Machine Learning Data Imputation and Classification in a Multicohort Hypertension Clinical Study

Libertas Academica
FREEDOM TO RESEARCH

William Seffens[1], Chad Evans[1], Minority Health-GRID Network and Herman Taylor[2]

[1]Physiology Department, Morehouse School of Medicine, Atlanta, GA, USA. [2]Director of Cardiovascular Research Institute (CVRI), Morehouse School of Medicine, Atlanta, GA, USA.

Supplementary Issue: Current Developments in Machine Learning Techniques in Biological Data Mining

ABSTRACT: Health-care initiatives are pushing the development and utilization of clinical data for medical discovery and translational research studies. Machine learning tools implemented for Big Data have been applied to detect patterns in complex diseases. This study focuses on hypertension and examines phenotype data across a major clinical study called Minority Health Genomics and Translational Research Repository Database composed of self-reported African American (AA) participants combined with related cohorts. Prior genome-wide association studies for hypertension in AAs presumed that an increase of disease burden in susceptible populations is due to rare variants. But genomic analysis of hypertension, even those designed to focus on rare variants, has yielded marginal genome-wide results over many studies. Machine learning and other nonparametric statistical methods have recently been shown to uncover relationships in complex phenotypes, genotypes, and clinical data. We trained neural networks with phenotype data for missing-data imputation to increase the usable size of a clinical data set. Validity was established by showing performance effects using the expanded data set for the association of phenotype variables with case/control status of patients. Data mining classification tools were used to generate association rules.

KEYWORDS: artificial neural network, data imputation, machine learning, hypertension

SUPPLEMENT: Current Developments in Machine Learning Techniques in Biological Data Mining.

CITATION: Seffens et al. Machine Learning Data Imputation and Classification in a Multicohort Hypertension Clinical Study. *Bioinformatics and Biology Insights* 2015:9(S3) 43–54 doi: 10.4137/BBI.S29473.

TYPE: Original Research

RECEIVED: June 05, 2015. **RESUBMITTED:** September 21, 2015. **ACCEPTED FOR PUBLICATION:** September 23, 2015.

ACADEMIC EDITOR: J. T. Efird, Associate Editor

PEER REVIEW: Nine peer reviewers contributed to the peer review report. Reviewers' reports totaled 3133 words, excluding any confidential comments to the academic editor.

FUNDING: Supported by 8U54MD007588, G12MD007602, P50HL117929, P30 HL107238, and 1RC4MD005964 (MH-GRID) grants from NIH/National Institute on Minority Health and Health Disparities. The content is solely the responsibility of the authors and does not necessarily represent official views of the respective institutions. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: wseffens@msm.edu

Introduction

A major conclusion from the Genome-Wide Association Study (GWAS) by Adeyemo et al.[1] for hypertension in African Americans (AAs) was that "alternate strategies … to identify rare variants, are clearly needed." As a result, genomic analysis of the Minority Health Genomics and Translational Research Repository Database (MH-GRID) study, which was designed to focus on rare gene variants, included a large array of phenotype variables and exome variants. MH-GRID is a multicohort hypertension clinical study of self-report AAs living in the southeastern US, a geographic area of high incidence of hypertension and stroke called the *Stroke Belt*.[2] The cohort partners in the MH-GRID study include Genetics of Left Ventricular Hypertrophy (Hyper-GEN), Reasons for Geographic and Racial Differences in Stroke, Genetics of Hypertension Associate Treatment, Jackson Heart Study, and Howard University Family Study. The data set contains >1,600 participants with extensive blood and urine analyses, self-report questionnaires on socioeconomic status, family history, medical records, and genomic exome data. Data mining methods, such as neural networks and matrix factorization, have recently been shown to uncover relationships in complex phenotypes, genotypes, and clinical data.[3,4] We used an artificial neural network (ANN) to impute MH-GRID missing phenotypic data and then converted the expanded data set into a data mining software suite to detect associations between phenotypes and case/control hypertension status.

The implementation of an ANN was selected not only for data imputation but also for detecting patterns in hypertensive patient data as *endophenotypes*.[4] Complex diseases are caused by multiple genetic, environmental, and behavioral factors. If a disease has heterogeneous etiologies, then the detection of operable genes is difficult as one set of genes can be important for one etiology but not for another. This could be an underlying difficulty with hypertension in prior GWAS efforts.[5] Endophenotype is an intermediate phenotype that combines genetic factors associated with a disease to reduce genetic heterogeneity. The MH-GRID data set is structured like an endophenotype since there is a control set and two cases, severe controlled and severe resistant hypertension (SRH).[6]

To get maximal statistical power, all cohort data are desired, as missing values degrade association studies.[7] The problem of databases containing missing values is common in biomedical informatics.[8] This issue arises from various reasons; it may be that the medical procedures were not needed clinically, that the procedure was not available in a cohort study, or that the measurements were taken but not recorded perhaps due to time constraints in the medical records or flagged in an upstream data cleaning operation. Missing data are a part of almost all clinical research, and investigators have to decide how to deal with them. For complete case analysis, only rows with all the values are used, this reduces the statistical power by lowering *N*. For available case analysis, substitution of missing values by one or more imputation methods increases *N*. Typical imputation methods include mean value method, replacing the missing value with mean value for that particular attribute; regression substitution method, replacing the missing value with historical value from similar cases; and matching imputation method, and for each unit with a missing *y*, finding a unit with similar values of *x* in observed data and taking its *y* value. Other methods include maximum likelihood and expectation-maximization (EM)., where some data mining models can deal with missing data better than others.,[9] Any decision regarding which technique to adopt really depends on the data set. Researchers must determine data-appropriate ways to incorporate incomplete data into their data set or lose statistical power. Performance of ANNs is degraded when coded missing values are used,[10] and when databases are highly skewed; ANNs have difficulty in identifying factors, leading to a rare outcome. The advantage that ANNs offer over statistical techniques is that the model does not have to be explicitly defined before the experiment begins. ANNs can capture the relevant data to develop robust models, whereas to derive a statistical model, prior knowledge of the relationships between the factors under investigation is required.[11]

We show here the use of ANNs for pattern detection in clinical phenotype data for use with hypertension data across more than six different data sources. Utilization of ANNs in clinical bioinformatics has increased recently with successful Big Data projects.[12] For instance, an NN model was developed to predict the risk of in-hospital mortality using various physiological measurements from the intensive care unit.[13] Another study found that ANN performed better than binary logistic regression models in the detection of diabetes status from clinical data.[14] A study related to the prediction of heart disease using ANNs with a clinical data set of 13 variables is similar to this study, and they found higher accuracy, sensitivity, and specificity than a state vector machine comparison.[15] These and other studies show the usefulness of ANNs for pattern detection in real-world clinical data sets.

