Challenges associated with missing data in electronic health records: a case study of a risk prediction model for diabetes using data from Slovenian primary care

Abstract

The increasing availability of data stored in electronic health records (EHRs) brings substantial opportunities for advancing patient care and population health. This is however fundamentally dependant on the completeness and quality of data in these EHRs. We sought to use EHR data to populate a risk prediction model for identifying patients with undiagnosed type 2 diabetes mellitus. We however found substantial (up to 90%) amounts of missing data in some healthcare centres. Attempts at imputing for these missing data or using reduced dataset by removing incomplete records resulted in a major deterioration in the performance of the prediction model. This case study illustrates the substantial wasted opportunities resulting from incomplete records by simulation of missing and incomplete records in predictive modelling process. Government and professional bodies need to prioritise efforts to address these data shortcomings in order to ensure that EHR data are maximally exploited for patient and population benefit.

Keywords: Databases and data mining, Electronic health records, Primary care, Quality control, Type 2 diabetes, Missing data.

Introduction

Fundamental to the concept of the learning healthcare system is creating the infrastructure to use the data generated as a by-product of healthcare to promote continuous quality improvements. Collection of structured data is of high importance in derivation of effective chronic disease prediction algorithms such as QDRisk or QRisk2 for predicting the development of type 2 diabetes mellitus (T2DM) and cardiovascular disease, respectively [1,2]. A major challenge to achieving this however is the recording of high quality data in electronic health records (EHRs) [3].

A recent study by Creswell et al. [4] identified a range of micro, meso and macro level factors that contribute to better use and repeated reuse of demographic, process and healthcare data to improve the quality and safety of care. The lack of motivation and prioritisation by professionals to enter data were identified as key barriers that need to be overcome. As observed by Kratz [5] it is expected that more and more technologies and novel approaches will become available to minimize the need for manual data entry. O'Brien et al. [6] proposed the redesign of the documentation process for nurses, where they concluded, that one can expect the value of nursing data will increase in the future and can represent a key differentiator for the economic success of the healthcare institutions by higher effectiveness and lower costs.

This case study aims to raise awareness about a major challenge in exploiting EHR data for patient and population health. It is based on an empirical study using data collected in three healthcare centres in Slovenia to observe the impact of missing data on predictive performance in detecting undiagnosed T2DM. To estimate the effect of missing data, we simulated the effects of different amounts of missing values in a complete dataset from two out of three healthcare centres included in the study.

Methods

Study design and population

We conducted a cross-sectional study using anonymised EHR data from three healthcare centres in Slovenia. EHRs from primary healthcare level providers were used to extract the routinely collected data from the Finnish Diabetes Risk Score (FINDRISC) [7] questionnaire (an internationally used risk prediction model) including physiological data in numerical format. Initially, we extracted records for 17,761 regular health check-ups in all participating healthcare centres (HCs) of which 13,072 records were not associated with any diagnosis and were used in further analysis. Table 1 presents the information on missing data from the three healthcare centres (HC1, HC2 and HC3). Out of the three healthcare centres, HC1 contained the highest fraction of records with at least one missing value (n=1,760, 99.7 %).

We observe less records with missing values in HC2 (n=8,751, 95.9 %) and HC3 (n=1,603, 73.6 %) with significantly higher percentages of complete records in comparison to HC1. Since our aim was to simulate missing and incomplete records while building a predictive model, we kept only complete records containing all FINDRISC variables, present fasting plasma glucose level (FPGL) measurement, and no indication of type 2 diabetes mellitus (T2DM) diagnosis (n = 952) for further analysis. Due to the extremely high number of records (1760, 99.7 %), with missing data, we excluded HC1 from further analysis. Out of 952 participants from the two remaining institutions, 477 (50.1%) were female with mean age of 55.4 (95% CI: 54.3-56.5) and 475 (49.9%) male with an average age of 55.3 (95% CI: 54.3-56.3) years.

