Multiple Imputation of a Large Cardiac Care Clinical Audit Dataset

BY BRIAN A. CATTLE

Biostatistics Unit, University of Leeds, Leeds LS2 9JT
e-mail: *b.a.cattle@leeds.ac.uk*

Keywords: Multiple imputation

1. Introduction

Modern healthcare systems involve the collection of a significant amount of routine data for audit purposes. Audit data has a significant advantage over clinical trial data in that it can show how treatments and procedures work in everyday practice with the full spectrum of patients rather than in carefully controlled trials with strict patient eligibility criteria. Clinical audit data therefore has the potential to be a valuable resource in many aspects of front line health care provision.

The greatest concerns in using audit data for research is missing and implausible data, which are inevitable when large volumes of data are collected. Missing and implausible data have the potential to bias statistical analysis if they cannot be appropriately addressed. Biases in clinical applications are particularly problematic since treatment strategies or healthcare policy could be based on flawed findings. The potential for missing data to undermine the validity of results has often been overlooked in medical applications (REF WOOD & WHITE 2004), partly because statistical methods to tackle missing data problems have not been widely available to medical researchers.

A number of methods have been proposed to deal with missing data including using only complete cases; using a missing category indicator (REF VACH & BLETTNER) and replacing missing values with the last measured value (REF: CARPENTER). None of these approaches is statistically valid and they will often lead to very serious bias. Single imputation of missing values usually causes standard errors to be too small, since it fails to account for the fact that we are uncertain about the missing values. Specifically single mean imputation shrinks standard errors unacceptably. Single regression imputation, whilst marginally better than mean imputation, tends to grossly exaggerate correlations. Complete case analysis is also unsatisfactory because even if missingness is not severe in any single variable, the combined effect when a number of variables are used in an analysis can be very large.

For some time *multiple imputation* has been suggested as a promising approach for dealing with missing data (REF RUBIN & LITTLE), although it is not until recently that coherent guidelines for its use have been suggested in the medical literature (REF STERN, CARPENTER & WHITE). Multiple imputation allows for the uncertainty about the missing data by creating several plausible imputed datasets and combining the results from each of them. Because we can never know the true values of the missing data the multiple imputation procedure must create multiple copies of the data from the empirical predictive distributions of the observed values. As such multiple imputation is based on a Bayesian

approach. Standard statistical methods are then used to fit the model of interest in each dataset, and the results differ because of the uncertainty about the imputed values. The results are only meaningful when averaged together to give overall estimated associations. Estimates and standard errors are calculated using Rubin's rules (REF RUBIN 1987) which take into account the within and between imputation variation. Recent developments in statistical software permit some degree of automation of the process of multiple imputation; see ICE in Stata (REF ROYSTON) or MICE in R (REF VAN BUREN).

In this paper we study the Myocardial Infarction National Audit Project (MINAP) database (REF RCP REPORTS). MINAP was first collected data in 2000 and originally recorded information on each patient presenting to hospitals in England and Wales suffering with a myocardial infarction. The dataset has since expanded so that it covers all aspects of patient care of patients having acute coronary syndromes (ACS). Data is now collected from 234 acute hospitals in England and Wales, and the priority is to provide useful data with which to analyse patient care (REF UCL DATA COLLECTION MANUAL). The MINAP dataset considered in this paper spans the four calendar year period between 2004 and 2007 inclusive, in which there are a total of 340,983 admissions recorded. Of these admissions 290,483 ultimately had a diagnosis of ACS, 32,385 patients had non-ACS diagnoses such as chest pain with uncertain cause whilst 10,115 patients had no diagnosis recorded.

We determine the extent of missing data and implausible data in key fields for risk modelling. Risk modelling is an important part of ACS care because it can be used to guide treatment strategies depending upon the classification of the patients degree of risk assessed using established risk factors. There are several important risk score in ACS management, including the Global Registry of Acute Coronary Events (GRACE) score (REFS); the TIMI Risk Score for Unstable Angina/Non-ST Elevation myocardial infarction (REFS) and the Evaluation of Methods and Management of Acute Coronary Events (EMMACE) risk score (REF EMMACE & EMMACE II). These risk scores provide the basis of our decisions about which variables to impute to allow the datasets to be used for risk modelling. We have also included variables that were identified as clinically important for ACS management specifically in the UK.

We aim to improve the usability of MINAP data beyond its current level by taking positive action to address the concerns relating to missing data. We develop a multiple imputation scheme to mitigate the effects of missing data in MINAP following the guidelines suggested in (REF STERN, CARPENTER & WHITE). We investigate the effects of multiple imputation on the dataset, and perform an example analysis based on the EMMACE risk score (REF EMMACE), to demonstrate the differences in the numerical results.

2. Multiple imputation

(*a*) *Multiple imputation using chained equations*

The multiple imputation procedure consists of several independent steps that when combined produce the multiply imputed analysis results of interest. Beginning with the incomplete data, a number of imputations (i.e. multiple imputations) are performed yielding a collection of imputed data sets. The analysis of interest is then performed in each of the imputed data sets. The results of the analysis in each imputed data set are then combined to produce the final multiply imputed analysis result.

Let $Y_j,\ j = 1, \ldots, p$ be one of p incomplete variables and $Y = (Y_1, \ldots, Y_p)$. The observed and missing parts of Y_j are Y_j^{obs} and Y_j^{miss}, respectively. Let the number of

imputations be $m \geq 1$. The i-th data set is denoted $Y^{(i)}$ with $i = 1, \ldots, m$. Let $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ denote the collection of $p-1$ variables in Y except Y_j. Let Q denote the quantity of interest (a regression coefficient, say), but more generally any quantity of interest.

Imputation is achieved using *chained equations* to allow imputation of variables which themselves depend on missing values. Hypothetically let the complete data Y be a partially observed random sample of the multivariate distribution $P(Y|\theta)$. Assume further that the multivariate distribution of Y is specified completely by θ, a vector of unknown parameters. The objective is to find the multivariate distribution of θ, either explicitly of implicitly. The chained equations obtains an estimate of the posterior distribution of θ by sampling iteratively from conditional distributions of the form

$$\begin{gathered} P(Y_1|Y_{-1}, \theta_1) \\ \vdots \\ P(Y_p|Y_{-p}, \theta_p) \end{gathered}$$

Starting from a random draw from the observed marginal distributions, the k-th iteration of chained equations is a Gibbs sampler that successively draws

$$\begin{aligned} \theta_1^{*(k)} &\sim P(\theta_1|Y_1^{\text{obs}}, Y_2^{(k-1)}, \ldots, Y_p^{(k-1)}) \\ Y_1^{*(k)} &\sim P(Y_1|Y_1^{\text{obs}}, Y_2^{(k-1)}, \ldots, Y_p^{(k-1)}, \theta_1^{*(k)}) \\ &\vdots \\ \theta_p^{*(k)} &\sim P(\theta_p|Y_p^{\text{obs}}, Y_2^{(k)}, \ldots, Y_p^{(k)}) \\ Y_p^{*(k)} &\sim P(Y_p|Y_p^{\text{obs}}, Y_1^{(k)}, \ldots, Y_p^{(k)}, \theta_p^{*(k)}) \end{aligned}$$

where $Y_j^{(k)} = (Y_j^{\text{obs}}, Y_j^{*(k)})$ is the k-th imputed variable at iteration k. Observed that previous imputations only enter the present imputation through it's relationship with other variables and not directly, meaning that unlike many applications of MCMC convergence can be very fast.