Methods

Data source is from a multicohort consortium called the MH-GRID Network that is a catalog of AA genomic data funded by the National Institute of Health. The purpose of MH-GRID is to collect and analyze biospecimen samples to define genetic, personal, and social–environmental determinants of severe hypertension, specific to people of African ancestry. Eligibility criteria included AA ethnicity and age between 30 and 55 years at baseline. Exclusion criteria include patients with secondary forms of hypertension, primary forms of kidney disease, or major comorbidities such as diabetes, heart failure, end-stage renal failure, HIV, and liver disease. The estimated total enrollment of the MH-GRID study is 1,692 participants. Clinicians gathered demographic and anthropometric data as well as biospecimen samples from each participant. Age, sex, marital status, cigarette smoking status, and race and/or ethnicity were self-reported by the participant through patient health history survey using REDcap.[16] The project phenotype file of *critical* or clinically important hypertension-related variables contains mixed data types composed of continuous, binary, and categorical values. The phenotype file of critical variables contains participant age (mean 46.2 years old), gender (65% female), systolic blood pressure (SBP), diastolic blood pressure (DBP), glomerular filtration rate (GFR), albumin creatinine ratio (ACR), heart rate (HR), total cholesterol (Tot_Chol), high-density lipoproteins (HDLs), low-density lipoproteins (LDLs), triglycerides (Tri_Glyc), weight, height, body mass index (BMI), fasting blood glucose (glucose), blood urea nitrogen (BUN), marital status, and cigarette smoking status (CigSmoke). Summary statistics on these variables is presented in Table 1. Lipid variables, fasting plasma glucose, and estimated GFR were measured by standard laboratory techniques.[17] BMI was calculated as weight in kilograms divided by height in meters squared (kg/m^2). Each participant was scored by an MH-GRID physician as either control, severe controlled hypertension (SCH) case, or SRH case. SRH was defined as participants with blood pressures that remain above 140/90 mmHg while using three antihypertensive agents of different classes.[6] Those who had blood pressure levels below 140/90 mmHg were categorized as SCH if diagnosed as hypertensive and currently taking antihypertensive medication. Controls had normal blood pressures. The original MH-GRID human subject study was approved by MSM IRB 299568-4, and complied with the principles of the Declaration of Helsinki. Subjects gave their written, informed consent to participate in the research. The study presented here represents further analysis of data from that study, and was approved by MSM IRB 299568-14.

ANN implemented in this study was emergent[18] with backpropagation using default learning parameters. This package was selected for multiprocessing on a high-performance cluster, alleviating the need for extensive tuning of learning rates, momentum values, kernel, and activation functions to speed up overall processing. Hidden layer was ten neurons feed from one input neuron for each phenotype variable. Prediction error was measured as a sum of square of

Table 1. MH-GRID critical clinical variables for participants across all cohorts.

VARIABLE	DATA TYPE	MEAN	STD DEV	RANGE	UNITS	DESCRIPTION
Critical variable summary statistics						
Age	Continuous	46.2	6.6	26	years	At study visit
SBP	Continuous	115.5	13.6	118	mmHg	Systolic blood pressure
DBP	Continuous	74.8	9.4	78	mmHg	Diastolic blood pressure
GFR	Continuous	102.6	20	169.6	ml/min	Glomerular filtration rate
HR	Continuous	68.5	10.5	74	bpm	Heart rate
Tot_Chol	Continuous	190.8	38.5	268	mg/dL	Total cholesterol
HDL	Continuous	54.5	16	121	mg/dL	High- density lipoproteins
LDL	Continuous	117	35.7	254	mg/dL	Low- density lipoproteins
Tri_Glyc	Continuous	95.3	52.5	787	mg/dL	Triglycerides
Weight	Continuous	89.7	23.2	189.5	kg	Measured
Height	Continuous	169.1	9.3	55.4	cm	Measured
BMI	Continuous	31.4	7.9	56	unitless	Body mass index
Glucose	Continuous	90.8	16.4	305.5	mg/dL	Fasting blood glucose
Gender	Binary	0.65	NA	1	unitless	Male = 0, female = 1
Marital stat	Category	NA	NA	1	unitless	Marital status
CigSmoke	Category	0.31	0.46	1	unitless	Cigarette smoking status
Case/control	Binary			1	unitless	Control, SCH, and SRH

Note: Std Dev is standard deviation of mean for $n - 1$.

the difference between the output neuron activation and the expected value in the training set, described as a sum square error (SSE). Predictions were scored as correct if the output activation value was within 0.01 for single output neurons and 0.1 for three output neurons combined.

The scale of the data values was normalized to prevent artificial significance of high magnitude data types. For instance, the mean of each attribute of the transformed set of data points can be reduced to zero by subtracting the mean of each attribute from the values of the attributes and dividing the result by the standard deviation of the attribute. As seen in Table 1, large numeric values can be found with Tri_Glyc and LDL when compared to the categorical or binary variables. In this study, two data normalization methods were examined, *range* and *softmax*. *Range* normalization, sometimes called *min–max* normalization, scales the data between 0 and 1, by subtracting the minimum value of an attribute from each value of the attribute and then dividing the difference by the range of that attribute: $x' = [x - \min(X)]/[\max(X) - \min(X)]$, where X is the set of all values for that data variable. It has the advantage of preserving exactly all relationships in the data, without adding any bias. If the range of data values for a variable is large, then the data become compressed. *Softmax* is a way of reducing the influence of extreme values or outliers in the data without removing them from the data set. It is useful with outlier data that one desires to include in the data set while still preserving the significance of data within a standard deviation of the mean. As a result, we also examined *softmax*, which also scales the data between 0 and 1 by using a variance transform: $x\sim = [x - \text{mean}(X)]/\text{var}(X)$, where $\text{var}(X)$ is the variance of the data values for that variable. Then, the *softmax* value is calculated as: $x' = 1/[1 + \exp(-x)]$. Using the logistic sigmoidal *softmax* normalization[19] for input data, the data range for all variables was 0–1, with a mean of 0.5 for continuous data types. Gender was coded as a nominal binary variable with zero for female and one for male. Cigarette smoking status was coded 0.0 or 0.5 for nonsmoker and 1.0 for self-reported smoker. Marital status was coded nominal between 0.0 and 1.0 in evenly spaced values for six different categories or binary using indicator variables.

Imputation method was used as described by Jerez et al.[20] We removed 10 each of control, SCH, and SRH cases for test sets ($N = 30$), leaving a training set of $N = 1{,}266$ from the phenotype file of 1,296 complete entries. Imputation set depends on missing variables (from 1 to 334) and is not present in the training or test sets. We start with missingness = 1 and set up a neural network with one output neuron corresponding to the missing variable. We then started training on the training set while monitoring performance error with the test set every 100 epochs. When SSE leveled, we stopped training. This was repeated eight times, and the best-performing ANN was given the imputation set for predicted values. We then added the imputed set into the training set, repeated the training with greater missingness imputation sets, and added more output neurons (equal to missingness). Participants with imputed values were not used in subsequent test sets. Training errors increased with greater missingness, so we stopped the imputation process at missingness = 4.

During the ANN training process, the use of *k*-fold cross-validation makes the results of the testing process more reliable because it guarantees that all data are used for training and testing.[21] In *k*-fold cross-validation, the data are randomly divided into parts called folds, where each fold is equal to another. Among the folds, one fold is selected for testing and the other folds are used for training. This process is repeated *k* times. Finally, all testing results are averaged to produce a single estimation result. In this study, eightfold cross-validation is used ($k = 8$). Scoring prediction accuracy from the eight runs of test sets used thresholds for assessing the squared errors in the output. Participants in the test set with a square of the difference between the prediction and true value for one output neuron of <0.001 were scored *correct*, while errors <0.01 were scored *close*. Due to the data normalization used (*range* or *softmax*), these thresholds correspond to significance levels when the output values are close to 0.5, the mean of the data variable.