Predictor variables

Five numeric (age, height, weight, body mass index (BMI), waist circumference) and six dichotomous variables (sex and five FINDRISC questions) were used to build a model. The five FINDRISC questions were related to daily physical activity (more than 30 minutes), history of high blood pressure, history of high blood glucose, fruit and vegetable consumption, and diabetes in family. All 11 variables were used as input variables for derivation of a predictive model. Table 2 presents a summary of all predictor values in the two healthcare centres included in the study.

Table 1. Summary information for all variables used in the study including the percentage of records with missing values (in brackets) in three healthcare centers.

	Missing data (%)		
	HC1	HC2	HC3
Age (years)	0.0	0.0	0.0
Weight (kg)	34.4	18.9	39.3
Height (m)	44.7	21.7	48.5
BMI (kg/m^2)	46.1	21.9	48.7
Waist circumference (cm)	46.0	28.2	54.7
Female (%)	2.4	0.2	0.0
Physically active? (%)	99.6	95.3	66.3
High blood pressure history (%)	99.5	95.4	66.9
High blood glucose history (%)	99.5	95.3	66.5
Fruit and vegetables consumption (%)	99.5	95.3	66.2
Diabetes in family (%)	99.5	95.3	66.9
Fasting plasma glucose (mmol/l)	40.1	21.2	15.2
Complete records	0.3	4.1	26.4

Outcome

The outcome in the study was prediction of undiagnosed T2DM based on FPGL. We used a FPGL threshold of 7.0 mmol/l (126 mg/dl) resulting in a very imbalanced classification problem where number of negative samples strongly outweighs the number of positive samples (10.4 % positive samples, n = 99).

Table 2. Summary information presenting mean (with corresponding 95% CI) or frequency values for all predictor variables from complete records used in the study.

	HC2 (n=377)	HC3 (n=575)
Age (years) [95% CI]*	56.5 [55.3-57.6]	54.7 [53.7-55.6]
Weight (kg) [95% CI]	83.9 [82.2-85.5]	81.7 [80.4-83.1]
Height (m) [95% CI]	168.9 [168-169.9]	167.3 [166.4-168.1]
BMI (kg/m^2) [95% CI]	29.4 [28.8-29.9]	29.7 [28.6-30.8]
Waist circumference (cm) [95% CI]	98.4 [97-99.7]	94.4 [93.4-95.4]
Female [n (%)]	175 (46.4%)	302 (52.5%)
Physically active? [n (%)]	215 (57%)	292 (50.8%)
High blood pressure history [n (%)]	132 (35%)	206 (35.8%)
High blood glucose history [n (%)]	149 (39.5%)	131 (22.8%)
Fruit and vegetables consumption** [n (%)]	329 (87.3%)	510 (88.7%)
Diabetes in family [n (%)]	160 (42.4%)	172 (29.9%)
Fasting plasma glucose (mmol/l)	6.1 [6.0-6.2]	5.7 [5.6-5.8]

Statistical analysis

Least absolute selection and shrinkage operator (LASSO) regularisation based logistic regression as defined and implemented by Tibshirani [8] was used to build the predictive models. The LASSO was chosen to avoid multicollinearity problems, especially in cases where higher proportion of missing data were introduced.

The performance of the models was assessed using the following measures:

1. General predictive performance was tested using the C-statistic that measures the balance between sensitivity and specificity by calculating the area under the receiver operator curve (AUC) [9].
2. The accuracy of probabilistic predictions was measured using Brier score [10].
3. Additionally, we calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the models [11].

Repeated five-fold cross-validation was used to calculate the average value of performance measures with corresponding confidence intervals. In five-fold cross-validation evaluation, we divide the initial set of samples into five groups containing approximately equal number (20%) of samples. In each iteration, we selected four different groups of data for derivation of the predictive model and test the model performance on the remaining group of samples. Each five-fold cross-validation was repeated 100 times on randomly reshuffled data.