Once the m imputed datasets have been created the next step is to estimate $\hat{Q} = (\hat{Q}^{(1)}, \ldots, \hat{Q}^{(m)})$ which is the vector of imputed estimates, which differ because the imputations in each of the m datasets are different reflecting our uncertainty about the imputed value. The estimates $\hat{Q}$ are finally then pooled using Rubin's rules, which are described in Subsection 2b.

(*b*) *Estimation of parameters: Rubin's rules*

Following multiple imputation we have a number of plausible possibilities for each missing entry, reflecting our uncertainty about the imputed values. To generate estimates or models from the imputed data sets we perform out analysis in each of the imputed datasets and combine the results using Rubin's rules. Rubin's rules state that the average of the parameter of interest is the multiply imputed estimator and its sampling variance is the average of the completed data sampling variances inflated by the between completion variance (REF RUBIN).

Mathematically, let $\hat{\theta}_m,\ m = 1, \ldots, M$, be the set of complete estimates of the population quantity θ and let $\hat{s}_m^2$ be the estimates of the completed data sampling variances. Then the multiply imputed estimator of θ is defined by Rubin's rules as

$$\tilde{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{2.1}$$

and its sampling variance is estimated by

$$\tilde{s}^2 = \hat{U} + (1 + 1/M)\hat{B} \tag{2.2}$$

where

$$\hat{U} = \frac{1}{M} \sum_{m=1}^{M} s_m^2 \tag{2.3}$$

and

$$\hat{B} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \tilde{\theta}\right)^2. \tag{2.4}$$

The first term $\hat{U}$ estimates the sampling variance of $\hat{\theta}$ that would be realised if data were complete. The second term $\hat{B}$ is an inflation term due to missingness and gives the appropriate inflation for $M \to \infty$. The term $\hat{B}/M$ is a correction to account for having used a finite number of imputations.

3. Imputation scheme

(*a*) *Variable selection*

As a general rule using every bit of available information yields multiple imputations that have minimal bias (REF MENG 1995; COLLINS 2001). This principle suggests that the number of predictors should be as large as possible. Practically however, the imputation scheme should be at least as rich as the models that the analyst intends to use for their statistical modelling after the imputations: a property referred to as *congeniality* (REF MENG 1994). When the imputation and the analysis model are congenial then multiple imputation gives results that are asymptotically equivalent to maximum likelihood.

Because our intention is to provide imputed datasets that will be applicable for general modelling of risk of early mortality following heart attack, we have attempted to include as many variables as possible that might be of interest to analysts. Using several risk scores as a guide we constructed a list of variables of interest to impute and we have also included additional variables, known as *auxiliary variables*, that can improve prediction of the missing values in the variables of interest. The subset of variables to be imputed is summarised in Table 1.

Table 1 has been divided into *patient demographics*, *medical history*, *drug usage* and *diagnostic variables* for convenience. The index of multiple deprivation score is a measure of the social status of the patient and is based on the patients address (REF IMDdocs online). There are several ACS diagnoses that are of importance in MINAP. ST-elevation myocardial infarction (STEMI) is diagnosed using the electrocardiograph (ECG) due to a characteristic elevation of the ST-segment on the ECG trace, (REF). Any myocardial

Table 1. *Variables and summary of missing data. Key:* †*primary percutaneous coronary intervention surgery;* ‡*coronary artery bypass graft surgery;* ⋆*angiotensin-converting enzyme inhibitor;* ⋆⋆*angiotensin II receptor antagonist.*

Variable	Variable type	% missing	Imputation method
Demographics			
Age	Continuous	0.4	Predictive mean matching
Mortality (alive/dead at 30 days)	Binary	4.7	Predictor only
Month of admission	Categorical	2.9	Polytomous regression
Ethnicity	Categorical	23.8	Polytomous regression
Index of multiple deprivation (IMD)	Continuous	16.2	Predictive mean matching
Total ACS related admissions	Continuous	4.9	Polytomous regression
Hospital of admission	Categorical	0.0	Predictor only
Medical history			
Hypertension	Binary	4.2	Default imputation
Stroke	Binary	9.8	Default imputation
Peripheral vascular disease	Binary	9.7	Default imputation
Heart failure	Binary	9.5	Default imputation
Renal failure	Binary	9.9	Default imputation
Myocardial infarction	Binary	5.0	Default imputation
Hyperlipidemia	Binary	10.6	Default imputation
Previous/existing angina	Binary	6.4	Default imputation
Previous PCI†	Binary	8.9	Default imputation
Previous CABG‡	Binary	8.6	Default imputation
History of smoking	Categorical	11.1	Polytomous regression
History of diabetes	Categorical	3.8	Polytomous regression
Drug usage			
Thiazide	Binary	20.6	Default imputation
Beta blocker	Binary	43.7	Default imputation
Loop dieuretic	Binary	19.9	Default imputation
ACE inhibitor⋆	Binary	19.2	Default imputation
ACEARB⋆⋆	Binary	19.1	Default imputation
Spironolactone	Binary	21.1	Default imputation
Diagnostics			
Electorcardiograph (ECG)	Categorical	9.6	Polytomous regression
Final diagnosis	Categorical	3.0	Polytomous regression
Cholesterol	Continuous	34.0	Predictive mean matching
Troponin biomarker concentration	Categorical	14.5	Polytomous regression
Glucose concentration	Continuous	46.9	Predictive mean matching
Heart rate	Continuous	14.4	Predictive mean matching
Systolic blood pressure	Continuous	15.8	Predictive mean matching

infarction that does not show this characteristic ST-segment elevation is referred to as non-STEMI (NSTEMI), and there are several subcategories of NSTEMI depending upon chemical biomarkers, such as Tropoinin (REF DIAGNOSIS GUIDANCE). Troponin is a chemical biomarker that is released when cardiac muscle dies, and is a positive indication that a myocardial infarction has been experienced by the patient even when the ST-elevation is not present. Full details of the diagnostic guidance we have used can be found in (REF EHJ PAPER).