Data mining software used was the Weka suite[22] for classification and rule generation. Setting for models used M5 pruned model rules (using smoothed linear models), Weka classifier Ridor -F 3 -S 1 -N 2.0, classifier OneR -B 6, and classifier PART -M 2 -C 0.25 -Q 1. MODLEM -RT 1 -CM 1 -CS 8 –AS produces a minimal set of rules from numeric data without discretization.[23]

Results

Our goal was to link phenotypic data from the MH-GRID hypertension study to hypertension disease progression by implementing the concept of endophenotype using the largest possible data set for pattern analysis. Since MH-GRID is a multicohort study, not all data variables are consistent across the entire data set. Raw MH-GRID phenotypic data contain a significant amount of missing data, which precludes the use of all variables for all participants. The MH-GRID phenotype file of critical variables includes 21 data types for 1,692 participants. We apply ANN approaches to remedy the missing data issue and then use data mining techniques to identify the patterns of hypertension status understandable by biomedical researchers.

The MH-GRID phenotype file of critical variables contains, in addition to the variables listed in Table 1, ACR and BUN. If we exclude ACR and BUN variables due to systematic absence from several of the cohorts, the remaining data contain 1,296 complete entries for 17 relevant phenotype variables. Of the others with missing data, 396 entries have 1–9 missing variables: 23 entries are missing 1 variable (heart rate × 8, marital status × 6, LDL × 4, glucose × 2, CigSmoke × 2, and height × 1), 7 are missing 2 (HDL and LDL × 2, weight and BMI × 2, marital status and CigSmoke × 2, and weight and height × 1), 24 are missing 4 (combinations of HDL, LDL, Tot_Chol, Tri_Glyc, and glucose), while 334 are missing 5 or 7 variables (lipids and others above), and 8 missing 9 variables. Our goal was to impute up to four missing variables, allowing us to include 54 more MH-GRID entries for study.

Complete case analysis of case/control status. We trained a neural network on control/SCH/SRH status with ten hidden and three output neurons on the 1,296 complete entries by splitting the set into a 1,266 training set and eight ($k = 8$ on SRH) different 30 sample test sets. The test sets were composed of 10 participants each of control, SCH, and SRH cases randomly selected without replacement. The limiting indicator variable was SRH with 88 participants, while control and SCH had 711 and 497 participants, respectively.

We compared *range* and *softmax* data normalization methods with the MH-GRID data using the complete case analysis of control/SCH/SRH. S*oftmax* in this study was observed to have slight increases in ANN prediction accuracy. Discrimination between control and SCH is harder since both have normal BP and the phenotype data did not contain medications, necessary for case/control discrimination. *Softmax* performed better between control and SCH, but surprisingly poorer for SCH vs SRH in this preliminary test. We then used *softmax* normalized data for training networks to impute missing variables for the incomplete entries owing to the wide ranges compared to standard deviations of many of the variables in Table 1.

Control, SCH, and SRH were study classification-derived variables that were clinically judged by study criteria and certified by MH-GRID physicians. Hence, these variables were not missing but were the phenotype classes used for hypertension association studies. The complete data set comprised of control ($N = 716$), SCH ($N = 504$), and SRH ($N = 90$). *k*-Fold sampling used a test set composed of 10 randomly selected participants from each class, using binary coding that required three output neurons. Training to 14K epochs yielded minimal overtraining effects. Since the case/SCH/SRH data are binary, and the ANN output is continuous, rounding (winner takes all) is used to score the accuracy of prediction. We scored correct if the SSE was <0.5 combined for the three outputs. This yielded an accuracy of 89.2% for control, 74.6% for SCH, and 78.3% for SRH. The difficulty in discriminating SCH from control stems from the absence of medications in the phenotype data. We will determine if increasing the data set size by imputing phenotype values will enhance the performance of case/control discrimination from the above baseline.

Missing 1 variable. Phenotype data imputation on variables with missingness of 1 was performed using *k*-fold cross-validation shown in Table 2. Initial runs were monitored for overtraining and epoch limits, labeled *final epoch*, set for maximal accuracy in the validation test sets. Initial average SSE for continuous variables in untrained ANNs was 0.023–0.043 (LDL, height, glucose, and HR), while nominal variables (CigSmoke and marital status) had higher starting SSE since the mean distributions are not normal. A correct prediction in the validation set was scored if the SSE of the output was below a threshold value corresponding to a significance level. With *softmax* normalization, two significance levels were used to evaluate accuracy prediction (more stringent 0.05%

Table 2. Phenotype data imputation for missingness = 1.

MISSING VARIABLE	INITIAL AVG.SSE	FINAL AVG. SSE	FINAL EPOCH	ACCRCY @ 10%	STD DEV	ACCRCY @ 5%	STD DEV	ACCRCY @ 1%
LDL	0.039	0.0014	80	100	0	100	0	96.7
Height	0.043	0.0003	5800	100	0	100	0	100
CigSmoke	0.056	0.022	17K	75	10.5	61.3	6.2	39.6
Glucose	0.023	0.015	20K	89.5	4.5	78.1	7.7	43.3
Heart rate	0.043	0.018	33K	81.7	8.7	66.7	11	34
Martial status	0.059	0.046	150K	73.3	14.1	62.8	12	23.3

Notes: Initial is at epoch = 0. Test set accuracy (Accrcy) at three different significance levels, within 10%, 5%, and 1% of *softmax* true value. Std Dev is $n - 1$ standard deviation over the k-fold validation test sets ($k = 8$).

and relaxed at 0.10%). Accuracy was measured as the average correct prediction percentage of the $k = 8$ test sets of 30 or 160 complete entries. The case/control run was for baseline measurement and comparison and not for data imputation. These runs had much higher starting SSE from three outputs that were also nominal data types.

Missing LDL. Mean LDL in the training set was 117 mg/dL with a range of 17–271 mg/dL (Table 1). Training the ANN for missing LDL had very quick convergence, requiring only 80 epochs for 100% accuracy in the test sets at 0.05 significance level. This effect is due to the inclusion of total cholesterol (Tot_Chol), HDL, and triglycerides (Tri_Glyc) in the training set with LDL being a derived variable. There exists a known relationship between the following four variables: LDL = Tot_Chol – HDL – Tri_Glyc/5 from Friedewald et al.[24], and hence the ANN training converges on a simple pattern discovered with those variables. The Friedewald relationship is considered valid only for Tri_Glyc <400, and it is based on a population of European descent, not AA. There are four missing values of LDL in the MH-GRID data set, presumably due to Tri_Glyc >500 for those four participants. An Iranian population formula[25] has been recently described: LDL = Tot_Chol/1.19 + Tri_Glyc/1.9–HDL/1.1, but it is based on a smaller number of non-AA samples. Recently, the Chen modified formula has been described: LDL = 0.9 * (Tot_Chol–HDL)–0.1 * Tri_Glyc, which was validated in 2,180 cases from a Chinese population.[26] They found that LDL calculated using both Friedewald's formula and Chen modified formula correlated well with directly measured LDL when Tri_Glyc was <400 mg/dL, but when Tri_Glyc was >400 mg/dL, the modified formula correlated better with directly measured LDL. Different modified formulas have been individually validated in different populations, and each formula was suitable for a particular population,[27] but none have been validated for AA, especially with high Tri_Glyc. A recent work suggests that the Chen formula is of utility in AA for Tri_Glyc >400.[28] Table 3 lists the results of estimates based on the three formulas and the resultant ANN predictions for the MH-GRID participants with missing LDL values. Calculated values vary greatly between estimator equations and trend inconsistently (Table 3). ANN prediction of LDL is consistently higher than the Chen formula by 7%–35% for these extreme Tri_Glyc >300 participants. There is also consistency with the Friedewald's formula, but further away than the Chen formula. The ANN does not base predictions on statistically expected values, because the mean LDL values in the training and test sets are 0.494 and 0.559 *softmax* normalized, respectively, while the mean of the four imputed values is 0.672, which is significantly higher.