The initial set of experiments measured the predictive performance of the model built on all data available during the cross-validation process. This step was followed by two alternative experiments that simulated two approaches to missing and incomplete data. The first approach ("exclude") simulated a scenario where the predictive model was built only from complete records. The second approach used the well-known missing value imputation method called missForest [12, 13] to allow inclusion of all records in the model-building phase. The missing value imputation by missForest was conducted using the default settings proposed by the authors of missForest in [12] with an exception in number of decision trees built for each variable that was set to 20. In both cases, we injected the missing data by:

1. Randomly selecting different percentage of samples (ranging from 5 to 90 in steps of 5).
2. Randomly selecting affected variables in samples selected in step 1. Variables were selected randomly using a weighting scheme that took into account the percentage of missing values for specific variable in the initial dataset (Table 1). By increasing the probability of missing value injection in variables with higher percentage of missing data, we achieved a realistic distribution of missing values.
3. Injecting a missing value for each selected (sample, variable) data point.

Ethics and reporting

The institutional ethics committee approved this study. We have followed the “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis” (TRIPOD) initiative guidelines [14].

Results

In this section, we present the results of the predictive performance evaluation for all three scenarios discussed above. As it can be observed from Figure 1 and as expected, the exclusion of incomplete records resulted in a significantly stronger decrease in AUC compared to the alternative scenario of replacing missing values using missForest. For each of 100 runs of the experiment, we initially built a LASSO model on full training sets that achieved a mean AUC of 0.850 (95% CI: 0.846-0.854) on test sets of the 5-fold cross-validation. The 95% CIs of predictive performance using LASSO on a full datasets compared to results of the missForest approach overlapped up to the point with 20% of injected missing values. At 20% of missing data, the missForest approach achieved and AUC of 0.841 (95% CI: 0.837-0.846). With ≥20% of missing values injected, the loss of performance was significant when missForest based models were used.

For “exclude” scenario, this value lay at 35% of missing values. There were no significant differences in AUC between the missForest and “exclude” approaches all the way up to 75%

of injected missing values where the confidence intervals stop overlapping and “exclude” starts to return significantly inferior results.

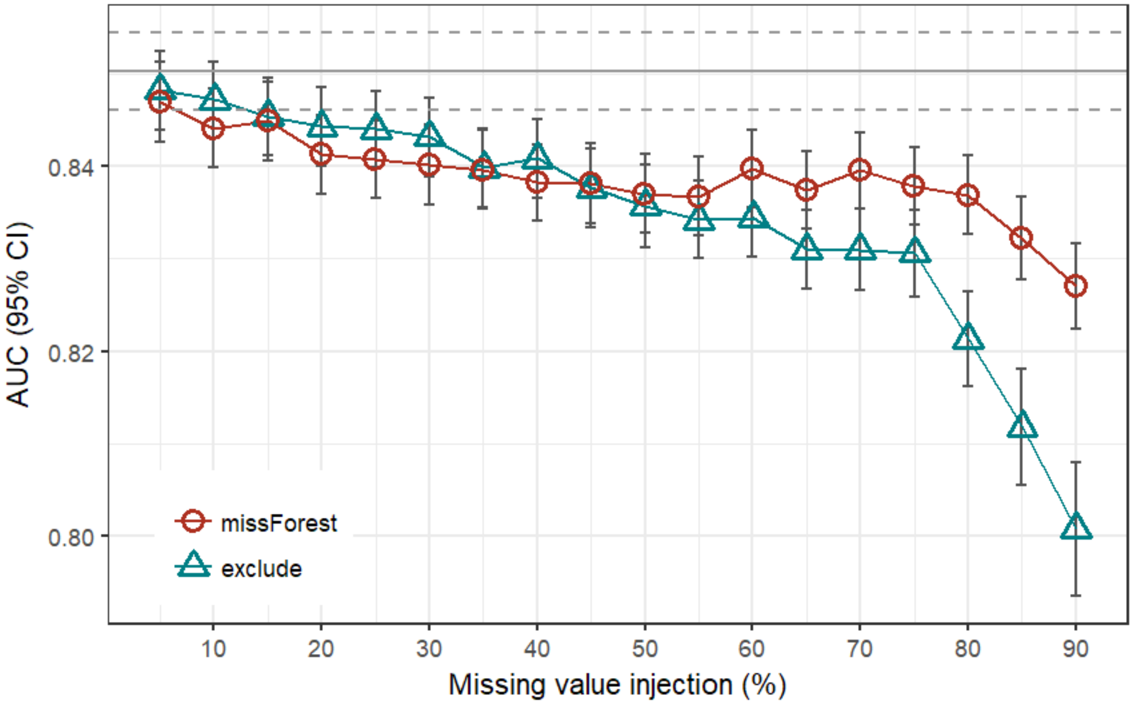

Figure 1. Comparison of AUC for two approaches to classification in presence of randomly injected missing values compared to a model built on a complete dataset (grey line with 95% CI in dashed lines).

