression models previously obtained and, significantly, we can identify specific patient subpopulations that do not fit the expected behavior, a key issue for clinical decision-making.

2. Diabetes clinical data

2.1. The AEGIS diabetes study

The AEGIS diabetes population study, conducted in the Spanish town of A Estrada (Galicia), aims 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 Gude, Díaz-Vidal, Rúa-Pérez, Alonso-Sampedro, Fernández-Merino, Rey-García, Cadarso-Suárez, Pazos-Couselo, García-López and Gonzalez-Quintela (2017), 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 1 shows the baseline 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 339 patients. Complete details about the study design and the measurement methodology protocol can be found in Gude et al. (2017).

2.2. Glucodensity

We adopt a novel functional representation for CGM data, termed glucodensity, that allows us to obtain a personalized functional profile of patient glucose homeostasis Matabuena, Petersen, Vidal and Gude (2020). Glucodensity is a natural extension of TIR metrics, the current gold standard for representing CGM data. TIR measures the proportion of time that a person spends with their blood glucose levels in a target range Battelino, Danne, Bergenstal, Amiel, Beck, Biester, Bosi, Buckingham, Cefalu, Close et al. (2019); Beck, Bergenstal, Riddlesworth, Kollman, Li, Brown and Close (2019), for example, in the hypo-hyper glycemic range. TIR is an intuitive metric but has two main disadvantages; first, the range will vary depending on the characteristics of the population examined; and second, 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.

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,

$$\hat{f}(x) = \frac{1}{m} \sum_{j=1}^m \frac{1}{h} k\left(\frac{x - x_j}{h}\right), \quad (1)$$

where $h > 0$ is the smoothing parameter and $k(\cdot)$ denotes a non-negative real-valued integrable function (Figure 1).

A critical point in kernel analysis is to measure the difference between two density functions. In this paper, 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

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

where \hat{Q}_{f_1} , and \hat{Q}_{f_2} are the corresponding quantile functions. Significantly, since 2-Wasserstein distance between two densities depends only on their quantile functions, it is not necessary to resort to density estimation methods, and we can approximate this distance using quantile-function estimations through empirical distributions.

Intuitively, glucodensity is more sensitive than previous CGM summary metrics. We then explore its use in modeling long-term glucose changes, and compare it with TIR metrics Hirsch, Sherr and Hood (2019); Battelino et al. (2019). For such purpose, we use some common glucose ranges that appear in the literature: i) $70 - 140 \text{ mg/dL}$; ii) $70 - 180 \text{ mg/dL}$; and iii) $> 180 \text{ mg/dL}$, that we denote as TIR^{70-140} , TIR^{70-180} , and $\text{TIR}^{>180}$, respectively. TIR^{70-140} has been used for non diabetic individuals and as secondary range for regulatory issues and comparability studies, TIR^{70-180} corresponds to the recently released international consensus in TIR metrics for diabetic patients, and $\text{TIR}^{>180}$ is a consensus metrics for assessing hyperglycemia.

Besides, we also consider different summary metrics derived from CGM data Gómez, Henao, Imitola Madero, Taboada, Cruz, Robledo Gomez, Rondon, Munoz-Velandia, Garcia-Jaramillo and Leon Vargas (2019); Rodbard (2018): CONGA (continuous overall net glycemic action), MAGE (mean amplitude of glycemic excursions), and MODD (mean of daily differences).

3. Methods

3.1. Preliminaries

We shall 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 $\mathbf{X} = (X^1, \dots, X^p)$ denotes the explanatory variables, Y the response variable, and R a binary random variable that indicates whether the response is missing or not. \mathcal{X} denotes a general 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 independent and identically distributed observations, where y_i is missing if $r_i = 0$. We assume R to be distributed according to $R \sim \text{Ber}(\pi(X))$ with $\pi(\cdot) = P(R = 1 | \mathbf{X} = \cdot)$, and hence some of the explanatory variables 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 that the probability

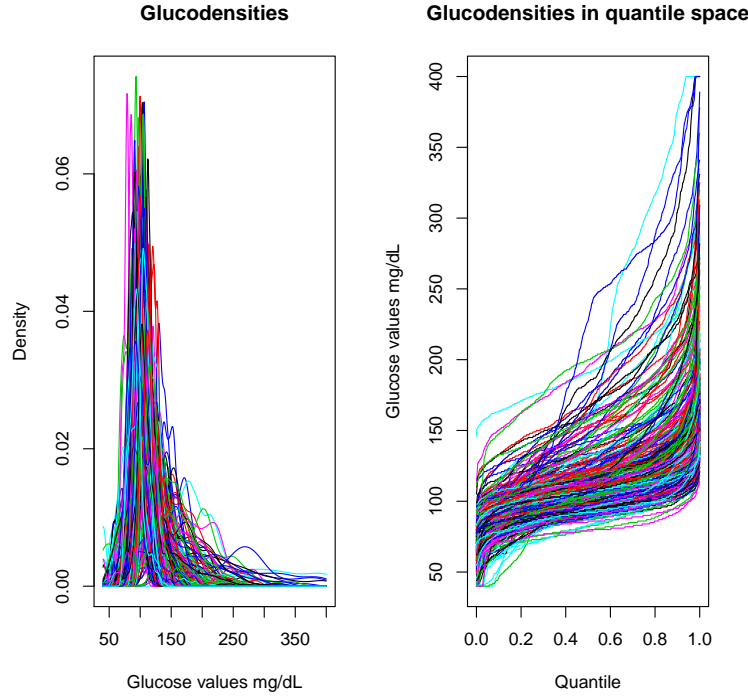


Figure 1: Glucodensities estimated from a random sample of the AEGIS study on diabetic and normoglycemic patients are shown. Left: glucose representation estimates the proportion of time spent by a patient at each glucose concentration over a continuum. Right: the representation of the glucodensities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