Discovering the relationships acquired by training ANNs is a difficult endeavor, and no method is generally applicable.[29] There are two types of approaches to extract rules from multilayer ANNs: local and global. In the global methods,[30] the neural network is treated as a black box, in which sets of global rules characterize the output classes directly in terms of inputs. Decompositional or local approaches go into the details of neural network structure, describing each neuron separately in terms of rules, followed by a concatenation algorithm.[31,32] For a first look, we used data mining techniques to suggest the structure of the ANN knowledge training.

Here, we reformatted the ANN training file for input into the Weka data mining suite. Using M5 pruned model rules (using smoothed linear models) produced the following two rules with a correlation coefficient of 0.99:

Table 3. LDL data imputation for 4 MH-GRID missing values with high triglycerides (Tri_Glyc).

TRI_GLYC	LDL CALCULATORS			ANN PREDICTION
	FE	I	CF	
543	86	408	121	143
525	86	396	119	148
502	140	426	166	179
800	9	524	72	110

Notes: Friedewald equation (FE): LDL = Tot_Chol – HDL – Tri_Glyc/5, Iranian (I): LDL = Tot_Chol/1.19 + Tri_Glyc/1.9–HDL/1.1, and Chen Formula (CF): LDL = 0.9(Tot_Chol–HDL)–0.1Tri_Glyc. MH-GRID participant's Tri_Glyc with missing LDL value. All units in mg/dL.

Rule 1: IF Tot_Chol < = 0.49 THEN LDL = 1.11 * Tot_Chol–0.43 * HDL–0.31 * Tri_Glyc

Rule 2: LDL = 1.08 * Tot_Chol–0.46 * HDL–0.32 * Tri_Glyc

The variables in the data mining rules are *softmax* normalized, so direct comparison to the formulas in Table 3 is not direct but is approximate for values near the 0.5 mean. These missing LDL data are classed as Missing Not at Random (MNAR) since the values were not calculated for the phenotype file if Tri_Glyc >400 in the MH-GRID data recording operation. Therefore, the imputation is biased, but the bias may be small if the model is well behaved for large values of triglycerides.

Missing height. Mean height in the training set was 168.8 cm with a range of 142–200 cm (Table 1). Training the ANN for height also had very quick convergence, requiring only 40 epochs for a typical plateau-shaped learning curve (Fig. 1A). This rapidity is due to inclusion of weight and BMI in the training set variables. Since BMI is also a derived variable, there exists a defined relationship between the three variables: Height = $[Weight/BMI]^{1/2}$ and hence the ANN training converges on a simple mathematical relationship discovered within those variables. But upon further learning, a secondary plateau forms at 2,000 epochs (Fig. 1B), again for 100% accuracy in the test sets at >1% significance. There was only one missing value for height in the data set. The BMI formula estimates that height as 190.2 cm, while the 2,000 epoch-trained ANN predicts 192 cm, or just 0.9% too high. An analogous study used ANNs to predict body weights of rabbits from body measurements[33] and concluded that the ANN model is better than multivariate linear regression.

The later convergence in Figure 1B may be due to discovery of patterns with outlier or spurious values within the dependent variables. To examine this, we loaded the ANN training file into the Weka data mining suite. Using M5 pruned model rules produced 16 rules with a correlation coefficient of 0.98, the most significant rule, in *softmax* values, being:

Rule 1: Height = –1.55 * BMI + 1.54 * Weight–0.13 * Gender

Again, the variables in the data mining rules are *softmax* normalized, so direct comparison to the correct formula is not direct but is approximate for values near the 0.5 mean. This suggests that ANN training discovers the known gender differences in BMI, along with other factors present in the phenotype data giving rise to the two-phase learning curves in Figure 1.

In a recent study,[34] BMI showed significant correlations with SBP, DBP, and HR after controlling for age and physical activity status in both genders, and SBP indicated the strongest association with BMI. Again, the ANN does not base predictions on statistically expected values, because the mean height values in the training and test sets are 0.501 and 0.531 in *softmax*, respectively, while the mean of the imputed value is 0.917 (corresponding to 192 cm), which is significantly higher than the training set mean. This missing value for height was probably due to a random data processing omission, and therefore the data class is Missing Completely at Random. This implies that the ANN prediction is unbiased since the probability that the observation (height) is missing is unrelated to the value of height or to the value of any other phenotype variables in the training set.

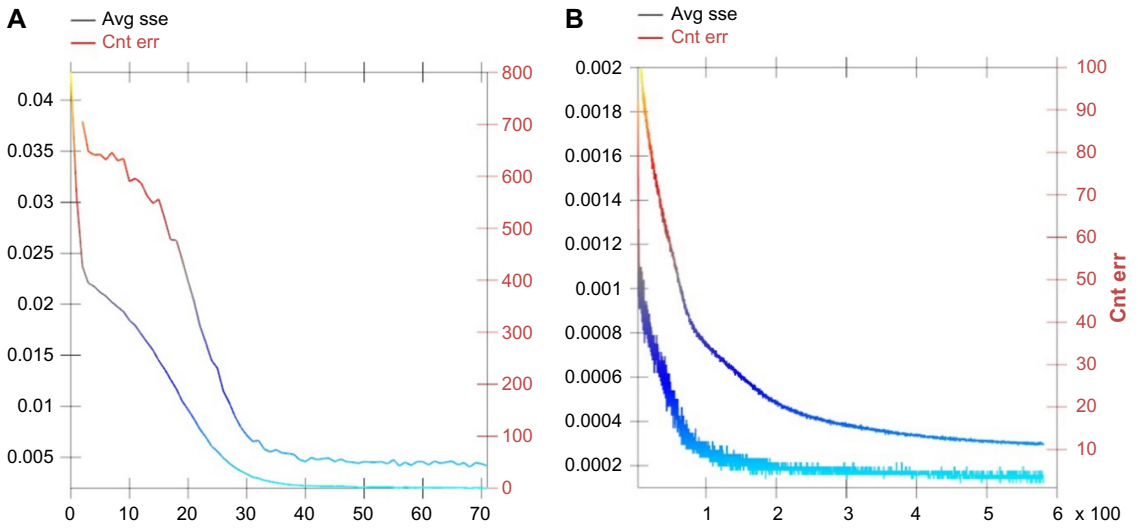

Figure 1. ANN learning curve on height.

Notes: Upper curves are the number of training set instances that are predicted incorrectly above a program set point, while the lower curves show learning by a decrease in SSE of the output neuron layer.