The results from the “exclude’ scenario, represented by triangles in Figure 1 observed from the right to the left can be used to estimate the relation between the sample size and performance of the classification model. This way the rightmost point represents a result where only 10% (n = 76) of the available data were used to build a model that achieved an average AUC of 0.801 (95% CI: 0.794-0.808). The second rightmost point represents a model performance when the number of available samples doubled (n = 152) resulting in an average AUC of 0.812 (95% CI: 0.805-0.818). Following the increase in performance from the right to the left (Figure 1) we can estimate the potential of improving the prediction performance by adding the records beyond 100 % of all available records. At the same time, we can observe the confidence intervals that are becoming narrower with the rising number of samples.

Observing the alternative performance metrics, Figure 2 shows no significant differences between the tested scenarios, except in cases with extremely high number of missing values. Figure 2 also shows the PPV metric with relatively low values for both approaches, a result commonly seen in imbalanced classification problems. PPV fell significantly below the performance of the basic model when the number of missing values passed 40%. The true effectiveness of the missing value imputation versus exclusion can be observed in the last plot of Figure 2, where we can observe the fraction of persons selected for screening by the predictive model. The missForest based model is not only better in AUC, but also selects a lower number of persons for further examinations at the same time, especially with the highest fraction of missing values.

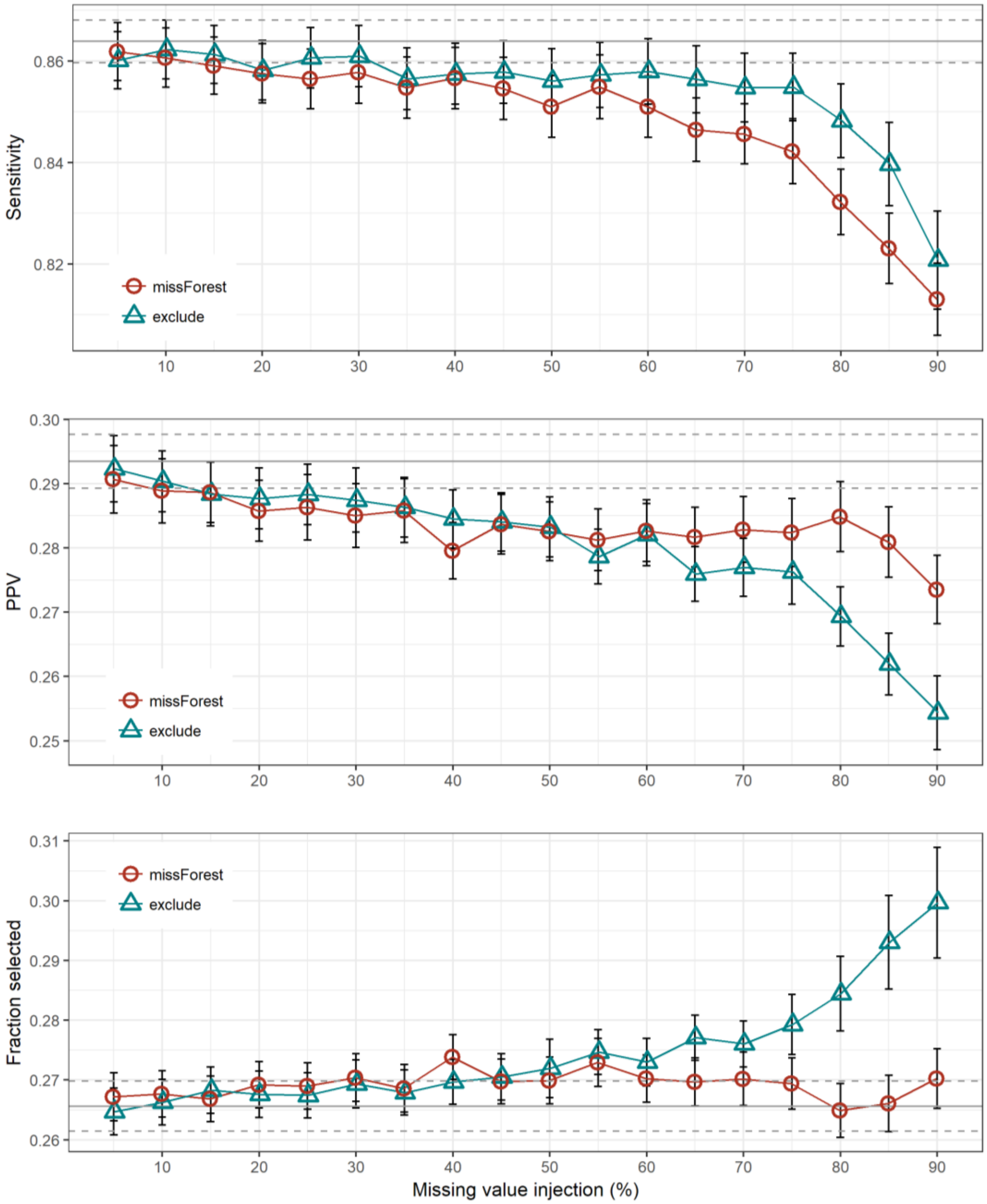

Figure 2. Sensitivity, positive predictive value (PPV) and fraction of selected samples by the predictive model for different levels of missing values.

Discussion

In this study, we demonstrate the effects of incomplete and missing primary healthcare data in developing a predictive model for undiagnosed T2DM. We observed a significant drop in predictive performance for the data imputation approach with only 20% of missing data. In data obtained from three healthcare centres in Slovenia, we observed high fractions of incomplete data ranging from 73.6% up to 99.8% observed in one of the healthcare centres. Compared to results achieved in a similar study [15] using a more common FINDRISC questionnaire to predict presence of undiagnosed T2DM, a model built on full dataset in this study performed very well. However, there was a significant difference in approach to data collection in both studies. There are some very different challenges