of not observing a patient increases with age. We also assume missing at random (MAR) mechanism in the response Y , namely R and Y are conditionally independent given \mathbf{X} , $R \perp\!\!\!\perp Y | \mathbf{X}$.

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

$$Y = f(\mathbf{X}) + \epsilon, \quad (3)$$

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 those datasets where some values for Y are not observed. To this aim we provide: 1) a method for univariate analysis based on testing statistical independence between each explanatory variable and the response variable; 2) a method for selecting the subset of explanatory variables that best predict the response variable; and 3) methods for predicting the response variable and for inferring the uncertainty in the predictions.

These methods are based on Reproducing Kernel Hilbert Space (RKHS) learning paradigm. The core element of this paradigm is a positive definite kernel function $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which allows us to measure the similarity between any data points $x, y \in \mathcal{X}$. The positive definiteness of the kernel function guarantees the existence of a dot product space \mathcal{H} and a feature mapping $\phi : \mathcal{X} \rightarrow \mathcal{H}$ 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 a linear problem Li (2018), allowing computational algorithms to easily determine optimal solutions.

3.2. Testing statistical independence

We wish to test whether two random variables $X \sim P_X$ and $Y \sim P_Y$ are not independent, i.e., 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 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 Muandet et al. (2017), 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 Gretton, Fukumizu, Teo, Song, Schölkopf and Smola (2007). We shall extend this test to the 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 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 $\phi(P) = \int_{\mathcal{X}} k(\cdot, x) dP(x)$, 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 Gretton, Borgwardt, Rasch, Schölkopf and Smola (2012):

$$\text{MMD}_{k_{\mathcal{X}}}(P, Q) = \|\phi(P) - \phi(Q)\|_{\mathcal{H}_{k_{\mathcal{X}}}}. \quad (4)$$

For the class of *characteristic* kernels the embeddings are injective, i.e., $\text{MMD}_k(P, Q) = 0$ 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 $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}$ (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 Gretton et al. (2012), defined as

$$\text{HSIC}(P_{X,Y}, P_X P_Y) = \|\phi_{X,Y} - \phi_X \otimes \phi_Y\|_{\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}}^2 \quad (5)$$

and it can be shown that when $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are characteristic kernels then $\text{HSIC}(P_{X,Y}, P_X P_Y) = 0$ if and only if $X \perp\!\!\!\perp Y$. Expanding Equation 5 we have

$$\begin{aligned} \text{HSIC}(P_{X,Y}, P_X P_Y) &= \langle \phi_{X,Y} - \phi_X \otimes \phi_Y, \phi_{X,Y} - \phi_X \otimes \phi_Y \rangle_{\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}} \\ &= \langle \phi_{X,Y}, \phi_{X,Y} \rangle + \langle \phi_X \otimes \phi_Y, \phi_X \otimes \phi_Y \rangle - 2\langle \phi_{X,Y}, \phi_X \otimes \phi_Y \rangle, \end{aligned} \quad (6)$$

where we drop $\mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_{\mathcal{Y}}}$ 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

$$\begin{aligned} \text{HSIC}(P_{X,Y}, P_X P_Y) &= \mathbb{E}_{X,Y,X',Y'}[k_{\mathcal{X}}(X, X')k_{\mathcal{Y}}(Y, Y')] + \mathbb{E}_{X,X'}[k_{\mathcal{X}}(X, X')]\mathbb{E}_{Y,Y'}[k_{\mathcal{Y}}(Y, Y')] - \\ &\quad - 2\mathbb{E}_{X,Y}[\mathbb{E}_{X'}[k_{\mathcal{X}}(X, X')]\mathbb{E}_{Y'}[k_{\mathcal{Y}}(Y, Y')]], \end{aligned} \quad (7)$$

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

$$\begin{aligned} \widehat{\text{HSIC}}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_{\mathcal{X}}(x_i, x_j)k_{\mathcal{Y}}(y_i, y_j) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_{\mathcal{X}}(x_i, x_j) \sum_{i=1}^n \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_{\mathcal{Y}}(y_i, y_j) - \\ &\quad - \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n k_{\mathcal{X}}(x_i, x_j)k_{\mathcal{Y}}(y_i, y_k). \end{aligned} \quad (8)$$

