# Missing Data Imputation for Individualised CVD Diagnostic and Treatment

Sitalakshmi Venkatraman [1], Andrew Yatsko [2], Andrew Stranieri [2], Herbert F Jelinek [3]

[1] School of Engineering, Construction and Design, Melbourne Polytechnic, Preston, Australia
[2] Centre for Informatics and Applied Optimisation, Federation University, Mt Helen, Australia
[3] School of Community Health, Charles Sturt University, Albury, Australia

## Abstract

*Cardiac health screening standards require increasingly more clinical tests consisting of blood, urine and anthropometric measures as well as an extensive clinical and medication history. To ensure optimal screening referrals, diagnostic determinants need to be highly accurate to reduce false positives and ensuing stress to individual patients. However, the data from individual patients partaking in population screening is often incomplete. The current study provides an imputation algorithm that has been applied to patient-centered cardiac health screening. Missing values are iteratively imputed in conjunction with combinations of values on subsets of selected features. The approach was evaluated on the DiabHealth dataset containing 2800 records with over 180 attributes. The results for predicting CVD after data completion showed sensitivity and specificity of 94% and 99% respectively. Removing variables that define cardiac events and associated conditions directly, left 'age' followed by 'use' of anti-hypertensive and anti-cholesterol medication, especially statins among the best predictors.*

## 1. Introduction

Missing values (MVs) may appear in database records for many reasons, presenting an obstacle for data processing. Numerous statistical methods have been advanced to deal with missing data, however most are deployed during the model-building phase, which implies making assumptions about variables and how they are related [1,2].

In clinical practice MVs are unavoidable, presenting a problem for disease classification or diagnosis [2,3]. It is sometimes possible to run a data mining algorithm in a mode that sidesteps MVs, but most methods perform better with complete data. Simple deletion of attributes and instances containing MVs is often not viable as this is able to distort perception of the data. A pre-processing step proposed in this paper, which makes surrogate entries, offers a systematic approach to the MV imputation problem.

MVs are often substituted by their attribute mode or mean, depending on whether the attribute type is categorical or numerical. However, by constraining other involved variables, the mode or mean can be evaluated more specifically [3,4]. Albeit, this generic method does not guarantee that classification is optimal as the substituted values are dependent on population cohorts from which they are evaluated.

In this paper, an approach is presented where a MV imputation algorithm employs well established risk factors as categorical types to guide the selection of substitute values. This improves on the current MV imputation paradigm for classification problems [3].

## 2. Context

Residents of South-East Australia have an opportunity to attend an annual Diabetic Health screening for type 2 diabetes mellitus (T2DM), cardiovascular disease (CVD) and hypertension (HT). Data on over 180 variables was collected, though many variables have MVs [4]. The MVs hamper application of data mining algorithms enabling diagnostic of CVD, HT or T2DM, and removal of records and attributes that have MVs has a very limited application, as previously stated. The MV imputation operates on different subsets of features that are all well-established risk factors for T2DM, CVD and HT. We call each subset a selector set. Selector set features were: the disease statuses, medication use, fasting glucose, systolic blood pressure (SBP), age, waist-circumference-to-height ratio (WCHR) and some others.

## 3. Methods

The MV imputation algorithm is represented in Figure 1. For example, if CVD, T2DM and HT were identified as the selector set, each combination of values of the three variables would be attempted in turn. For instance, one combination of the selector feature values is {yes, yes, yes} drawn from the domain of CVD: {yes, no}, T2DM : {yes, no} and HT: {yes, no}. The records with values on the selector features {yes, yes, yes} are then assembled.

Within the assembly, each non-selector feature that has a MV is then replaced with the mean for that set if the feature is continuous or the mode if it is categorical. This is repeated for all combinations of the selector set features to comprise a single pass.

