

Data Quality Improvement, Data Linkage and Multiple Imputation in the UK National Vascular Database

Brian A. Cattle¹, Thomas J. Fleming¹, Christopher P. Gale¹, David C. Mitchell², Mark S. Gilthorpe¹, D. Julian A. Scott³, Christopher McCabe⁴, Carolyn Czoski-Murray⁵ & Paul D. Baxter⁶

¹ Centre for Epidemiology and Biostatistics, University of Leeds, Leeds, UK

² Southmead Hospital, Bristol, UK

³ Cardiovascular and Diabetes Research, Leeds Institute of Genetics Health and Therapeutics, University of Leeds, Leeds, UK

⁴ Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada

⁵ Leeds Institute for Health Sciences, Charles Thackrah Building, University of Leeds, Leeds, UK

⁶ Centre for Epidemiology and Biostatistics, University of Leeds, Leeds, UK

Correspondence: Brian A. Cattle, Centre for Epidemiology and Biostatistics, University of Leeds, Leeds LS2 9JT, UK. Tel: 44-113-343-8913. E-mail: b.a.cattle@leeds.ac.uk

Received: May 13, 2012 Accepted: May 30, 2012 Online Published: September 18, 2012

doi:10.5539/ijsp.v1n2p137 URL: <http://dx.doi.org/10.5539/ijsp.v1n2p137>

This work has been supported by the National Institute for Health Research under Programme Development Grant RP-PG-1209-10059

Abstract

The National Vascular Database (NVD) is a prospective audit database collecting information of the quality of care and outcomes of patients admitted to acute hospitals in England, Wales, Scotland and Northern Ireland with several vascular disorders. The NVD has proved to be an important resource for clinical audit but by contrast its potential as a valuable research tool remains under exploited.

We demonstrate proof-of-principle linkage of the NVD to Hospital Episode Statistics (HES) and UK Statistics Authority data. We present and validate Multiple Imputation (MI) methods to address problems with missingness in the linked dataset, focusing on a specific risk model. MI is applied to these linked data to extend the chosen risk model to long term mortality outcomes.

Keywords: case-mix adjustment, data linkage, missing data, multiple imputation by chained equations, The National Vascular Database

1. Introduction

The National Vascular Database (NVD) is a prospective audit database collecting information about the quality of care and outcomes of patients admitted to hospitals in England, Wales, Scotland and Northern Ireland with

- 1) Abdominal Aortic Aneurysms (AAA)-the focus of this paper
- 2) Lower limb ischaemia requiring bypass
- 3) Carotid Endarterectomy
- 4) Amputation.

The NVD has proved to be an important resource for clinical audit (Prytherch et al., 2001; McCollum et al., 1997) by contrast its potential as a valuable research tool remains under exploited. Use of audit data for research is dependent on the ability to adjust for case-mix, which in turn is dependent on the completeness and quality of data collected.

Hospital Episode Statistics (HES) is a data warehouse containing details of all admissions to National Health Service hospitals in England. HES information is stored in separate records: one for each period of care. Each

HES record contains a wide range of information about an individual patient admitted to a hospital. For example, clinical information about operations, patient information such as age and gender, administrative information such as date of admission and geographical information on where the patient was treated. In addition UK Statistics Authority Mortality data are available listing date and cause of death for patients in the HES data. Using HES data in conjunction with the NVD offers the prospect of augmenting the range of variables and outcomes that can be included in case-mix modelling.

Where data are incomplete or missing, patients cannot contribute information to a particular research question. This can lead to gross lack of generalisability and substantial bias in estimates, to the extent that invalid conclusions may be drawn and incorrect policy recommendations made.

A number of methods have been proposed to deal with missing data (e.g. Moons et al., 2006; Arnold & Kronmal, 2003; Donders, 2006), including using a missing category indicator for categorical variables (Vach & Blettner, 1991) and replacing missing values with the last measured value for longitudinal missing data (Carpenter & Kenward, 2008). Many of these approaches are not generally statistically valid, and they can lead to serious bias.

Single imputation of missing values may cause standard errors to be too small since it fails to account for uncertainty about the imputed values. For some time multiple imputation (MI) has been suggested as a promising approach for dealing with missing data (Little & Rubin, 1987), although it is only relatively recently that coherent guidelines for its use have been suggested in the medical literature (Sterne et al., 2009).