Under MAR 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 methods. 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 Tsiatis (2007), given by

$$w_i = \frac{r_i}{n\pi(x_i)}, \quad i = 1, \dots, n, \quad (9)$$

which results in assigning large w_i values as the probability of observing a response decreases. With 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

$$w_i^* = \frac{w_i}{\sum_{i=1}^n w_i}, \quad i = 1, \dots, n. \quad (10)$$

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 $1/n$ of the empirical distribution with the normalized weights $\hat{W}^* = (\hat{w}_1^*, \dots, \hat{w}_n^*)$ in the Equation 8. Thus, we obtain

$$\begin{aligned} \widehat{\text{HSIC}}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y) &= \sum_{i=1}^n \sum_{j=1}^n \hat{w}_i^* \hat{w}_j^* k_X(x_i, x_j) k_Y(y_i, y_j) + \sum_{i=1}^n \sum_{j=1}^n \hat{w}_i^* \hat{w}_j^* k_X(x_i, x_j) \sum_{i=1}^n \sum_{j=1}^n \hat{w}_i^* \hat{w}_j^* k_Y(y_i, y_j) - \\ &\quad - \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \hat{w}_i^* \hat{w}_j^* \hat{w}_k^* k_X(x_i, x_j) k_Y(y_i, y_k). \end{aligned} \quad (11)$$

Calibration under the null hypothesis with the precedent statistic is not trivial and the permutation approach is generally not valid, since the response Y is not exchangeable due to non-homogeneous missing data mechanism. To overcome this difficulty, we propose a novel bootstrap approach, which properly deals with distributional representations and, in general, with other complex statistical objects Efron and Tibshirani (1994).

Under the 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,

$$\begin{aligned} \widehat{\text{HSIC}}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y) &= \langle \hat{\phi}_{X,Y} - \hat{\phi}_X \otimes \hat{\phi}_Y, \hat{\phi}_{X,Y} - \hat{\phi}_X \otimes \hat{\phi}_Y \rangle_{H_X \otimes H_Y} \\ &= \langle \hat{\phi}_{X,Y} - \phi_{X,Y} + \phi_X \otimes \phi_Y - \hat{\phi}_X \otimes \hat{\phi}_Y, \hat{\phi}_{X,Y} - \phi_{X,Y} + \phi_X \otimes \phi_Y - \hat{\phi}_X \otimes \hat{\phi}_Y \rangle. \end{aligned} \quad (12)$$

Then, a natural bootstrap procedure that allows us 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 $\widehat{\text{HSIC}}^{j*}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y)$ as

$$\begin{aligned} \widehat{\text{HSIC}}^{j*}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y) &= \langle \hat{\phi}_{X,Y} - \hat{\phi}_{X,Y}^{j*} + \hat{\phi}_X^{j*} \otimes \hat{\phi}_Y^{j*} - \hat{\phi}_X \otimes \hat{\phi}_Y, \\ &\quad \hat{\phi}_{X,Y} - \hat{\phi}_{X,Y}^{j*} + \hat{\phi}_X^{j*} \otimes \hat{\phi}_Y^{j*} - \hat{\phi}_X \otimes \hat{\phi}_Y \rangle_{H_X \otimes H_Y}, \end{aligned} \quad (13)$$

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

$$p\text{-value} = \frac{1}{m} \sum_{j=1}^m I\left(\widehat{\text{HSIC}}^{j*}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y) \geq \widehat{\text{HSIC}}(\hat{P}_{X,Y}, \hat{P}_X \hat{P}_Y)\right). \quad (14)$$

The bootstrap consistency with missing data can be proved by using some standard tools of empirical process theory Van de Geer (2000), and it is provided as supplementary material.

3.3. 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 Guyon and Elisseeff (2003).

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 Chen, Stern, Wainwright and Jordan (2017); and second, identifying those variables with non-zero gradient function Yang, Lv and Wang (2016). 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 Yang et al. (2016), we propose to identify the relevant predictors by learning the gradient of the regression function f . 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 $\mathbf{g}(\mathbf{X}) = \nabla f(\mathbf{X}) = (g_1(\mathbf{X}), \dots, g_p(\mathbf{X}))^T$ 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 close enough to \mathbf{x}_i then $f(\mathbf{x}_j) \approx y_i + \mathbf{g}(\mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)$. 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 \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 neighborhood 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 should be adjusted to warrant asymptotic estimation consistency.

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 counterpart

$$\hat{\mathcal{E}}(\mathbf{g}) = \frac{1}{n^2} \sum_{i,j=1}^n \omega_{ij} (y_j - y_i - \mathbf{g}(\mathbf{x}_i)^T (\mathbf{x}_j - \mathbf{x}_i))^2, \quad (15)$$

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 (15) as a functional in the RKHS \mathcal{H}_k^p , so $\mathbf{g} \in \mathcal{H}_k^p$ and $\mathcal{E} : \mathcal{H}_k \times \overset{p}{\cdot} \times \mathcal{H}_k \rightarrow \mathbb{R}^+$, induced by a pre-specified positive kernel k . Thus, we propose the following optimization formula to learn the gradient vector:

$$\arg \min_{\mathbf{g} \in \mathcal{H}_k^p} \frac{1}{n^2} \sum_{i,j=1}^n \omega_{ij} (y_i - y_j - \mathbf{g}(\mathbf{x}_i)^T (\mathbf{x}_i - \mathbf{x}_j))^2 + J(\mathbf{g}). \quad (16)$$