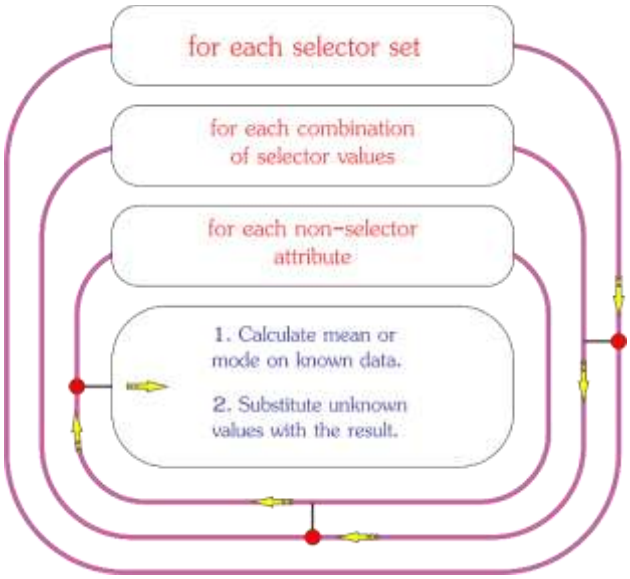


Figure 1. Schematics of the proposed algorithm for imputation of missing data

Some combinations of selector features resulted in datasets that were too small or too riddled with MVs so not all MVs could be typically set in a single pass. Therefore, the selector set was dynamically modified by removing some features and returning some previously removed features until all selector combinations had been tried. Any features that still have MVs are then filled using the global mean or mode from the whole data, but this was not required.

The selection of features to form the selector set at each pass is performed using an information measure used in the Info-Neighbour algorithm [3]. Selector features with higher Information Gain (IG) persist longer in the selector set through multiple passes than those with lower IG. When continuous features such as age, SBP, fasting glucose and WCHR were included in the selector set, they were required to be discretised so that a dataset pertaining to each combination of selector set values could be assembled. This was achieved by subdividing the range into a number of intervals of fixed frequency [3] with boundaries then adjusted to match thresholds known from the literature for these variables.

MVs were required to be assigned where not all subject's disease statuses were known. For this CVD, T2DM and HT statuses were set using other features or related disorders following clinical guidelines or practices accepted in the field [5].

The MV imputation method was first developed for T2DM classification [4]. The same algorithm can be used to cater for CVD and HT classification, which are connected classification problems. We imposed a minimum on the size of a data subset drawn to calculate substitute values for missing entries. Within these subsets, the means of continuous attributes are recalculated if some of the input is vastly different from the mean and therefore potentially erroneous.

Upon data imputation, our primary interest was to be able to classify data according to CVD status. To verify the predictability of CVD, the Info-Neighbour and the Naïve Bayesian classifiers were deployed [3]. The classification accuracy was evaluated by using leave-one-out cross-validation [6] in each of the eight years recorded and sensitivity and specificity determined. Balanced accuracy is the mean of the two quantities [3].

Classification of CVD after data completion was performed with 205 attributes including the status for each record listed. This number does not include features that are conventionally used to set CVD status. Specifically, the information about cardiovascular events and cardioneuropathic disorders as well as use of anticoagulant and antiplatelet medications [5] was withheld. The classification task was also attempted with a reduced set of just over 90 features with high IG.

Additionally, a number of features were individually evaluated for the ability to predict CVD status. This was done by calculating optimal cut-off levels that maximise the balanced accuracy of classification, which is beyond the scope of this account. The accuracy as the objective of optimisation was evaluated on the full dataset (all years). The accuracy of classification using the obtained cut-off levels was then tested by distributing the records randomly five times into two equally sized folds proportionally to class memberships in the full dataset, which is regarded an unbiased approach [4].

## 4. Results

The data of 6776 records for 847 participants over 8 years was assembled presenting a missing value rate of 36%. The results plotted as the receiver operating characteristic chart are presented in Figure 2. The high accuracy of the Info-Neighbour and Naïve Bayesian classifiers is evident. There are clearly two clusters, one for each classifier regardless of whether the full or reduced set of features was used.

Evaluation of classification accuracy is specific to the dataset where the eight subsets corresponding to the eight years recorded contain the same number of records with no participant repeated in any given year. Also the composition of the data ensures that the subsets are independent from each other as much as possible. With all features, the Info-Neighbour method showed specificity, sensitivity and balanced accuracy of 99±1%, 94±3% and 97±2%, respectively. For the reduced feature-

set the results were 99±1%, 95±4% and 97±2%, respectively. By the Naïve Bayesian method the means and standard deviations for specificity, sensitivity and balanced accuracy were 84±3%, 82±1% and 83±2%, respectively, for all features, and 85±3%, 84±2% and 84±2%, respectively, with the reduced set of top performing features. It is evident that the reduced feature-set is no less accurate than the full set.

The best performing continuous features are given in Table 1 together with the accuracy achieved with optimal cut-off levels. Assuming that levels increase when corresponding cut-offs are reached, across the boundary the decision (i.e. CVD status) is either 'yes' or 'no', as shown.

Values of any nominal attribute can be distributed into two groups in a number of ways to classify a binary output, which was in the current case CVD status, which can be either 'yes' or 'no'. These groups can be given appropriate names and one-to-one correspondence established between the groups and the two data classes. Obviously, no grouping is required for attributes with only two values.