MI allows for the uncertainty about the imputed values by creating several plausible imputed datasets and combines the results from each of them. Because we do not know the true values of the missing data the multiple imputation procedure creates multiple copies of the data from the empirical predictive distributions of the observed values. Standard statistical methods such as regression are then used to fit the model of interest in each dataset, with different results because of the variation introduced in the imputation of the missing values. The results are only meaningful when combined to give overall estimated associations. Point estimation and estimates of standard errors are calculated using Rubin's rules (Rubin, 1987) taking into account the between imputation variation. Recent developments in statistical software permit some degree of automation of the process of multiple imputation; see ICE in STATA (Royston, 2004) or MICE in R (van Buuren & Groothuis-Oudshoorn, 2011).

In this proof-of-principle study we focus on the Abdominal Aortic Aneurysms (AAA) component of the National Vascular Database (the NVD AAA data). As the largest component of the database (approximately 60% of patient episodes relate to AAA admissions) this provides a test bed for developing vascular multiple imputation models. Analysis focuses on the Vascular Biochemistry and Haematology Outcome Model (VBHOM) (Tang et al., 2007). This binary logistic regression model of in hospital mortality was built using National Vascular Database items that contained complete data on *Urea, Sodium, Potassium, Haemoglobin, White Cell Count, Age on and Mode of Admission*. As a model originally developed for the NVD VBHOM provides a useful basis on which to build and validate the concept of multiply imputed vascular case-mix models whilst recognising this is only a step towards robust vascular case-mix modelling.

In Section 2 of this paper we describe the degree of missingness present in the NVD and the possibility for HES to help address this missing data. In Section 3 we describe the methods and approach to linking NVD and HES data. In Section 4 we present and validate Multiple Imputation by Chained Equations (MICE) as a technique to address the problems of missing data in the linked HES-NVD data in the context of case-mix adjustment. In Section 5 we apply MICE to the linked HES-NVD data to demonstrate proof-of-principle in modelling long term patient outcomes. Section 6 gives our conclusions in terms of future utility of the NVD and approaches to vascular case-mix modelling.

2. Analysis of Missing Data

2.1 Scope of Missingness

Missingness can affect the validity of a database in a number of different ways. Firstly, all of the data can be missing for a patient, either because a particular hospital does not contribute to the database or because certain patients do not appear in the database even where a hospital is contributing data on some patients. Both cause bias in the representativeness of the database to the patient population. We found that HES contains a higher volume of cases than the NVD, but as the NVD is not yet a complete census of all acute hospitals this was expected.

A second possibility for missingness occurs when a database does not record variables on a patient that have utility in adjusting to case-mix as either a predictor or an outcome measure. In Section 5 of this paper we discuss a

proof-of-principle study for the use of HES data to augment the range of variables available in the NVD.

Recent literature from the USA and Canada (Osborne et al., 2010; Dimick & Upchurch Jr, 2008; Dueck et al., 2004; Vogel et al., 2011) suggests a focus on including more patient demographic information in case-mix modelling such as deprivation/income (Osborne et al., 2010) or distance travelled to treatment (Dueck et al., 2004). Emphasis is also placed on using patient, surgeon and hospital characteristics (Dueck et al., 2004) for case-mix modelling. Co-morbidities also feature strongly in the list of possible predictors (Osborne et al., 2010; Vogel et al., 2011). Although usually based on selective cohorts this suggests important variables may not be adequately represented in the NVD. Evidence from the UK literature (Hadjianastassiou et al., 2007; Jibawi et al., 2006; Prytherch et al., 2005; Sobocinski et al., 2011) is more varied. Work building on data from the NVD (Prytherch et al., 2005) models case-mix largely using patient clinical characteristics, extending or augmenting the VBHOM model (Hadjianastassiou et al., 2007). Previous use of HES data without linkage (Jibawi et al., 2006) focuses on hospital level characteristics without individual level patient characteristics to model. Long term survival outcomes have been studied (Prytherch et al., 2005) but only in the context of trial data for a selected patient group in a single hospital.

The third and most straightforward scope for missingness occurs where, for some of the variables recorded in the database, for some of the patients entered into the database, measurements are missing or implausible. It is this scope of missingness that is described and evaluated for the NVD and can be addressed using multiple imputation.

2.2 Analysis by Variable

An overview of the NVD AAA dataset shows that there is a range of data missingness across variables (median proportion of missing data = 22%, interquartile range = 10% to 64%). Each variable in the NVD is assigned a status, namely *required* or *preferred*. Although many fields are incomplete, often this coincides with variables that are neither described as required nor preferred for entry into the database.

Required variables have much reduced data missingness (typically less than 5%) but constitute only a small proportion (17%) of variables. They are typically related to dates of treatment, admission and discharge.

Around 20% of variables in the database are described as *preferred*. These typically relate to clinical measurements taken on patients (e.g. creatinine biomarker concentration and sodium levels) or drugs taken by patients (e.g. beta blockers and statins) that would form important potential variables in any risk model of mortality. Data missingness amongst these variables ranges from around 5% to 50%.