Under MAR assumption 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 (10). The variable selection expression can be rewritten as

$$\arg \min_{\mathbf{g} \in \mathcal{H}_k^p} \frac{1}{n^2} \sum_{i,j=1}^n \hat{\omega}_{ij}^* (y_i - y_j - \mathbf{g}(\mathbf{x}_i)^T (\mathbf{x}_i - \mathbf{x}_j))^2 + J(\mathbf{g}). \quad (17)$$

The representer theorem states that the minimizer of (17) can be represented as a finite linear combination of kernel products evaluated on the samples of the dataset Schölkopf, Herbrich and Smola (2001):

$$g_r(\cdot) = \sum_{i=1}^n \alpha_i^r k_{\mathcal{X}}(\cdot, \mathbf{x}_i), \quad r = 1, \dots, p, \quad (18)$$

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

Several regularization terms have been considered in the bibliography. We adopt the Group Lasso penalty Fukumizu and Leng (2012); Yang et al. (2016):

$$J(g_r) = \inf \left\{ \|\alpha^r\|_2 : g_r(\cdot) = \sum_{i=1}^n \alpha_i^r k_{\mathcal{X}}(\cdot, \mathbf{x}_i) \right\}, \quad (19)$$

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

$$\arg \min_{\alpha^1, \dots, \alpha^p} \sum_{i,j=1}^n \hat{\omega}_{ij}^* (y_i - f^*(\mathbf{x}_i, \mathbf{x}_j))^2 + \lambda_n \sum_{r=1}^p \eta_r \|\alpha^r\|_2, \quad (20)$$

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 $\mathbf{k}_i = (k(\mathbf{x}_i, \mathbf{x}_1), \dots, k(\mathbf{x}_i, \mathbf{x}_n))^T$ 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 (16) 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 Yang et al. (2016).

3.4. Prediction and uncertainty analysis

Let us recall that the ultimate goal is to predict Y using the information provided by the explanatory variables \mathbf{X} . To develop this aim, we adopt the kernel ridge regression approach proposed by Liu and Goldberg Liu, Goldberg et al. (2020). 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 Liang, Rakhlin et al. (2020). 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 Lei and Candès (2020).

Let us assume a linear regression model:

$$y_i = f(\mathbf{x}_i) + \epsilon = \mathbf{x}_i^T \beta + \epsilon \quad i = 1, \dots, n, \quad (21)$$

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:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \|\beta\|_2^2, \quad (22)$$

which is solved by $\hat{\beta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$, where $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, $\mathbf{y} = (y_1, \dots, y_n)^T$, 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 (22), exploiting the linear structure of problem, 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 $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{n \times n}$.

In Liu et al. (2020), the authors propose two different estimators for missing data. In both cases, the solution has the same closed-form expression given by representer theorem. 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:

$$\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} (\mathbf{W} \mathbf{y} + (\mathbf{I} - \mathbf{W}) \mu(\mathbf{x})), \quad (23)$$

where $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$ denotes a diagonal matrix containing the weights (see Equation 9) and $\mu(\mathbf{x}) = (\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n))^T$ denotes the imputation function.

Doubly robust estimators achieve optimal asymptotic variance when their weights w_1, \dots, 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

Kang, Schafer et al. (2007); Vermeulen and Vansteelandt (2015), thereby failing to provide real advantages with respect to IPW estimator.

The impact of the smoothing parameter on model generalization is an essential issue for the ensuing performance, and is strongly connected with the minimum-norm interpolating problem in the context of RKHS. Therefore, we propose to select the smoothing parameter through *leave-one-out* cross-validation, by adapting the estimators to missing data Liang et al. (2020).

In order to supply a prediction interval for the response with a confidence level of $1 - \alpha$, we provide a novel algorithm to perform conformal inference Lei and Candès (2020); Lei, G'Sell, Rinaldo, Tibshirani and Wasserman (2018), 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 23.
2. Compute the residuals $\hat{e}_i = |y_i^{test} - \hat{f}(\mathbf{x}_i^{test})|/\hat{\sigma}(\mathbf{x}_i^{test})$, for every $i = 1, \dots, n_2$ with $r_i^{test} = 1$. Value $\hat{\sigma}(\mathbf{x}_i^{test})$ 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:

$$\hat{F}_{n_2+1}^e(x) = \frac{1}{\sum_{i=1}^{n_2+1} \hat{w}_i^{test}} \sum_{i=1}^{n_2} 1\{\hat{e}_i \leq x\} \hat{w}_i^{test} + \hat{w}_{n_2+1}^{test}, \quad (24)$$

where we incorporate also the weight of \mathbf{x}_{n+1} , $\hat{w}_{n_2+1}^{test}$ in the estimation.

4. Compute the $1 - \alpha$ quantile, $\hat{q}_{1-\alpha}$, from $\hat{F}_{n_2+1}^e$.
5. Finally, return $[\hat{f}(\mathbf{x}_{n+1}) - \hat{q}_{1-\alpha} \hat{\sigma}(\mathbf{x}_{n+1}), \hat{f}(\mathbf{x}_{n+1}) + \hat{q}_{1-\alpha} \hat{\sigma}(\mathbf{x}_{n+1})]$ as the required prediction interval.