(*b*) *Missingness mechanisms*

Missingness mechanisms are assumptions about the data that describe the way in which we believe the missing data are, or are not, related to the observed data. It should be noted that they are assumptions, which justify the analysis and are not themselves properties of the data. Four mechanisms are acknowledged to exist, and these are as follows:

- *Missing completely at random*: there are no systematic differences between the observed values and the missing values.
- *Missing at random*: any systematic difference between the missing values and the observed values can be explained by differences in observed data
- *Missing not at random*: even after the observed data are taken into account, systematic differences remain between the missing values and the observed values.
- *Missing by design*: the data are missing because of the design of the questionnaire or data collection strategy.

If data are assumed to be missing completely at random then the observations constitute a random sample of the 'full' dataset and no bias will be present in analyses as a consequence of missing data. Multiple imputation assumes that data are missing at random: that is the observed variables are predictive of the missing values. Analyses based on multiply imputed data will avoid bias only if enough variables that predict the missing values are included in the imputation models. Failure to do so may render the missing at random assumpition implausible and analyses based on the data may be biased. If the data are assumed to be missing not at random, then the missing values depend on some additional factor that has not been observed and which cannot be used to predict the missing values. Data that are missing by design can be dealt with, for example by inverse probability weighting methods (REF), but are not considered further in this paper.

The important issue is that there is is sufficient plausibility in the missing at random assumption to justify the belief that the missing values can reasonably be predicted by the observed values. Including as many predictors as possible tend to make the missing at random assumption more plausible (REF SCHAFER 1997), although including more than 15 to 25 predictors gives a negligible increase in the variance explained in the prediction equations (REF VAN BUREN 1999).

A common misunderstanding of multiple imputation is that it is restricted to data assumed to be missing at random. There are techniques to deal with data that have been assumed to be minning not at random, details of which may be found in (REF: RUBIN 87; LITTLE 2009; ALBERT & FOLLMAN 2009).

(*c*) *Missingness patterns*

We examined the pattrerns of missingness to gain further insight into the possible ways by which the data came to be missing. In data which is collected longitudinally through the patients stay we looked for monotone missingness: that is Y_j is observed only if Y_{j-1} is observed. In doing so no bobvious patterns were found. We also looked at which variables might have an effect on missingness in other variables. The most significant observation was that when missing values occur, systolic blood pressure and heart rate are most often missing sumultaneously as shown in Table 2.

Table 2. *Missing value patterns in systolic blood pressure and heart rate.*

Percent	Systolic BP (mmHg)	Heart rate (bpm)
83	recorded	recorded
13	missing	missing
2	recorded	missing
<1	missing	recorded
100%		

Table 2 shows that where either of systolic blood pressure or heart rate is missing, the most common situation is that both will be missing. Simultaneous missingness constitutes 13% of all of the cases, but over 80% of those cases in which one or both values is missing.

We used logistic regression with the outcomes 'missing' and 'not missing' for each variable that we imputed to identify potential predictors. Where non-linear realtionships were found using powers of predictors we also examined these relationships in more detail using generalised additive models: see Subsection 3g.

(*d*) *Handling interactions: imputation by splitting*

The most important interaction that we discovered in the MINAP data is between age and sex. Because of the potential importance of this interaction in risk modelling it must be included in the imputation scheme. Generally speaking, female patients tend to have their heart attacks later in life compared to males: see Figure 1. There is also a dependence of 30 day mortality on the patient's year of admission, partly as a result of changing processes of care and treatments (REF THROMBO/PCI). The dependence of mortality rate on year is considered further in Subsection 3h.

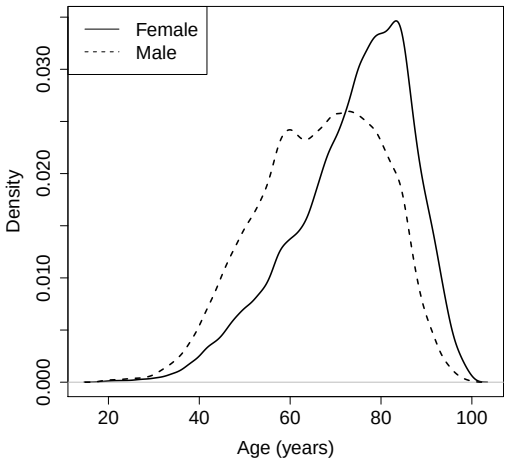

Figure 1. Age and sex interaction. Histograms showing the age distribution by sex.

An additional factor to consider is that the amount of missing data in most variables decreases over time. Table 3 tracks the missingness of several key variables over the four year period considered. It is evident that the overall trend is a reduction in the amount of

Table 3. *Missingness by year in a selection of variables*

	2004 missing %	2005 missing %	2006 missing %	2007 missing %
Age	1.9	0.9	0.2	0.2
Systolic blood pressure	18.8	16.4	13.6	13.9
Heart rate	16.0	14.2	13.1	13.9
Glucose	79.6	50.7	29.0	21.8

missing data year-on-year but that in some variables, most noticeably glucose, there is a marked variation between the years. The variation in glucose missingness can be explained by changing practices in data collection during the four year period.

Because MINAP is a very large database and imputation of the whole dataset is computationally challenging, we have addressed imputation of these interactions by splitting the dataset according to sex and year of admission. This constitutes *imputation by splitting*. Splitting reduces the computational burden, ensures the imputations respect the age–sex interaction, the variations of outcome with year and the annual variation of missing data.

(*e*) *Default imputations*

Conditions that the patient has previously suffered from can affect their risk of mortality, as can the various standard drug therapies on admission. In MINAP both the medical history and drug therapies are reported using 'yes/no' fields. After discussions with clerical and medical staff we determined that the pragmatic approach to imputing the missing 'yes/no' responses was to impute a 'no' response if the information was missing, provided that other information did not imply that 'no' would be the wrong response. We chose this approach because it is more likely that a condition or treatment would go unrecorded (i.e. missing) if the patient had no history of that condition or did not receive the treatment. In instances where the patient had been given a treatment (i.e. the entry should be 'yes') it would be negl01igent to have not recorded this information. Table 1 shows the imputation scheme used for each variable including those for which default imputation has been used.