Missing cigarette smoking. Cigarette smoking has well-known impacts on cardiovascular function and health, but conversely, the variables in the phenotype file may not be predictive.[35] This variable among the MH-GRID cohorts was inconsistently defined and scored, some with a simple binary yes/no, while others with gradations and time frame qualifiers, and some that differentiated the tobacco product. Cigarette smoking status was initially coded 0.5 for nonsmoker and 1.0 for self-reported smoker, with 0.0 reserved for missing. We chose a nonbinary coding scheme to allow for more categories that were present in some of the cohort phenotype data and later compared to binary indicator variable coding. For the "0.5/1.0" coding, the learning curve plateaus ~8,000 epochs, and good test accuracy occurs at 17,000 epochs (Fig. 2). During learning, the fraction of correct predictions in the training set goes from 5% at 0 epochs to 65% at 17,000 epochs. The stringency is set at 1% on the SSE during learning in the ANN. Again, the ANN does not base predictions on statistically expected values, because the mean values in the training and test sets are 0.657 and 0.600 *softmax* normalized, respectively (medians are both 0.500), while the mean of the two imputed values is 0.833, which is significantly higher. The two imputed values correspond to *smokers* by rounding. So the missing data class is most likely Missing At Random (MAR) and at worst MNAR if smokers tend not to self-report on questionnaires. Some clinical studies suggest high veracity and accuracy on self-reports,[36] while other studies in minority populations have reported substantial refusal of confirmation and disconfirmation rates in both intervention and control groups[37,38]

Data mining with Weka produced sets of many rules, which did not have ready interpretations to physiological models. The previous imputed variables discussed were continuous, while cigarette smoking is binary. In this case, rounding for the final ANN prediction would be appropriate, making the *relaxed* accuracy measure in Table 2 appropriate. We examined whether the values of the data coding influenced learning or pattern detection. Setting nonsmoker status from 0.5 to 0.0 increases the distance between training and test set values and the coding becomes properly binary. This alternate coding scheme did not significantly change the performance of the learning. The test accuracy values for 1% and 0.1% significance levels were 10.7% (±6.6) and 39.6% (±7.0), correct for 0.5/1.0 coding, and 37.6% (±7.9) and 40.5% (±8.7) for "0.0/1.0" coding, respectively. Therefore, at the most stringent level, test accuracy was nearly identical, while at 1% level, the greater numeric distance between the coded values yields better accuracy (37.6% vs 10.7%).

Missing glucose. Mean glucose in the training set was 90.9 mg/dL, with a range of 54–360 mg/dL (Table 1). Fasting glucose levels in blood have a well-known connection to health, but again, the variables in the phenotype file may not be predictive.[39] The learning curve plateaus ~10,000 epochs (Fig. 3), and good test set accuracy occurs at 20,000 epochs (Table 2). The long learning curve past 10K epochs provides only minor increases in accuracy. At 5K epochs and 10% significance level, the accuracy is 89.3%, while at 20K, it is 89.5%, which is not significantly different. As a function of testing significance level, the accuracy values at 20K epochs are 78.1% (±7.7 at 5% significance level), 43.3% (±8.2 at 1% significance level), and 15.7% at 0.1% significance level. As for smoking status, data mining with Weka produced sets of many rules, which did not have ready interpretations to physiological models.

The ANN does not base predictions on statistically expected values, because the mean glucose values in the training and test sets are 0.487 and 0.529 in *softmax*, respectively, while the imputed value is 0.393.

Missing HR. Mean HR in the training set was 68.5 bpm, with a range of 38–112 bpm (Table 1). There are two different factors involved in HR management: intrinsic and extrinsic controls. Intrinsic regulation of HR is the result of the unique nature of cardiac tissue – it is self-regulating and maintains a rhythm. Extrinsic controls are those that come from both

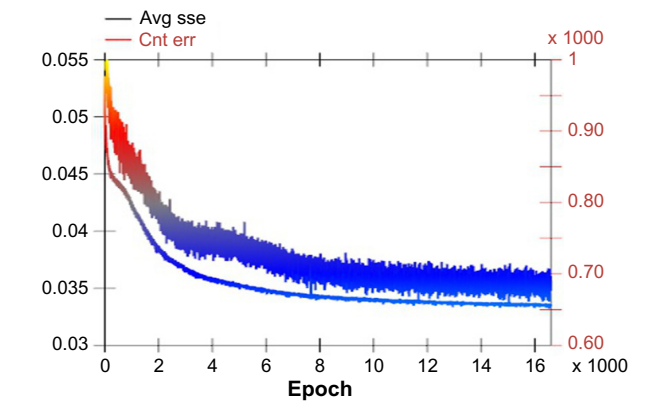

Figure 2. ANN learning curve on cigarette smoking.
Notes: Upper curve is the number of training set instances out of 1,266 that are incorrect above an error set point (0.01), while lower curve is SSE of the output neuron layer.

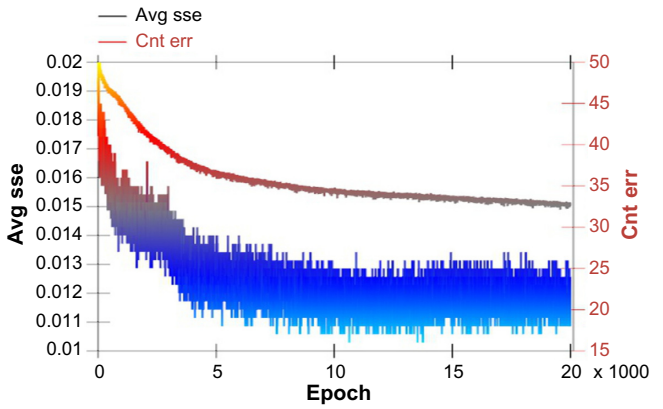

Figure 3. Fasting glucose ANN learning curve.
Notes: Upper curve is SSE of the output neuron layer, while lower curve is the number of training set instances that are incorrect above an error set point.

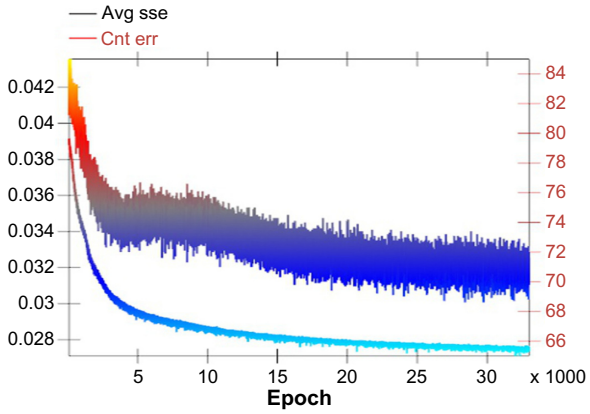

Figure 4. Learning curve for heart rate.
Notes: Upper curve is the number of training set instances that are incorrectly predicted, while lower curve is SSE of output.

hormonal responses and commands from the central nervous system and autonomic nervous system. Extrinsic regulation can cause the HR to change rapidly because of chemicals that circulate in the blood or by direct action of nerves that go to the heart. Due to the complex nature of the system, the relationship of HR to other phenotype file variables would be tenuous.

Typically, HR is 60–80 bpm, with influences from gender and age; men aged 36–45 years average 71–75 bpm and those aged 46–55 years average 72–76 bpm and women aged 36–45 years average 74–78 bpm and those aged 46–55 years average 74–77 bpm. Figure 4 after 33K epochs shows a gradual decline in SSE error and an accuracy of about 63%, correct in the test set at 5% stringency and close to 82% for 10% stringency (Table 2).

Again, the ANN does not base predictions on statistically expected values, because the mean HR values in the training and test sets are 0.494 and 0.435 *softmax* normalized, respectively, while the mean of the eight imputed values is 0.370, which is significantly lower and within a 0.03–0.78 range.