3.5. Handling multiple sources with a kernel

RKHS offers a powerful and natural data analysis paradigm that is able to cope with data of different nature Borgwardt, Gretton, Rasch, Kriegel, Schölkopf and Smola (2006). 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, $\mathbf{x} = (x^{gluco}, x^{real}, x^{categ})$. 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 $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sigma})$. Here, we propose to use the Laplacian kernel with the standard Euclidean distance as 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 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 Berg, Christensen and Ressel (1984); Sejdinovic, Sriperumbudur, Gretton and Fukumizu (2013), we propose to use a simple and global Laplacian kernel that integrates the three sources:

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

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

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 hemoglobin A1c is a hemoglobin-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?*

Variable	<i>p</i> – 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 2

Estimated raw p-values of A1c total variation vs each biomarker using the me proposed in Section 3.2 with normoglycemic patients.

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 interpretation, it is well-known that these variables are highly correlated. *Can we identify a reduced subset of relevant explanatory variables to predict five-year A1c changes?*
3. CGM technology may provide a more suitable tool for assessing glucose homeostasis than traditional diabetes biomarkers. *How do CGM data impact on improving 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 those individuals whose future glucose behavior cannot be precisely predicted?*

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

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 $A1c_{5years} - A1c_{initial}$ and those predictor variables shown in Table 1.

For this purpose, we use the Hilbert-Schmidt independence criterion that we propose in the context of missing data (Section 3.2), together with a specific bootstrap approach designed for this task. The underlying missing data mechanism is estimated using a univariate logistic regression.

Results in Table 2 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.

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 explanatory variables and the response variable to improve predictions. However, a key point to increase 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 adjust the method proposed in Section 3.3 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 1 except sex. We additionally include those TIR metrics specified in Section 2.2. In order to avoid overfitting and to improve the reproducibility of results, we select model parameters by cross-validation. We estimate the underlying missing data mechanism via lasso logistic regression.

Finally, the explanatory variables selected by the algorithm are: Age, $A1c_{initial}$, FPG, BMI, and MAGE. Notably, CGM contribution is made through the specific MAGE index, leaving aside fine-grained TIR information.

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

To answer this question, we fit several kernel ridge regression models (Section 3.4) for predicting $A1c_{5years}$: 1) Excluding CGM data as an explanatory variable; 2) Including CGM data through the MAGE index; 3) Including CGM data through the above mentioned TIR metrics (TIR^{70-140} , TIR^{70-180} , and $TIR^{>180}$); and 4) Including CGM data through glucodensity representation. Both of them share Age, $A1c_{initial}$, FPG and BMI as covariates. Kernel selection and parameter tuning have been calibrated following Section 3.5.

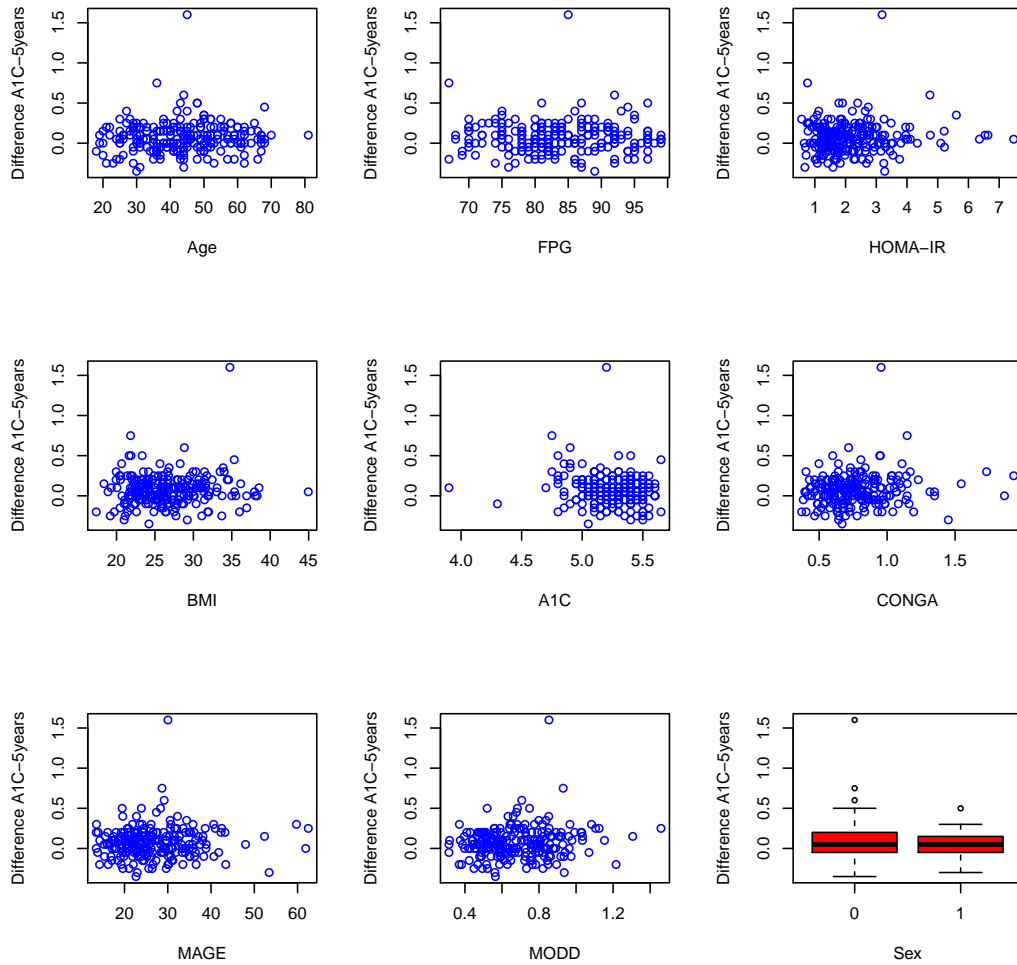


Figure 2: Marginal dependence relation between examined variables in the AEGIS database.