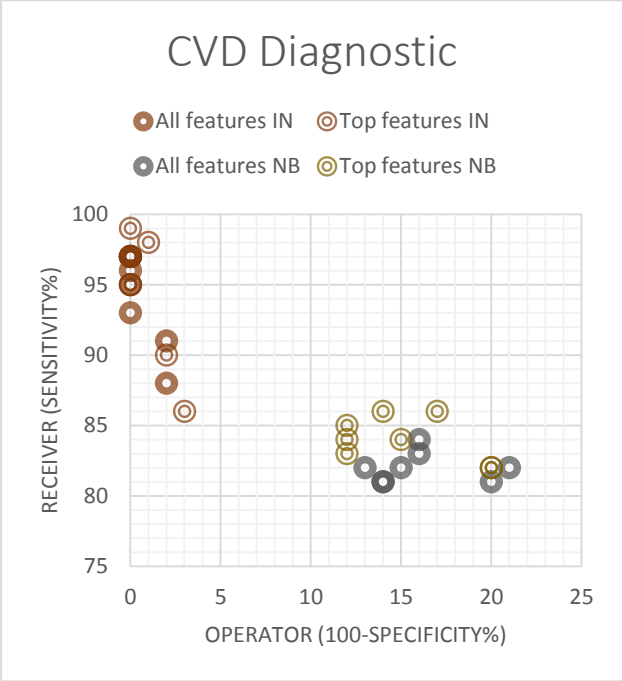


Figure 2. Classification accuracy using all or only top features by two classifiers on eight samples

Table 2 lists the nominal group values that optimally correspond to a specific CVD status for the best performing variables of the nominal type. The heart function can be graded either as 'normal' or 'abnormal' with a number of levels to the latter. The foot reflex is either 'absent' or 'reduced' – the two grouped as 'weak', or else 'present'.

From Tables 1 and 2 it is evident that age and use of antihypertensive medication are the main contributors of CVD diagnostic outcomes, while also being included in the Framingham CVD risk calculation [7]. From Table 2, HT (as a substitute for SBP) and T2DM statuses are also influential factors used in the formula. To enter the SBP term for the purpose of evaluation we regarded participants with no CVD as untreated and those with CVD as treated for high blood pressure.

Table 1. CVD diagnostic cut-off levels and balanced accuracy (BA) for continuous features

| Name * | Cut-off | CVD | BA % |
|---|---|---|---|
| CVD Risk (%) by BMI | 26.86 | yes | 72±1 |
| CVD Risk (%) by TC&HDL | 18.23 | yes | 71±1 |
| Age (years) | 62.50 | yes | 69±0 |
| GSSG (μmol/L) | 356.7 | yes | 68±0 |
| LDL (mmol/L) | 2.687 | no | 66±0 |
| Homocysteine (mmol/L) | 9.215 | yes | 64±1 |
| IL-6 (pg/mL) | 17.28 | yes | 64±0 |
| GSH (μmol/L) | 1658 | yes | 63±1 |
| D-dimer (mmol/L) | 0.6161 | yes | 63±1 |
| HbA1c (%) | 6.020 | yes | 63±1 |
| Standing 3min PP (mmHg) | 49.95 | yes | 62±1 |
| HRV DFA32 | 85.59 | no | 62±1 |
| HRV RR (msec) | 1010 | yes | 59±0 |
| HRV HR (1/min) | 59.95 | no | 58±1 |

*BMI – body mass index; TC – total cholesterol, HDL – high density lipoprotein cholesterol, LDL – low density lipoprotein cholesterol; GSSG - glutathione disulphide, GSH - reduced glutathione; HbA1c - glycated hemoglobin; PP = SBP - DBP – pulse pressure; HRV – heart rate variability, DFA32 – detrended fluctuation analysis using 32 instance segments of a transformed time series of heart beats, RR – mean time between consecutive heart beats by ECG R waves.

Table 2. CVD status optimal value correspondence and balanced accuracy (BA) for nominal features

| Name | Value | CVD | BA % |
|---|---|---|---|
| Antihypertensive meds | taken | yes | 69±1 |
| Heart Function | normal | no | 69±1 |
| Antilipidemic meds | taken | yes | 66±1 |
| Foot Reflex | weak | yes | 64±1 |
| HT Status | yes | yes | 63±1 |
| T2DM Status | yes | yes | 61±1 |

## 5. Discussion

Recent research has been on the rise in applying different data mining techniques for CVD classification. Applying hybrid models such as genetic algorithms with neural network weights an accuracy of 89% has been reported [8]. These studies have predominantly replaced MVs with the global mean for each attribute. Our research is unique as the novel imputation algorithm is based on dynamic clustering of several cardiac risk factors, including biomarkers and is capable of imputing MVs on an individual basis and spans several years. The proposed Info-Neighbour method [3] achieved an accuracy of 97%.