a researcher will have to confront when the data are collected using a paper-and-pencil questionnaire in comparison to analysis of routinely collected data from EHRs. In this study, we focused on missing data that frequently originates from lack of motivation to enter the data [4]. The fraction of missing data observed in this study was high, especially for variables where reporting to the National Institute of Public Health was not mandatory.

High variance of results that was observed in some results (Figure 1) can be attributed to different factors, where class imbalance represents an important factor. With only 9.1% of positive samples, where we could confirm the undiagnosed T2DM, our problem could be classified as highly-imbalanced problem. On the other hand, in most similar studies one can meet even lower fractions of positive class, especially in cases where the sample is limited to younger population groups [16, 17].

Our observations from three healthcare centres in Slovenia suggest that there are extremely large proportions of incomplete records stored in medical information systems on the primary healthcare level. To demonstrate the impact of incomplete records on the final predictive performance of the models build on such data, we conducted an experimental study. By randomly injecting missing variable values we were able to show a significant drop in performance when 35% of records were either missing or were incomplete. The impact of missing values on predictive performance can be reduced by using a missing value replacement using methods such as missForest, but their contribution seems to significantly affect the predictive performance only in cases with an extremely high percentage of missing values. This is also true in cases where the fraction of persons selected for further examinations is observed. By selecting more cases than missForest model, the usage of "exclude" model leads to higher costs for the healthcare system. When observing the performance of the classifier built in exclusion scenario, one can observe a positive trend when the percentage of missing values decreases, meaning access to more records would further improve the performance of the classifier.

However, even data imputation techniques can result in misleading conclusions, especially in case of so called missing at random (MAR) scenarios [18]. On the other hand, researchers also report misleading results in case of using simple approaches for dataset modification such as discarding observations or ad-hoc replacement of missing values [19]. To some extent, we can solve this kind of problems by using classification models that can implicitly handle missing values (e.g. decision trees [20]).