In order to compare the performance of these regression models, we use R^2 after including the specific missing data mechanism:

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

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.

Performance results, by using leave-one cross-validation, are: 1) $R_{noCGM}^2 = 0.61$; 2) $R_{MAGE}^2 = 0.65$; 3) $R_{TIR}^2 = 0.64$; and 4) $R_{gluco}^2 = 0.71$. Figure 3 depicts the residuals versus $A1c_{initial}$ 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 information in predicting long-term A1c changes.

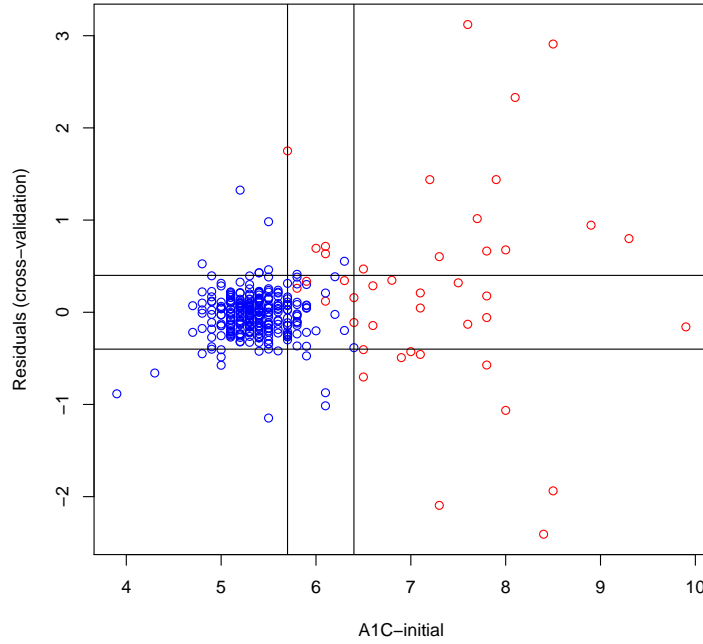


Figure 3: Residuals vs. $A1C_{initial}$ for the model that includes glucodensity as a covariate in the AEGIS database. Red circles correspond to diabetic patients

4.4. *Can we provide a characterization of those individuals whose future glucose behavior cannot be precisely predicted?*

Figure 4 depicts prediction intervals at a confidence level of 90%, after applying conformal inference (Section 3.4) 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 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 to whom the model provides an unreliable prediction, and therefore, a more personalized follow-up is advisable. Particularly, Figure 5 shows that long-term changes cannot be adequately predicted for individuals with an elevated FPG. 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.

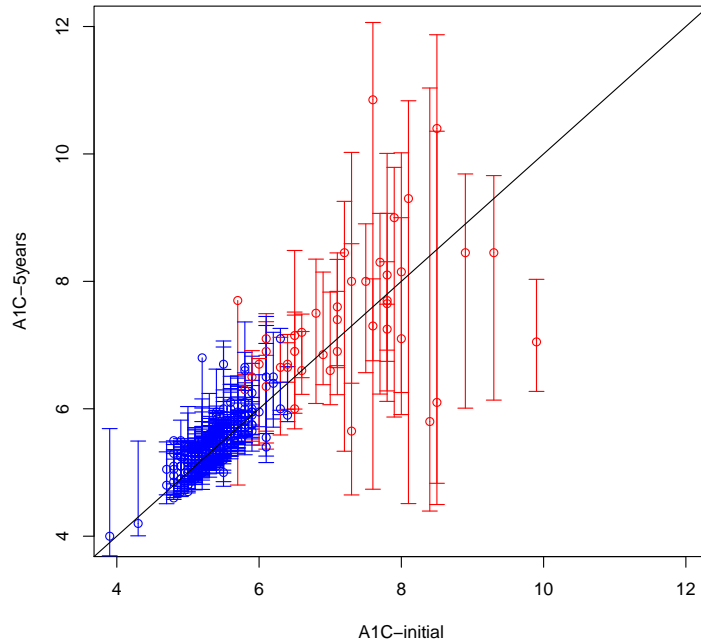


Figure 4: Prediction intervals for each response observed in the AEGIS database (90% confidence level). Red circles correspond to diabetic patients.

5. Discussion

The incidence and proliferation of diabetes are 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 with a logistic regression Makrilakis, Liatis, Grammatikou, Perrea, Stathi, Tsiligros and Katsilambros (2011). 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 Mühlenbruch, Paprott, Joost, Boeing, Heidemann and Schulze (2018). In contrast, some authors argue against using thresholds and categorizing patients into different ranges of glucose levels, and hence against defining diabetes as a homogeneous disease, resulting in an oversimplistic approximation for a heterogeneous metabolic disorder Gale (2013); Zaccardi and Khunti (2018). In this sense, some recent contributions have been made to modeling blood glucose dynamics as a function of time, with an application in predicting A1c in the short term Gaynanova, Punjabi and Crainiceanu (2020); Zaitcev, Eissa, Hui, Good, Elliott and Benaissa (2020). Furthermore, this line of research assigns a key role to the analysis of glucose excursions from CGM data, in search of a better phenotyping and corresponding progress towards the implementation of a personalized intervention Hall et al. (2018); Tsiatis (2019).