The Framingham CVD risk [7] featured in Table 1 has proven its utility internationally [9] and is the best CVD predictor in our study. The Framingham CVD risk accounts for age, SBP and treatment for it, smoking and diabetic status, and additionally either body mass index (BMI) or cholesterol levels (TC&HDL). It is calculated separately for men and women. It is often recommended that an all-inclusive treatment for CVD were started as early as at 20% risk in clinical practice. Here we find the optimal level by BMI higher at 27%. It is likely that due to antilipidemic medication the optimal risk by TC and HDL instead of BMI is much lower, demonstrating the effect of risk reduction. Statins by far are the best representatives of the antilipidemic group and are known to significantly reduce LDL (Table 1) and TC, while in absence of medication cholesterol levels are known to correlate strongly with BMI [4].

In the current study BMI was not shortlisted as a strong predictor of CVD. WCHR, which is invariant with respect of gender, similar to BMI, performed much better in agreement with previous research [4].

There are some highly anticipated and performing correlates of CVD status such as the ECG (heart function) abnormality and the weakness of reflex in feet (Table 2) that are not in the formula of CVD risk. Although, with the exception of CVD risk, Table 1 and 2 results cannot be regarded as strong, suggesting that these attributes may not be directly related to CVD if taken individually but may confer better accuracy if applied in combination with other attributes. Some variables from Table 1 are regarded as emergent markers for CVD. For instance, D-dimer is a possible biomarker of endothelial dysfunction associated with atherosclerosis and CVD [10]. Heart rate variability (HRV) measures obtained from the raw ECG were also identified [11].

## 6. Conclusion

Missing values (MV) are often guessed/imputed for convenience of dealing with complete data when applying classification or other data mining methods, which is a topic we previously extensively discussed [3]. Here we modified a dynamic clustering approach previously adopted by us [4]. The MV imputation algorithm is an important addition to individualised healthcare in the current population screening environment since it allows the prediction of CVD in patients who do not provide all the required cardiac health predictor variables.

## References


[1] Little RJA, Rubin DB. Statistical analysis with missing data. 2002. 2nd ed. Wiley.

[2] Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338:b2393. doi:10.1136/bmj.b2393.

[3] Jelinek HF, Yatsko A, Stranieri A, Venkatraman S, Bagirov A. Diagnostic with incomplete nominal/discrete data. Artificial Intelligence Research 2015; 4(1):22-35.

[4] Stranieri A, Yatsko A, Jelinek HF, Venkatraman S. Data-analytically derived flexible HbA1c thresholds for type 2 diabetes mellitus diagnostic. Artificial Intelligence Research 2016; 5(1):111-134.

[5] Wilson PWF. Estimation of cardiovascular risk in an individual patient without known cardiovascular disease. www.uptodate.com, March 2016.

[6] Shao Z, Er MJ, Wang N. An Efficient Leave-One-Out Cross-Validation-Based Extreme Learning Machine (ELOO-ELM) With Minimal User Intervention. IEEE Transactions on Cybernetics 2015; (in print).

[7] D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008; 117:743.

[8] Amin SU, Agarwal K, Beg R. Genetic neural network based data mining in prediction of heart disease using risk factor. IEEE Conference on Information and Communication Technologies (ICT); 2013; 1227–1231.

[9] Selvarajah S, Kaur G, Haniff J, Cheong KC, Hiong TG, van-der-Graaf Y, Bots ML. Comparison of the Framingham Risk Score, SCORE and WHO/ISH cardiovascular risk prediction models in an Asian population. International Journal of Cardiology 2014; 176:211–218.

[10] Nwose EU, Richards RS, Jelinek HF, Kerr PG. D-dimer levels reflect progression of diabetes mellitus and likelihood of cardiovascular complications. Pathology 2007; 39(2):252-257.

[11] Cornforth D, Jelinek HF, Tarvainen M. A comparison of nonlinear measures for the detection of cardiac autonomic neuropathy from heart rate variability. Entropy 2015; 17(3):1425-1440.



Address for correspondence:
Herbert F. Jelinek
Charles Sturt University, NSW 2640 Australia
E-mail: hjelinek@csu.edu.au