(*f*) *Non-normal variables*

Multiple imputation assumes of normality of the variables being imputed, and it is important to check that this assumption will be approximately satisfied. For those variables that are found to have a non-normal distribution a transformation to approximate normality is required. A Box–Cox transformation or logarithmic transformation will usually suffice (REF WHITE). After imputation the transformation can be reversed to recover the distribution of the imputed variable on its natural scale. In this work we have opted for the logarithmic transformation for the variables *heart rate*, *systolic blood pressure* and *glucose*, all of which are sufficiently non-normal to cause concern about the validity of the normality assumption.

Figure 2 shows the QQ-plots for heart rate and its logarithmic transform. The heart rate variable shows non-normality over much of the distribution, whereas the logarithmic distribution, although not perfect in the left tail, is somewhat closer to normality than the untransformed variable.

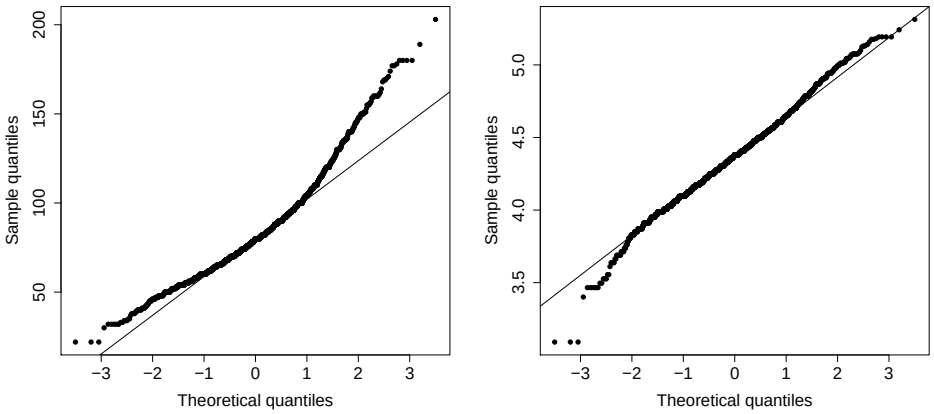

Figure 2. QQ-plot of heart rate (left) and the logarithm of heart rate (right).

(*g*) *Non-linear associations*

It is important to consider the possibility of non-linear associations between the continuous variables as an incorrect assumption of linearity will bias the higher-order terms towards the null. We did extensive analysis for each variable to identify potential predictors, and to quantify the associations between the variables and the potential predictors. We used generalised additive models (GAMs) (REF HASTIE &/or WOOD) to explore potential predictors for the continuous variables and to identify non-linear relationships. A generalised additive model is a generalised linear model with a linear predictor involving a sum of smooth functions, usually cubic splines, of the covariates. Because the generalised additive model is a sum of smooth functions it is able to identify characterize non-linear regression effects. A further advantage of generalised additive models is that there is a degree of automation to the fitting of the smooth functions (REF WOOD & R LIBRARY), and so the specification of the model in the computer software is no more challenging than an ordinary linear model.

The generalised additive model output shown in Figure 3 suggests by inspection that the age of male patients depends upon the square of their systolic blood pressure, whilst it depends on the cube of heart rate.

There are a selection of methods for imputing variables that have non-linear relationships with their predictors. In this paper we have opted to use predictive mean matching (PMM) to deal with non-linearity in the prediction equations (REF LITTLE 1998). PMM is a general-purpose semi-parametric imputation method in which the imputations are confined to the observed distribution. PMM can also preserve non-linear relations even if the structural part of the model is not correctly specified. Furthermore PMM can handle non-mormal distributions (see Subsection 3f), although for robustness we have used the logarithmic transformation.

A possible disadvantage of PMM is that it may fail to produce enough between imputation variation when the number of predictors is small (REF VAN BUREN). As the sample sizes in MINAP are very large and the number of predictors is also large, we believe that PMM offers a useful method of imputing continuous variables and preserving non-linear relationships in the prediction equations. Moreover, partly to mitigate concerns

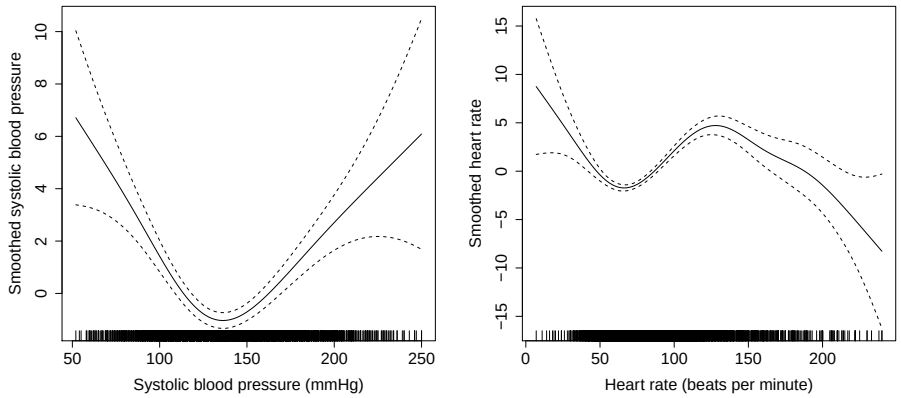

Figure 3. Generalised additive model for male age showing that age has different non-linear associations with systolic blood pressure and heart rate.

regarding insufficient between imputation variation, we have elected to use 25 imputations in our work rather than the 'standard' five imputations suggested in some of the literature (REFS). (REF WHITE) also advises using more than five imputations, and suggests that at least 20 imputations is advisable.

(*h*) *Effect of seasonal variation*

It is generally accepted in the cardiology literature that admissions for heart attack show seasonal variations (REFS). In most places in the world there are substantially more admissions with generally poorer outcomes in the winter months, although the reasons for this are not yet clearly understood. By analysing admissions and mortality data using time series, we have also discovered an important seasonal variation in the MINAP data.

Figure 4 summarises the time series analysis of 30-day mortality. The four plots in Figure 4 from top to bottom show the original mortality data; the seasonal component of the mortality data; the year-on-year trend and the unexplained (random) variation. There is a cyclic seasonal component showing peaks of mortality at the start of each year (winter months) and troughs near the middle of each year (summer months). There is also a generally decreasing trend in mortality year-on-year as shown by the trend plot. To capture the seasonal variation in the imputations we included the month of admission to hospital in the imputation scheme, and the variation between years is captured by the splitting; see Subsection 3d.