Missing marital status. Marital status variable is categorical, with six possibilities; so we examined different coding schemes. Initially, marital status was coded 0.0 for never married, 0.2 for single, 0.4 for married, 0.6 for separated, 0.8 for divorced, and 1.0 for widowed. The data set contains 19% never married, 27% single, 33% married, 3% separated, 15% divorced, and 3% widowed. Here, the ANN has a prediction close to statistically expected values, because the mean values in the training and test sets are 0.435 and 0.507, respectively, while the mean of the imputed value is 0.383, which would be rounded to 0.4 as a prediction. Married is coded 0.4, the nearest round value for the prediction, which is the most frequent class in the data set (33% married). Throughout the training, the learning curve had a gradual downward slope, and hence training was followed out to 150K epochs (Fig. 5). The accuracy was unexpectedly good at 73% (within 10% of true values) across the eight test runs, considering only indirect linkage to any of the other phenotype variables.

Sorting the assignment of categorical values to different marital status examines sensitivity and is done to allow information content comparison with the ordinal variables using the same SSE measure. Changing the order of assignment for categories and using odd values evenly spaced between 0 and 1 had minor effect on the test set performance. Compared to the accuracy values shown in Table 2, the rearranged order yielded 46.1% accuracy at 10% significance level, 32.2% at 5%, and 18.3% at 1% significance level, all with similar standard deviations. These differences with the prior assignment order are not significant. We then converted marital status into six indicator binary variables. Using indicator variables for each marital status category requires six output neurons instead of one for nominal continuous variables and alters SSE comparison with the ordinal variables. Binary variables allow a winner-take-all outcome for predictions, which lead to increases in prediction accuracy to 53.4% for indicator variable marital status.

Missing 2 variables. The abovementioned results can be expanded into examining greater missingness in the MH-GRID

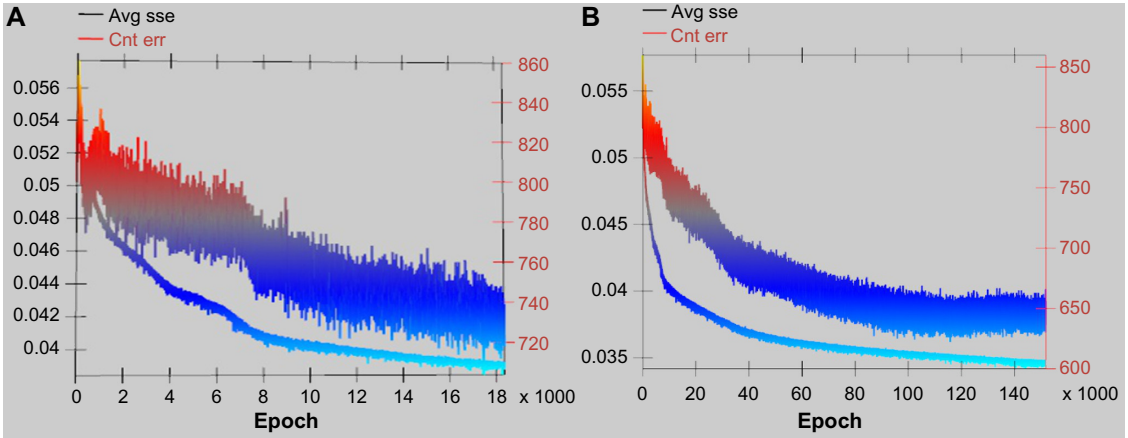

Figure 5. Learning curve for marital status. A is early phase out to 18K epochs, while B is extended out to 150K epochs.
Notes: Upper curves are the number of training set instances that are incorrectly predicted, while lower curves are SSE of output.

Table 4. Summary of imputation accuracy averages.

IMPUTED VARIABLE	MISSING 1			MISSING 2				MISSING 4				
	CORRECT	CLOSE	EPOCH	MISSING	CORRECT	CLOSE	EPOCH	MISSING	MISSING	CORRECT	CLOSE	EPOCH
Marital	8.6	23.3	152219	CigSmoke	6.7	26.7	32200	–	–	–	–	–
CigSmoke	37.6	40.5	16636	Marital	10	33.3	32200	–	–	–	–	–
Height	93.3	96.7	5804	Weight	20	50	30600	–	–	–	–	–
Weight	–	–	–	Height	46.7	80	30600	–	–	–	–	–
Weight	–	–	–	BMI	10	40	43500	–	–	–	–	–
BMI	–	–	–	Weight	6.7	40	43500	–	–	–	–	–
LDL	86.7	100.0	80	HDL	23.3	80	33100	Tri_Glyc	Tot_Chol	10	43.3	538
HDL	–	–	–	LDL	3.3	40	33100	Tri_Glyc	Glucose	13.3	36.7	538
HDL	–	–	–	LDL	3.3	40	33100	Tri_Glyc	Tot_Chol	6.7	40	507
HDL	–	–	–	Glucose	–	–	–	Tri_Glyc	Tot_Chol	6.7	20	501
Glucose	15.7	43.3	20135	HDL	–	–	–	Tri_Glyc	LDL	10	23.3	538
Glucose	15.7	43.3	20135	HDL	–	–	–	Tri_Glyc	Tot_Chol	6.7	50	501
HR	12.0	34.0	33074	–	–	–	–	–	–	–	–	–
Tot_Chol	–	–	–	HDL	–	–	–	Tri_Glyc	Glucose	40	83.3	501

Notes: Average accuracy values of eight test set runs for the variable in first column. Additional missing variables are listed for missing 2 and 4. Accuracy drops with greater missingness, and more learning epochs are needed.

phenotype data set. With imputation of one missing data variable cases, the resulting training set grew from 1,296 to 1,319, and imputation of missingness = 2 variable cases adds seven more participants. For missingness = 2 cases, a test set of equal proportion of control, SCH, and SRH was used to assess a two-neuron output model of learning HDL and LDL. Learning was sampled at 3K epochs to 33K epochs (Table 4). The ANN predicted 73% of the time too high for HDL and 27% too low for LDL, for an overall 10% error at 15K epochs. The minimum prediction error for HDL was at 15K epochs, while for LDL, it was at 33K epochs in the same training session. The lowest absolute average prediction errors were +10% for HDL and –2.5% for LDL, implying that LDL is better predicted among the two missing variables. This performance difference is perhaps caused because LDL is a derived variable and its expression is simpler to model, whereas HDL is an actual measurement. Half of the HDL predictions and 80% of LDL predictions are within 10% of actual values.

An additional set of missingness = 2 variables involves the relationship among height, weight, and BMI (Table 4). We examined weight and height in the same manner as HDL and LDL mentioned earlier. The best performance for predicting weight occurs at 30,600 epochs with an overpredicted average of +0.34%. Height minimal prediction error occurred at 8,800 epochs at +1.6%. Surprisingly, the only direct information content in the phenotype data is from BMI. Similar to the HDL and LDL case, the accuracy of prediction for weight and height is different, with weight being more than twice as accurate. This suggests that weight has greater connection to the other phenotype variables than height, especially given the BMI. Next, we examined weight and BMI. The best individual performance was at 43,500 epochs, with overall error continuing to drop at that termination point. From the first case (weight and height), height was easier to learn but had greater error than weight, and in the second (Weight and BMI), both are difficult to predict (>43K epochs). Surprisingly, the error in weight prediction increases with loss of BMI more than that in height (10% vs 46.7%). Predicting marital status and cigarette smoking status also yielded low SSE in the test set at 32,200 epochs. Comparing the three example cases, HDL and LDL had the lowest SSE due to the inclusion of related variables of Tri_Glyc and Tot_Chol in the phenotype file, while weight and height showed asymmetry in ANN impact. Future experiments will incorporate other anthropometric variables, such as waist circumference and stress markers,[40] to study this effect further.