Conclusions

Recent studies emphasise the importance of motivation for entering the data to improve the learning health systems of the future. To raise awareness on this issue, the health data science community needs to organise a number of events that highlight both the clinical and research importance and value on striving towards more complete datasets. One such use case was demonstrated in this study. Using simulation of missing values, we were able to detect the significant drop of performance even in cases where only one-third of records were missing or incomplete.

Acknowledgments

We are grateful to Teo Bizjak and Jozica Leskovsek from Nova vizija d.d. for providing the technical and administrative support to access the EHR data used in this study.

References

[1] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. Bmj. 2008 Jun 26;336(7659):1475-82.

[2] Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. Br Med J. 2009;338:15.

[3] Jones KH, Laurie G, Stevens L, Dobbs C, Ford DV, Lea N. The other side of the coin: Harm due to the non-use of health-related data. International Journal of Medical Informatics. 2017 Jan 31;97:43-51.

[4] Cresswell K, Smith P, Swainson C, Timoney A, Sheikh A. Establishing data-intensive healthcare: the case of Hospital Electronic Prescribing and Medicines Administration systems in Scotland. Journal of Innovation in Health Informatics. 2016 Oct 4;23(3):572-9.

[5] Kratz A. Electronic reporting of all reference laboratory results: An important step toward a truly all-encompassing, integrated health record. Health informatics journal. 2015 Feb 19:1460458215569004.

[6] O'Brien A, Weaver C, Hook ML, Ivory CH. EHR documentation: the hype and the hope for improving nursing satisfaction and quality outcomes. Nursing administration quarterly. 2015 Oct 1;39(4):333-9.

[7] Makrilakis K, Liatis S, Grammatikou S, Perrea D, Stathi C, Tsiligros P, Katsilambros N. 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. 2011 Apr 30;37(2):144-51.

[8] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011 Jun 1;73(3):273-82.

[9] Pencina MJ, D’Agostino RB. Evaluating discrimination of risk prediction models: the C Statistic. JAMA. 2015 Sep 8;314(10):1063-4.

[10] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.). 2010 Jan;21(1):128.

[11] Simon R. Sensitivity, specificity, ppv, and npv for predictive biomarkers. Journal of the National Cancer Institute. 2015 Aug 1;107(8):djv153.

[12] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112-8.

[13] Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, Marrero J, Zhu J, Higgins PD. Comparison of imputation methods for missing laboratory data in medicine. BMJ open. 2013 Aug 1;3(8):e002847.

[14] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC medicine. 2015 Jan 6;13(1):1.

[15] Štiglic G, Fijačko N, Stožer A, Sheikh A, Pajnkihar M. Validation of the Finnish Diabetes Risk Score (FINDRISC) questionnaire for undiagnosed type 2 diabetes screening in the Slovenian working population. Diabetes Research and Clinical Practice. 2016 Oct 31;120:194-7.

[16] Vandersmissen GJ, Godderis L. Evaluation of the Finnish Diabetes Risk Score (FINDRISC) for diabetes screening in occupational health care. International journal of occupational medicine and environmental health. 2015 Jan 1;28(3):587-91.

[17] Stiglic G, Pajnkihar M. Evaluation of Major Online Diabetes Risk Calculators and Computerized Predictive Models. PloS one. 2015 Nov 11;10(11):e0142827.

[18] Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, Young BE, Graham CH, Costa GC. Imputation of missing data in life-history trait datasets: which approach performs the best?. Methods in Ecology and Evolution. 2014 Sep 1;5(9):961-70.

[19] Srinivasan K, Currim F, Ram S, Lindberg C, Sternberg E, Skeath P, Najafi B, Razjouyan J, Lee HK, Foe-Parker C, Goebel N. Feature Importance and Predictive Modeling for Multi-source Healthcare Data with Missing Values. InProceedings of the 6th International Conference on Digital Health Conference 2016 Apr 11 (pp. 47-54). ACM.

[20] Kokol P, Pohorec S, Štiglic G, Podgorelec V. Evolutionary design of decision trees for medical application. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012 May 1;2(3):237-54.