(*i*) *Inclusion of the outcome as a predictor*

It is important to include the outcome variable (in this case mortality status at 30 days) as a predictor in the imputation model. Failing to include the outcome will severely dilute the associations between the outcome and the other variables (REF MOONS ET AL 2006 & QRISK PAPERS). If imputation is being performed on data with a survival outcome, then both the event *and* the censoring must be imputed (REF. VAN BUREN 1999). Missing outcomes will also be imputed, but the results of the imputations are excluded in the final analyses. The imputed outcomes are discarded because a correctly imputed outcome adds

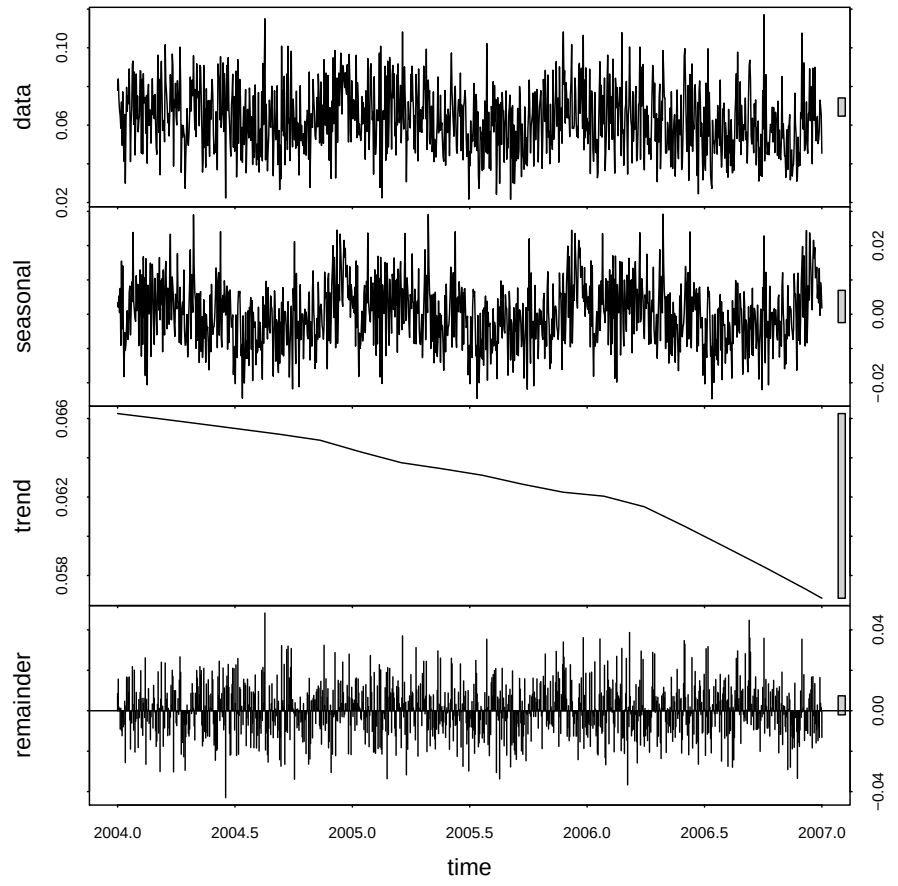

Figure 4. Time series decomposition of the number of deaths reported per day for the four year period January 2004 to December 2007.

nothing except Monte Carlo error whilst an incorrectly imputed outcome adds more error (REF CANTAB NOTES).

(*j*) *Clustering*

The MINAP data are collected from many hospitals, each of which may experience a different patient profile (i.e. clustering), and this effect could be taken into account by performing a multi-level imputation with hospital of admission included as a random effect. There are very few multi-level imputation packages available (REFS MLWIN & MICE), and this continues to be an active area of research.

Because the clusters (hospitals) are generally very large, we included the hospital of admission as a fixed-effect and therefore some allowance is made for the fact that MINAP is a multi-centre observational study. The errors induced by this decision are unlikely to

bias the results of analyses based on our imputed data, unless the clustering is explicitly the focus of the analysis.

(*k*) *Potential problems*

Collinearity might be encountered with variables that are linearly related, or almost so. The solution is to carefully redefine the imputation model to reduce or remove the duplicated information in the prediction equations. We encountered collinearity whenever using the hospital of admission and the index of multiple deprivation (IMD) score as predictors for the same variable. Because heart attack patients are most likely to be taken to a hospital close to their home address, the IMD score and the hospital of admission carry very similar information; that is the IMD scores of patients close to each hospital will be similar. The solution was to use the hospital of admission as a predictor for missing IMD scores, and then to use the IMD score, rather than hospital *and* IMD score, to predict other missing variables. This approach also had the very significant advantage of reducing the computational burden because the hospital of admission is now used only in the imputation of one variable.

Perfect prediction occurs during the imputation if one of the predictor variables always takes a particular value, i.e. the predictor has no variation. Several statistical packages that have multiple imputation libraries, including MICE in R (REF VAN BUREN R CODE) and Stata (REF ROYSTON), avoid this by using *augmented logistic regression*. Briefly, the augmented logistic regression procedure adds a small number of extra observations into the dataset so that no prediction is perfect, and then assigns very low weights to these observations thus ensuring successful draws from the predictive distribution.

Feedback is a potential problem with correlated predictors. For example, a higher imputed value of one predictor will produce a higher than average imputation for the correlated variables. If the correlations are strong these higher imputed values are fed back into the imputations for the original predictor and the cycle continues. Such behaviour should be identified and remedied.

Incompatibility occurs when the joint distributions of predictors do not exist in an analytical sense. For example two linear regression specify a joint multivariate normal given certain regularity conditions (ARNOLD & PRESS 1989). The joint distribution of a linear regression and a proportional odds regression model is unknown and yet is easily specified in multiple imputation software. The simulation work that is available suggests that incompatibility is not a serious problem in practice (REF VAN BUREN 2006; DRECHSLER & RASSLER 2008).

Failure to converge. It is of particular importance to check convergence of the Gibbs sampler when using the PMM algorithm as it can be sensitive to the imputation model and in the worst cases may become locked at the first imputation. Locking of the sampler should also be anticipated when using *passive imputation*, which is the imputation of values directly using a combination of other imputed variables. An example of passive imputation might be the imputation of body mass index (BMI) when mass or height are imputed. BMI could be imputed directly (i.e. passively) from the formula $BMI = mass/height^2$. Examples of poor convergence and locking of the sampler are shown in (REF VAN BUREN). Assessment of convergence for our imputations is covered in Subsection 4a.