The abovementioned data imputation efforts yielded seven new additional participants for downstream analysis. We then trained on control/SCH/SRH status and found 98% correct prediction of control status, 81% correct prediction of SRH status, but only 70% of SCH status. The latter performance is due to the lack of medications in the phenotype data file, necessary for clinical classification into the three endophenotype variables.

Missing 4 variables. The abovementioned data imputation effort yielded seven new additional participants for downstream analysis from missingness = 2 MH-GRID participants. These were added into the training set for imputation of the 24 MH-GRID participants who are missing four variables, combinations of HDL, LDL, Tot_Chol, Tri_Glyc, and glucose, allowing us to include 54 more entries for study. The 24 entries grouped into four classes of common missing variables, with HDL and Tri_Glyc occurring across all four classes as summarized in Table 4. Comparison among these variables allowed assessment

of the loss of ANN predictive accuracy with loss of more data variables. The variable LDL is the only instance that occurs in all three missingness cases (1, 2, and 4), allowing for a better understanding of its information contribution to the other variables. Missing Tri_Glyc and Tot_Chol on top of missing HDL drops the accuracy by half, but still results in 43.3% correct prediction within 1% of true value. The same drop in accuracy is seen in the missing HDL variable at 40% correct. Although the accuracies listed in Table 4 for missingness = 4 are low, the majority of test values were within 10% of true values, perhaps a more clinically appropriate significance level. Therefore, we again used the trained ANNs to impute missingness = 4 variables.

Our goal was to extend the imputation with missingness = 5 and 7 to achieve a 31% increase in training set size to 1,692, but the NNs performed poorly with imputing 7 out of 21 data variables. In these extreme imputation cases, the NN did not converge even out to 500K epochs and the test set accuracy decreased (data not shown) even for the reduced 10% significance level. Extended the training out to hundreds and thousands of epochs can be necessary for complex bioinformatics pattern detection,[41,42] or other machine learning techniques are required. For missingness up to four variables, only 0.6% of the data values are missing, facilitating rapid and accurate imputation. But for missingness = 5–7, 8.6% of data values become missing, a significant jump for this data set.

Available case analysis on case/control status. With 54 additional MH-GRID participant entries from the phenotype data imputation, we can assess the impact on downstream analysis by redoing the case/control ANN. As before with the complete case analysis, we trained a neural network on control/SCH/SRH status with ten hidden and three output neurons on 1,350 available entries by splitting the set into a 1,320 training set and eight ($k = 8$ on SRH as the limiting indicator variable) different 30 sample test sets *balanced* between control, SCH, and SRH. Again, the test sets were composed of 10 participants each of control, SCH, and SRH randomly selected without replacement. Training out to 4K epochs yielded minimal overtraining effects as judged by SSE in the test sets evaluated periodically. Predictive accuracy was higher at 85.8% with 2K epochs than the complete smaller set at 79.6%. A two-tailed paired t-test of test results from 500 to 4K epochs gave 0.020, demonstrating a significant performance increase using the imputed values. Selectivity between the cases for eight replicates yielded accuracy values of 91.3% for control, 77.5% for SCH, and 86.7% for SRH. The difficulty in discriminating SCH from control stems from the absence of medications in the phenotype data.

We changed the design of the test set to determine if the small observed ANN performance increase with the imputed data described above is sensitive to construction and size. The prior test set was balanced between control and the two cases, SCH and SRH, but could only be 30 instances in size due to the limited SRH cases. The test set was increased to 160 as 1/8 of the size of the overall data set, randomly selected without replacement, so the control/cases were not evenly balanced in numbers. Ten test sets were evaluated using t-test, but no significant difference was found between the training with or without the imputed values. Majority of the SRH test set instances numbered less than the balanced test set protocol described earlier.

Complete and imputed clinical variables from phenotype file were formatted for Weka input to predict the study status with machine learning classifiers. Java Repeated Incremental Pruning (JRIP) rules correctly classified 89% of the participants into the following seven rules:

(DBP > 0.84) = > class = SRH
(SBP > 0.85) = > class = SRH
(SBP > 0.58) = > class = SCH
(GFR < 0.35) = > class = SCH
(DBP > 0.67) = > class = SCH
(BMI > 0.55) and (Age > 0.55) and (Height > 0.57) and (SBP < 0.36) = > class = SCH= > class = Control

These rules show the classification of SRH by the elevated blood pressures (rules 1 and 2), while SCH separates from control by additional variables of GFR, BMI, age, and height.

Discussion

This study evaluated the use of ANNs for data imputation and pattern detection in clinical phenotype data, involving hypertension status across six different data sources. Comparison among the cohorts was not possible due to limited and unequal distribution of missing values in the data set. We can compare data class type and ANN error since the missing variables were continuous, binary, and categorical. The error for the continuous measured value of height is 0.043, and for the continuous derived variable LDL, it is lower at 0.0005. Thus, when HDL becomes missing with LDL, the error jumps to 1.03. While height has a 0.043 SSE, it jumps more with the loss of derived variable BMI (2.71) than with the loss of the measured value height (1.19). So in this case, the loss of a derived variable is more severe than the loss of a measurement. Moreover, if the lost variable has a content of a measured value, and two such variables become missing, then the SSE is even greater as shown by marital and cigarette status at more than twice the magnitude (3.42 SSE), although the latter effect could be due to involvement of a binary and categorical variable.

It is likely that the performance increase observed in control/SCH/SRH prediction is mainly due to an increase in the training set size when adding the available imputed values, rather than the incorporation of valuable patterns absent in the missing data. Thus, it would be unclear if just using mean values would make a difference in the case/control ANN outcome. Certainly, such an imputation method would skew the statistical distribution, which may alter outcomes. However, examining the accuracy as in Table 4 provides insight into the commonality or interaction of information content within the data set, which would not be uncovered by using statistical imputation

methods. Sensitivity to the loss of any variable indicates significant information content from that data element.

Conclusion

These results are presented as a use case of data imputation using ANNs to help merge the variables in a multicohort clinical study. We have shown that a neural network can predict one, two, and four missing variables from a data set containing 17 variables (for a 23% loss of variables and with only 0.6% total information missing). ANN predictive accuracy ranged from 100% to 23% for one missing value, and with four missing variables together, it was 87%, while it was 7% in some cases. We were not able to extend the range of missingness to 5–7 in order to capture a much larger data set for available analysis.

Acknowledgments

Minority Health-GRID Network Executive Team members: Rakale Collins Quarells, PhD, Morehouse School of Medicine, rquarells@msm.edu; Donna K. Arnett, PhD, MSPH, University of Alabama at Birmingham, arnett@uab.edu; Gary H. Gibbons, MD, National Heart, Lung, and Blood Institute, National Human Genome Research Institute, gary.gibbons@nih.gov; Robert L. Davis, MD, MPH, The University of Tennessee Health Science Center, rdavis88@uthsc.edu; Suzanne M. Leal, PhD, Baylor College of Medicine, suzannemleal@gmail.com; Deborah A. Nickerson, PhD, University of Washington, debnick@u.washington.edu; James Perkins, PhD, Clark Atlanta University, jperkins@cau.edu; Charles N. Rotimi, PhD, National Human Genome Research Institute, rotimic@mail.nih.gov, Joel H. Saltz, MD, PhD, SUNY Stony Brook, Joel.Saltz@stonybrook.edu; Herman A. Taylor, Jr., MD, MPH, FACC, Morehouse School of Medicine, htaylor@msm.edu; James G. Wilson, MD, University of Mississippi Medical Center, jgwilson2@umc.edu.