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. Firstly, glucodensity shows 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, $A1c_{initial}$, FPG, BMI, and MAGE) resulting from the detection of higher-order

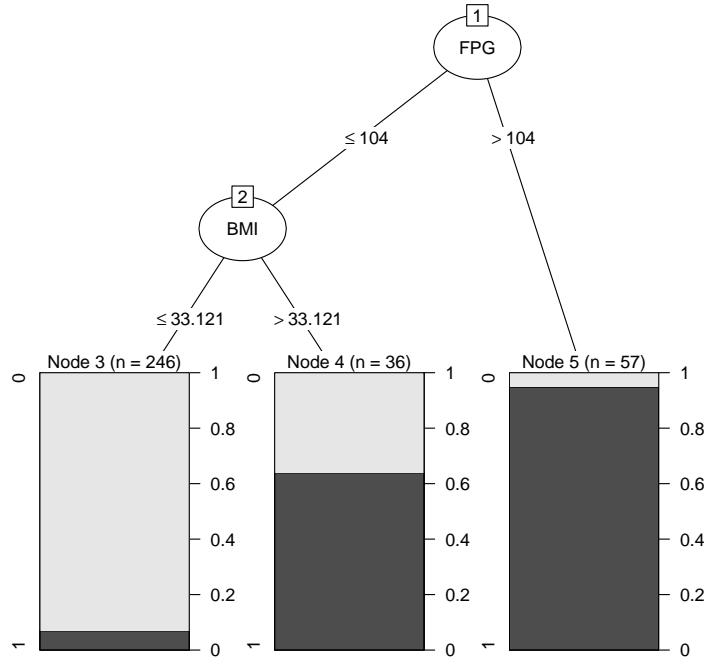


Figure 5: Clinical decision rules that allow us to identify those patients with a significant uncertainty in their $A1c_{5year}$ predictions.

interactions with $A1c_{5years}$. We then analyze the ability to predict $A1c_{5years}$ from this subset of biomarkers, with several nonparametric regression models differing on the way of including CGM data as a covariate. 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 the one reported by other authors *for short-term predictions* Gaynanova et al. (2020); Zaitcev et al. (2020). Furthermore, glucodensity demonstrates a positive impact on improving accuracy in predicting $A1c_{5years}$. Ultimately, these results enforce the prominent role of CGM data to provide a comprehensive picture of the glucose metabolism Cefalu et al. (2018), 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 Cefalu et al. (2018).

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 clinical practice biomarkers. 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.

6. 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 explanatory variables, 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 license¹.

References

- American Diabetes Association, 2019. 7. Diabetes technology: Standards of medical care in diabetes-2019. *Diabetes Care* 42, S71–S80.
- Battellino, T., Danne, T., Bergenstal, R.M., Amiel, S.A., Beck, R., Biester, T., Bosi, E., Buckingham, B.A., Cefalu, W.T., Close, K.L., et al., 2019. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes care* 42, 1593–1603.
- Beck, R.W., Bergenstal, R.M., Riddlesworth, T.D., Kollman, C., Li, Z., Brown, A.S., Close, K.L., 2019. Validation of time in range as an outcome measure for diabetes clinical trials. *Diabetes Care* 42, 400–405.
- Berg, C., Christensen, J.P.R., Ressel, P., 1984. Harmonic analysis on semigroups: theory of positive definite and related functions. volume 100. Springer.
- Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J., 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, e49–e57.
- Cefalu, W., et al., 2018. Need for regulatory change to incorporate beyond A1c glycemic metrics. *Diabetes Care* 41, e92–e94.
- Chen, J., Stern, M., Wainwright, M.J., Jordan, M.I., 2017. Kernel feature selection via conditional covariance minimization. *Advances in Neural Information Processing Systems (NIPS 2017)* 30, 6946–6955.
- Cirillo, D., Valencia, A., 2019. Big data analytics for personalized medicine. *Current opinion in biotechnology* 58, 161–167.
- Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.
- Fukumizu, K., Leng, C., 2012. Gradient-based kernel method for feature extraction and variable selection, in: *Advances in Neural Information Processing Systems*, pp. 2114–2122.
- Gale, E., 2013. Is type 2 diabetes a category error? *The Lancet* 381, 1956–1957.
- Gaynanova, I., Punjabi, N., Crainiceanu, C., 2020. Modeling continuous glucose monitoring (CGM) data during sleep. *Biostatistics*.
- Van de Geer, S.A., 2000. Applications of empirical process theory. volume 91. Cambridge University Press Cambridge.
- Gómez, A.M., Henao, D.C., Imitola Madero, A., Taboada, L.B., Cruz, V., Robledo Gomez, M.A., Rondon, M., Munoz-Velandia, O., Garcia-Jaramillo, M., Leon Vargas, F.M., 2019. Defining high glycemic variability in type 1 diabetes: comparison of multiple indexes to identify patients at risk of hypoglycemia. *Diabetes technology & therapeutics* 21, 430–439.
- Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 723–773.
- Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., Smola, A., 2007. A kernel statistical test of independence. *Advances in neural information processing systems* 20, 585–592.
- Gude, F., Díaz-Vidal, P., Rúa-Pérez, C., Alonso-Sampedro, M., Fernández-Merino, C., Rey-García, J., Cadarso-Suárez, C., Pazos-Couselo, M., García-López, J.M., Gonzalez-Quintela, A., 2017. Glycemic variability and its association with demographics and lifestyles in a general adult population. *Journal of diabetes science and technology* 11, 780–790.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 1157–1182.
- Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., McLaughlin, T., Snyder, M., 2018. Glucotypes reveal new patterns of glucose dysregulation. *Plos Biology* 16, e2005143.
- Hirsch, I.B., Sherr, J.L., Hood, K.K., 2019. Connecting the dots: validation of time in range metrics with microvascular outcomes. *Diabetes Care* 42, 345–348.
- Hofmann, T., Schölkopf, B., Smola, A.J., 2008. Kernel methods in machine learning. *The Annals of Statistics* 36, 1171–1220.
- Hu, F.B., Satija, A., Manson, J.E., 2015. Curbing the diabetes pandemic: the need for global policy solutions. *JAMA* 313, 2319–2320.
- Kang, J.D., Schafer, J.L., et al., 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science* 22, 523–539.
- Kosorok, M.R., Laber, E.B., 2019. Precision medicine. *Annual Review of Statistics and its Application* 6, 263–286.