4. Results

(*a*) *Checking the imputations: convergence*

There is no definitive method for checking the imputations or the within imputation iterations of the Gibbs sampler. The chain mean and standard deviation at each iteration can be plotted and on convergence the different streams should freely intermingle without showing any definite trends (REF VAN BUREN). In general convergence of the MICE

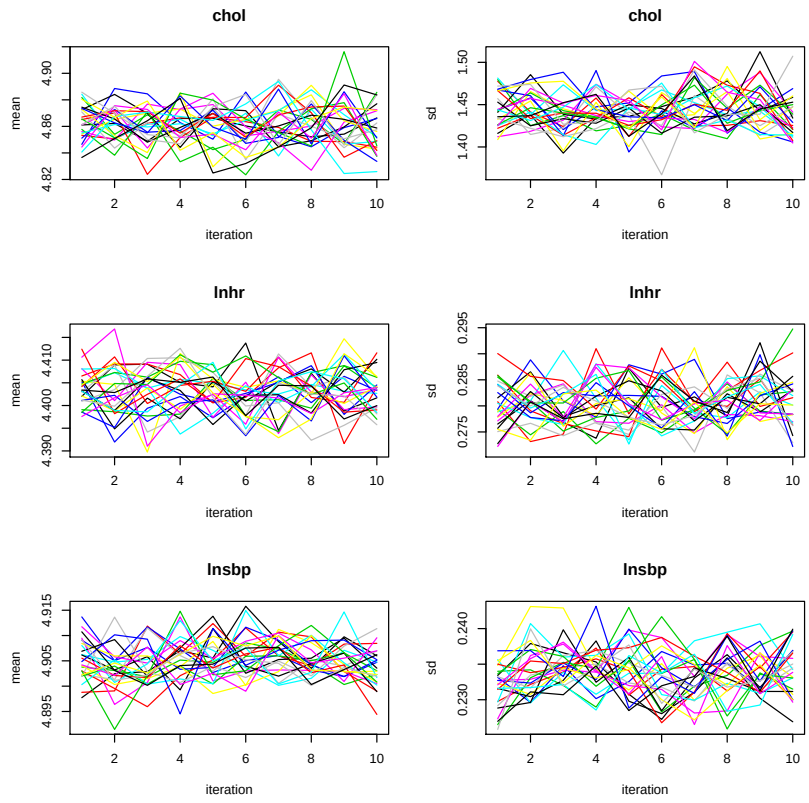

Figure 5. Examples of chain means and standard deviations for cholesterol, log heartrate and log systolic blood pressure.

algorithms in R, when healthy, is also rapid even if the starting imputation is poor. This is because previous imputations enter the current imputation only through their relations with the other variables and not directly. Although the default setting of five iterations is often sufficient, in this work we used 10 within imputation iterations for additional assurance of successful convergence.

The iterations shown in Figure 5 show healthy performance of the Gibbs sampler for cholesterol, log heartrate and log systolic blood pressure. The chain means and standard deviations in Figure 5 show healthy convergence of the sampler for each of the variables shown. Some variables, for example troponin, show a definite trend in the early iterations. This trend however, is replaced by the desirable freely intermingled chains after a small

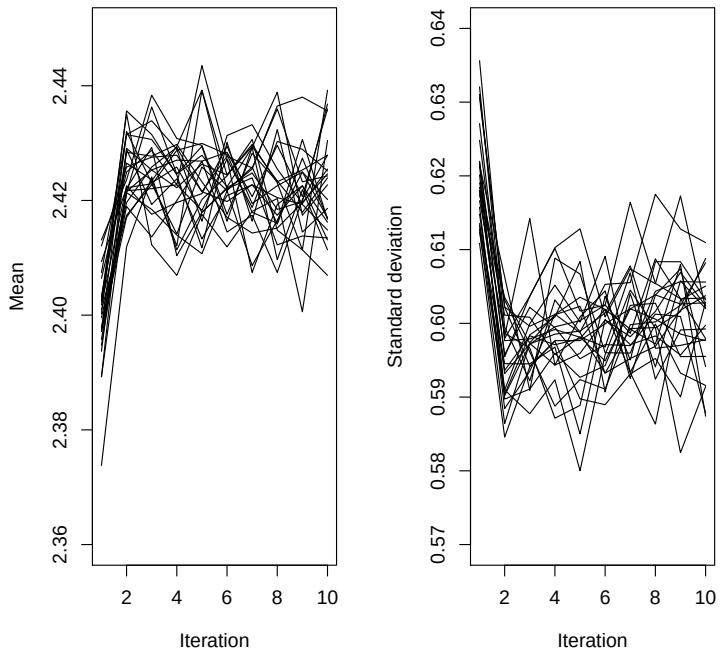

Figure 6. Examples of chain means and standard deviations showing an early trend becoming healthy convergence.

number of iterations. Figure 6 shows an example of this behaviour and highlights that for those variables for which there is an initial trend, the ultimate performance of the Gibbs sampler over the 10 iterations is satisfactory.

We have investigated the chains for each variable, and we are satisfied that convergence is satisfactory for each of our imputations.

(*b*) *Plausibility of missing at random assumption*

In Section 3b we mentioned that to make the missing at random assumption plausible it was important to include for each variable as many predictors as possible. In setting up our imputations we have ensured that we have included as many clinically relevant predictors as possible for each variable unless there was reason to do otherwise. For example, collinearity prevented hospital of admission and IMD score being used as predictors simultaneously for some variables; see Subsection 3k.

In general, a good imputed value is one that could have been observed had it not been missing. The missing at random assumption can *never* be tested on the observed data (REF: VAN BUREN), but we can check that the imputations are plausible by comparing the distributions of the observed and imputed values for each imputed data set. If there are large differences between the observed and imnputed values it would be important to consider why the differences have ocurred.