Author Contributions

Conceived and designed the experiments: WS. Analyzed the data: WS, CE. Wrote the first draft of the manuscript: WS. Contributed to the writing of the manuscript: CE. Agree with manuscript results and conclusions: WS, CE, HT. Jointly developed the structure and arguments for the paper: MH-GRID Network. Made critical revisions and approved final version: WS, HT. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Adeyemo A, Gerry N, Chen G, et al. A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet*. 2009;5:e1000564.
2. Howard VJ, Woolson RF, Egan BM, et al. Prevalence of hypertension by duration and age at exposure to the stroke belt. *J Am Soc Hypertens*. 2010;4(1):32–41.
3. Lynn KS, Li LL, Lin YJ, et al. A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data. *Bioinformatics*. 2009;25(8):981–8.
4. Wang HM, Hsiao CL, Hsieh AR, Lin YC, Fann CSJ. Constructing endophenotypes of complex diseases using non-negative matrix factorization and adjusted rand index. *PLoS One*. 2012;7(7):e40996.
5. Levy D, Ehret G, Rice K, van Duijn CM. Genome-wide association study of blood pressure and hypertension in six population-based cohort studies. *Nat Genet*. 2009;41:677–87.
6. Calhoun DA, Jones D, Textor S, et al. Resistant hypertension: diagnosis, evaluation, and treatment. A scientific statement from the American Heart Association Professional Education Committee of the Council for High Blood Pressure Research. *Hypertension*. 2008;51(6):1403–19.
7. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inform*. 2012;10(2):117–22.
8. Mackinnon A. The use and reporting of multiple imputation in medical research – a review. *J Intern Med*. 2010;268(6):586–93.
9. Kaiser J. Dealing with missing values in data. *J Syst Integr*. 2014;1:42–51.
10. Ennett CM, Frize M, Walker CR. Influence of missing values on artificial neural network performance. *Stud Health Technol Inform*. 2001;84(pt 1):449–53.
11. Livingstone DH, Manallack DT, Tetko IV. Data modelling with neural networks: advantages and limitations. *J Comput Aided Mol Des*. 1997;11:135–42.
12. Chen J, Xing Y, Xi G, et al. A comparison of four data mining models: Bayes, neural network, SVM and decision trees in identifying syndromes in coronary heart disease. Lecture notes in computer science. *Adv Neural Netw*. 2007;4491:1274–9.
13. Xia H, Daley B, Petrie A, Zhao X. A neural network model for mortality prediction in ICU. *Comput Cardiol*. 2012;39:261–4.
14. Rahman A, Nesha K, Akter M, Uddin S. Application of artificial neural network and binary logistic regression in detection of diabetes status. *Sci J Public Health*. 2013;1(1):39–43.
15. Chitra R, Seenivasagam V. Heart disease prediction system using supervised learning classifier. *Int J Software Eng Soft Comput*. 2013;3(1):2277–5099.
16. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377–81.
17. Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl*. 2013;3:1–150.
18. Aisa B, Mingus B, O'Reilly R. The emergent neural modeling system. *Neural Netw*. 2008;21:1146–52.
19. Pyle D. *Data Preparation for Data Mining*. Morgan Kaufmann Series in Data Management Systems Series, Morgan Kaufmann Publishers, San Francisco, CA; 1999.
20. Jerez JM, Molina I, Garcia-Laencina PJ, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif Intell Med*. 2010;50:105–15.
21. Baykan NA, Yilmaz N. A mineral classification system with multiple artificial neural network using k-fold cross validation. *Math Comput Appl*. 2011;16(1):22–30.
22. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten H. The WEKA data mining software: an update. *SIGKDD Explor*. 2009;11(1):10–8.
23. Stefanowski J. The rough set based rule induction technique for classification problems. In: Proceedings of 6th European Congress on Intelligent Techniques and Soft Computing. Vol. 1; 1998; Aachen:109–13.
24. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low–density lipoprotein cholesterol in plasma, without use of preparative ultracentrifuge. *Clin Chem*. 1972;18:499–502.
25. Ahmadi SA, Boroumand MA, Moghaddam KG, Tajik P, Dibaj SA. The impact of low serum triglycerides on low density lipoprotein cholesterol estimation. *Arch Iran Med*. 2008;11:318–21.
26. Chen Y, Zhang X, Pan B. A modified formula for calculating low density lipoprotein cholesterol values. *Lipids Health Dis*. 2010;9:52–7.
27. Nigam PK. Calculated low density lipoprotein-cholesterol: Friedewald's formula versus other modified formulas. *Int J Life Sci Med Res*. 2014;4(2):25–31.
28. Wang H, Becker DM, Vaidya D, Moy TF, Yanek LR, Kral BG. An improved equation for estimation of LDL cholesterol in European and African American adults. *Circulation*. 2011;124:A17196.
29. Yedjour D, Yedjour H, Benyettou A. Explaining results of artificial neural networks. *J Appl Sci*. 2011;11(15):2855–60.
30. Markowska-Kaczmar U. Evolutionary approaches to rule extraction from neural networks. *Studies Comput Intell*. 2008;82:117–209.
31. Kamruzzaman SM, Islam M. Extraction of symbolic rules from artificial neural networks. *World Acad Sci Eng Technol*. 2010;10:271–7.
32. Zhou ZH, Jiang Y, Chen SF. Extracting symbolic rules from trained neural network ensembles. *AI Commun*. 2003;16:3–15.
33. Salawu EO, Abdulraheem M, Shoyombo A, et al. Using artificial neural network to predict body weights of rabbits. *Open J Anim Sci*. 2014;4:182–6.
34. Dimkpa U, Oji JO. Relationship of body mass index with haemodynamic variables and abnormalities in young adults. *J Hum Hypertens*. 2010;24:230–6.
35. Papathanasiou G, Mamali A, Papafloratos S, Zerva E. Effects of smoking on cardiovascular function: the role of nicotine and carbon monoxide. *Health Sci J*. 2014;8(2):274–90.

36. Barruecoa M, Jiménez Ruizb C, Palomoc L, Torrecillad M, Romeroe P, Riescof JA. Veracity of smokers' response regarding abstinence at smoking cessation clinics. *Arch Bronconeumol*. 2005;41:135–40.
37. Schoenbach VJ, Tracy Orleans C, Quade D, et al. Effectiveness of a self-help quit smoking program for African Americans. 2000. Available at: www.epidemiolog.net/pub/qfl/QuitforLife.pdf
38. Lillington L, Royce J, Novak D, Ruvalcaba M, Chlebowski R. Evaluation of a smoking cessation program for pregnant minority women. *Cancer Pract*. 1995;3:157–63.
39. Bando Y, Ushiogi Y, Okafuji K, Toya D, Tanaka N, Fujisawa M. The relationship of fasting plasma glucose values and other variables to 2-h postload plasma glucose in Japanese subjects. *Diabetes Care*. 2001;24(7):1156–60.
40. Davis AR, Bekka-Coker W, Ahabue B, et al; MHGRID Network. Study of hypertension and neighborhood violence in the MHGRID dataset: a first look. In: NIMHD Conference; 2014; Baltimore, MD.
41. Pratt K, Seffens W. Evaluation of hidden layer architecture in neural networks for mRNA backtranslation. In: Proceedings of GSAC XV; 2003; Savannah, GA:83.
42. White G, Seffens W. Using a neural network to backtranslate amino acid sequences. *Electron J Biotechnol*. 1998;1:3.