¹<https://gitlab.citius.usc.es/marcos.matabuena/RKHSmissing>

- Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L., 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association* 113, 1094–1111.
- Lei, L., Candès, E.J., 2020. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*.
- Li, B., 2018. Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics* 46, 79–103.
- Liang, T., Rakhlin, A., et al., 2020. Just interpolate: Kernel “ridgeless” regression can generalize. *Annals of Statistics* 48, 1329–1347.
- Little, R.J., Rubin, D.B., 2019. *Statistical analysis with missing data*. volume 793. John Wiley & Sons.
- Liu, T., Goldberg, Y., et al., 2020. Kernel machines with missing responses. *Electronic Journal of Statistics* 14, 3766–3820.
- Makrilakis, K., Liatis, S., Grammatikou, S., Perrea, D., Stathi, C., Tsiligras, P., Katsilambros, N., 2011. 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* 37, 144–151.
- Matabuena, M., Petersen, A., Vidal, J.C., Gude, F., 2020. Glucodensities: a new representation of glucose profiles using distributional data analysis. *arXiv:2008.07840*.
- Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning* 10, 1–141.
- Mühlenbruch, K., Paprott, R., Joost, H.G., Boeing, H., Heidemann, C., Schulze, M.B., 2018. 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* 6, e000524.
- Peters, A., Ahmann, A., Battelino, T., Evert, A., Hirsch, I., Murad, M., Winter, W., Wolpert, H., 2016. Diabetes technology-continuous subcutaneous insulin infusion therapy and continuous glucose monitoring in adults: An endocrine society clinical practice guideline. *J Clin Endocrinol Metab.* 101, 3922–3937.
- Rodbard, D., 2018. Glucose variability: a review of clinical applications and research developments. *Diabetes technology & therapeutics* 20, S2–5.
- Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A.A., Ogurtsova, K., et al., 2019. 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* 157, 107843.
- Schölkopf, B., Herbrich, R., Smola, A.J., 2001. A generalized representer theorem, in: *International conference on computational learning theory*, Springer. pp. 416–426.
- Schölkopf, B., Smola, A.J., Bach, F., et al., 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
- Schork, N.J., 2015. Personalized medicine: time for one-person trials. *Nature* 520, 609–611.
- Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 2263–2291.
- Selvin, E., Crainiceanu, C.M., Brancati, F.L., Coresh, J., 2007. Short-term variability in measures of glycemia and implications for the classification of diabetes. *Archives of internal medicine* 167, 1545–1551.
- Topol, E.J., 2010. Transforming medicine via digital innovation. *Science Translational Medicine* 2, 16cm4.
- Tsiatis, A., 2007. *Semiparametric theory and missing data*. Springer Science & Business Media.
- Tsiatis, A.A., 2019. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC Press.
- Vermeulen, K., Vansteelandt, S., 2015. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* 110, 1024–1036.
- Yang, L., Lv, S., Wang, J., 2016. Model-free variable selection in reproducing kernel hilbert space. *The Journal of Machine Learning Research* 17, 2885–2908.
- Zaccardi, F., Khunti, K., 2018. Glucose dysregulation phenotypes - time to improve outcomes. *Nature Reviews Endocrinology* 14, 632–633.
- Zaitcev, A., Eissa, M.R., Hui, Z., Good, T., Elliott, J., Benaissa, M., 2020. A deep neural network application for improved prediction of HbA1c in Type 1 diabetes. *IEEE Journal of Biomedical and Health Informatics* 24, 2932–2941.
- Zheng, Y., Ley, S.H., Hu, F.B., 2018. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews Endocrinology* 14, 88–98.