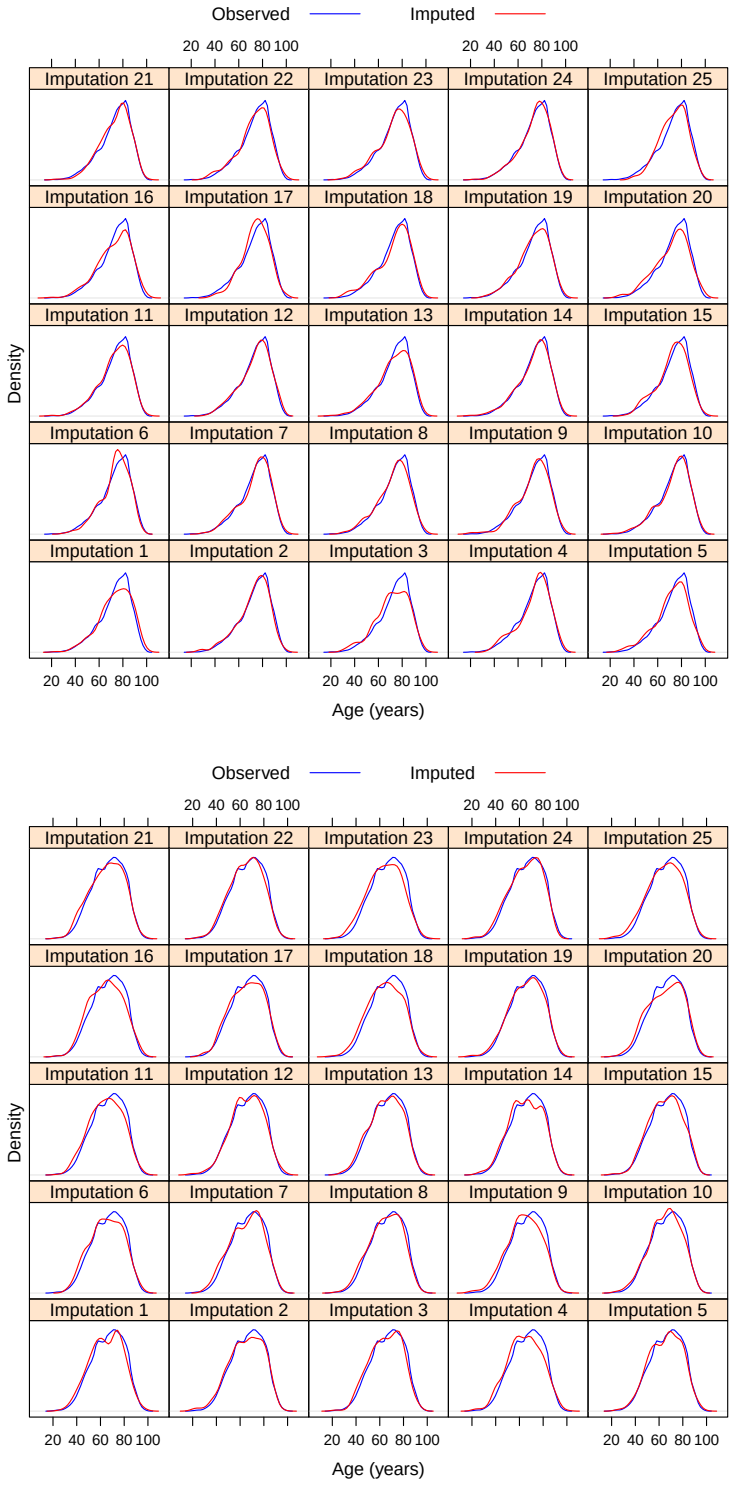

Figure 7. Examples of observed and imputed distributions for female age (top) and male age (bottom) in 2004.

Figure 7 shows the observed and imputed distributions for age for both sexes in each of the 25 imputed datasets (see also Figure 1). The observed and imputed distributions are similar in each of the imputed datasets, which suggests that the missing at random assumption is plausible. The left skew of the female distribution has also been correctly represented by the imputation scheme. Some imputations appear to predict a lower peak for the age distribution with the consequence of a higher density across the whole age range; for example imputation 3 of the females. A similar effect occurs in the males in, for example, imputation 20. Figure 7 also shows that the different distributions of the male and female ages are correctly represented, confirming that imputing by splitting has captured the age–sex interaction mentioned in Subsection 3d. In general the differences between the observed and imputed distributions are minor, and so the imputations are plausible.

(*c*) *Comparison of imputed data and complete case analyses*

An important consideration when considering the usefulness of multiple imputation is that the number of complete cases available diminishes with each variable added to an analysis. For example, the missingness of age is 0.4%, but if this were included in a complete cases analysis with cholesterol, the missingness for the analysis would be in excess of 34%. Therefore an imputation in one variable has value in all of the others because of the additional cases that can be reclaimed for the analysis.

We present a comparison between the results that would be realised using complete cases analysis and those obtained using our multiply imputed data sets. The imputed results are calculated using Rubin's rules as outlined in Subsection 2b. Table 1 of (REF MORROW CIRCULATION 2000) provides the univariate risk of 30-day mortality stratified by presenting characteristics. We have used this table as a guide for comparing demographics between our datasets, but have also included some additional variables that we included in our imputations. Table 4 compares complete cases and imputed data for those patients diagnosed with ST-elevation myocardial infarction in MINAP. There are some variables in Table 1 of (REF MORROW CIRCULATION 2000) which are not routinely recorded in MINAP, and these have been omitted from Table 4 for clarity.

Table 4 shows that there is little difference in mean age between the complete case and multiply imputed datasets. This is as expected since the overall missingness in age is very low (0.4%). The largest difference in mean age occurs in male patients and is 0.5 years. The multiply imputed data suggests more patients over 75 years than the complete cases data and vice-versa with patients over 65 years. There are minor differences between the means for glucose, systolic blood pressure, heart rate and IMD score. The multiply imputed data predicts nearly 6000 more smokers or ex-smokers than the complete cases data would suggest and 3000 more non-smokers. The factors in the cardiovascular risk section were all imputed using a negative default imputation and so the number of positive indications does not change between the datasets.

Table 5 compares complete and imputed MINAP data sets using the Evaluation of Management and Methods for Acute Coronary Events (EMMACE) risk score. Table 5 shows the odds ratio for mortality per standard deviation with 95% confidence intervals and t-values for complete case analysis and multiple imputation.

The most obvious observation is that there is a large increase in the number of cases available when the imputed data is used. For the four years the percentage increase in the number of cases included in the analyses are 22%, 19%, 17% and 24% respectively. The increase in the number of cases taken over all of the years is 20%. Therefore complete cases

Table 4. *Comparison of complete cases and imputed data characteristics for patients diagnosed with ST-elevation myocardial infarction. Key: *Imputed by default method; † Primary percutaneous coronary intervention surgery; ‡ Coronary artery bypass graft surgery.*

	number of complete cases	Complete cases MINAP data	Multiply imputed MINAP data (113,445 cases)
Age, years (population)	112,888	66.04 (65.96,66.12)	66.02 (65.94,66.10)
Age, years (female)	33,774	72.17 (72.03,72.32)	72.18 (72.04,72.32)
Age, years (male)	78,291	63.40 (63.31,63.50)	63.45 (63.36,63.54)
>75 years	32,697	33,438	33,608
>65 years	60,624	61,365	64,259
Heart rate	97,611	78.80 (78.66,78.94)	78.80 (78.66,78.93)
Systolic BP	96,298	137.75 (137.56,137.94)	137.60 (137.43,137.78)
Glucose	61,637	8.46 (8.44,8.50)	8.53(8.50,8.56)
Cholesterol	84,881	5.27 (5.26,5.28)	5.24 (5.23,5.26)
IMD score	94,759	21.59 (21.49,21.70)	21.57 (21.49,21.70)
Risk factors			
Smoking status	102,044		
Current		38,849	44,792
Past		30,760	36,747
Never		23,383	26,187
Not known		9052	11,185
Diabetes	106,362	14,518	18,010
Prior hypertension*	106,008	44,248	44,248
Cardiovascular history			
Peripheral vascular disease*	102,394	3525	3525
Cerebrovascular disease*	102,237	6221	6221
Prior PCI† *	103,493	5586	5586
Prior CABG‡ *	103,804	2850	2850
Prior angina *	105,591	21,256	21,256

analyses will contain only four fifths of the data that is available from the multiply imputed data sets, showing the value of multiple imputation to access information from cases that would otherwise be lost to missingness. Using more variables in a model would inevitably entail more data lost to missingness when analysing complete cases, and therefore a greater benefit, in terms of case recovery, from using the multiply imputed data.

The results for both the complete cases and multiply imputed analyses show that a one standard deviation increase in age (appromimately 13 years) increases the odds of death at 30 days by 10%. A one standard deviation increase in heart rate (approximately 22 beats per minute) increases the risk of mortality by 5%, whilst a one standard deviation increase in systolic blood pressure (typically 30 mmHg) results in a 5% decrease in mortality risk. The larger number of cases available in the multiply imputed data has the effect of shrinking the confidence intervals, which can quickly be assessed by noting the generally inflated t-values. This increased precision in the estimates reinforces the benefit of using the multiply imputed data.

5. Conclusions

We have presented multiple imputation for a national cardiac care routine audit database. Despite criticisms of the use of audit data for research where there are missing values, we

Table 5. *Comparison of complete case and impued results using the EMMACE risk score variables. Odds ratios quoted per standard deviation. Number of cases appear in bold at the top of each analysis.*

Mortality at 30 days	Complete OR	t value	95% Conf. Interval	Imputed OR	t value	95% Conf. Interval
2004	**23,525**			**30,217**		
Age	1.0954	3.47	(1.0404,1.1533)	1.1001	4.05	(1.0504,1.1520)
Heart rate	1.0335	1.27	(0.9824,1.0872)	1.0197	0.83	(0.9737,1.0679)
Systolic BP	0.9560	-1.71	(0.9081,1.0065)	0.9440	-2.39	(0.9003,0.9897)
2005	**24,294**			**29,843**		
Age	1.1452	5.00	(1.0859,1.2078)	1.1461	5.54	(1.0922,1.1444)
Heart rate	1.0380	1.44	(0.9866,1.0920)	1.0272	1.12	(0.9799,1.0768)
Systolic BP	0.9398	-2.33	(0.8919,0.9902)	0.9606	-1.61	(0.9147,1.0089)
2006	**23,309**			**28,157**		
Age	1.0866	2.87	(1.0193,1.1067)	1.0945	3.42	(1.0393,1.1527)
Heart rate	1.0510	1.79	(0.9998,1.0047)	1.0492	1.86	(0.9975,1.1037)
Systolic BP	0.9492	-2.37	(0.9963,1.0001)	0.9489	-1.95	(0.9003,1.0002)
2007	**19,068**			**25,228**		
Age	1.0579	1.68	(0.9906,1.1299)	1.0667	2.19	(1.0068,1.1301)
Heartrate	1.0916	1.79	(1.0254,1.1620)	1.0858	2.91	(1.0273,1.1477)
Systolic BP	0.9575	-1.82	(0.8970,1.0221)	0.9664	-1.15	(0.9117,1.0245)
All years	**90,196**			**113,445**		
Age	1.1020	6.80	(1.0716,1.1333)	1.1070	7.91	(1.0794,1.1352)
Heart rate	1.0471	3.34	(1.0192,1.0758)	1.0387	3.01	(1.0133,1.0648)
Systolic BP	0.9543	-3.31	(0.9282,0.9811)	0.9577	-3.32	(0.9336,0.9825)

have demonstrated that careful use of multiple imputation could improve the amount and quality of data in such databases.

We have attempted to impute a large number of variables that we feel are important to generiac risk modelling (Table 1). As an important caveat to our work however, it should be noted that *imputations are analysis specific*, and it is vital to check for each analysis that the imputation schemes used are robust and compatible with the proposed analysis. *Failure to check the compatibility of a proposed analysis with the imputation models and the use of an incompatible analysis model will almost certainly result in biased analyses*.

We considered each variable in terms of amount of missing data, and how that missingness was related to other variables (e.g. systolic blood pressure and heart rate are often simultaneously missing). We also considered important interactions such as age–sex, outcome–year, and missingness–year and accounted for these by splitting the MINAP dataset by sex and year: imputation by splitting.

We considered which variables we expected to have non-normal distributions, such as concentrations of biomarkers, and used transformations and appropriate imputation schemes to handle these. For continuous variables we also used an imputation scheme that accounted for non-linear behaviour between variables variable and their predictors. Because mortality and admissions for acute coronary syndromes are known to vary seasonally we also accounted for this in our imputations. As MINAP is a multi-centre observational study we have also included the hospital of admission as a fixed effect in order to take some account of clustering.

We have checked our imputations both to assess convergence and the plausibility of the missing at random assumption upon which the imputations are founded. Convergence was

shown to be satisfactory. Although it is impossible to check the missing at random assumption using the observed data, we are satisfied that we have included sufficient predictors to render the missing at random assumption plausible. Furthermore there are no significiant differences between the observed and imputed distributions which adds further evidence to the plausibility of the missing at random assumption.

We showed that there are minor differences between the means for the continuous variables (Table 4) although there are relatively insignificant. Table 4 does demonstrate the multiply imputed data predicts a larger number of smokers and diabetics (proportions?)

From our results using the EMMACE risk score variables, imputation improves information content of data by recovering missing cases. This is reflected in greater precision of estimates of odds ratios. Therefore the main improvement given by our multiply imputed data is an improvement in the number of cases that can be analysed in comparison to complete cases analysis. Table 5 shows that year-on-year the analyses based on imputed datasets contain approximately 20% more cases than the complete cases analysis.

In conclusion we beleive that our imputation scheme can be used to improve the number of cases available for analysis in the Myocardial Infarction National Audit Project. In so doing the robustness and utility of MINAP data is improved beyond its current levels. This will assist MINAP in achieving its priority goal of providing useful data with which to analyse patient care.

Acknowledgements

The funding of the British Heart Foundation is gratefully acknowledged. The dedicated assistance of John Birkhead during the early phase of this work is also gratefully acknowledged.