The remaining variables in the database (63%) have no required or preferred status. These include important co-morbidities such as previous occurrence of stroke, cardiac failure and impaired renal function that would inform a risk model. Over 40% of these variables are recorded with missingness over 50%.

2.3 Geographical Variation in Missingness

Data missingness by hospital is shown in Figure 1. The circles give locations of hospitals contributing to the NVD AAA data base and the radius of the circle is proportional to the percentage of missing entries in the indicated variables. Whilst there is little variation in missingness amongst required fields (median 7%, interquartile range 7% to 8%), there is marked variation across fields not described as required or preferred (median 40%, interquartile range 37% to 48%) and even greater variation amongst preferred fields (median 11%, interquartile range 6% to 22%). Thus any analysis excluding patients with missing data may lead to unacceptable geographical bias.

The variation in data collection and quality between hospitals means that there is a multilevel effect: that is variations at a hospital level may affect inferences made at the patient level. Consequently the imputation scheme must take some account of this, but it is argued that unless clustering is specifically the focus of the analysis model there is little disadvantage in regarding the clusters as a fixed effect in the imputations (van Buuren & Groothuis-Oudshoorn, 2011). Bona-fide methods of multi-level imputation can be found in, for example (Schafer & Yucel, 2002; Yucel, 2008; Goldstein et al., 2009).

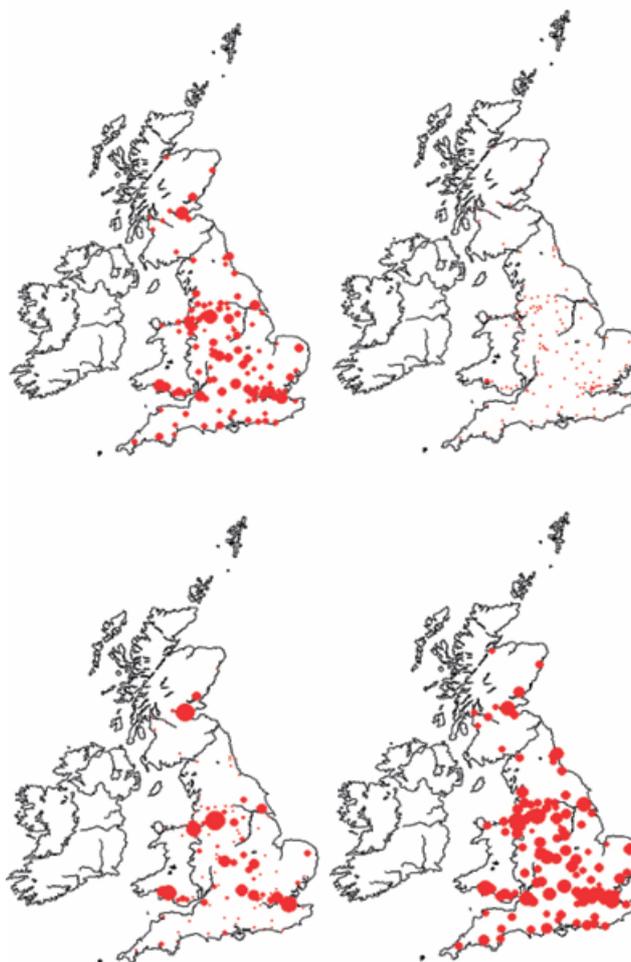


Figure 1. Missingness by hospital and type of variable in the NVD AAA database. The four maps correspond to: % missingness all variables (top left); % missing required variables (top right); % missing preferred variables (bottom left); % missing other variables (bottom right). All plots on the same scale

3. Linking Hospital Episode Statistics Data and the National Vascular Database

3.1 Introduction

HES is the data source for a wide range of healthcare analysis for the UK National Health Service (NHS), Government and many other organisations and individuals. It contains patient care data from 1989 onwards, with more than 12 million new records added each year, and outpatient attendance data from 2003 onwards, with more than 40 million new records added each year.

To identify patients across admissions, deterministic record linkage is used to assign each patient a unique identifier: the *HESID*. The *HESID* can identify patients across years (HES is organised into datasets by financial year) and across record type—inpatient, outpatient, Accident and Emergency, critical care, mortality. The mortality dataset is created by linking mortality data from the UK Statistics Authority (known as ONS because of its former name the Office of National Statistics) to patient information in HES. The HES database captures information on deaths only if it occurred in hospital.

The death record in HES can be analysed using the diagnoses which provide information on the condition or disease at the time of death, but does not provide any information on the actual cause of death. Linking mortality data from the UK Statistics Authority with HES created a richer dataset that captures mortality information for people who died both in and outside of hospital. ONS provides additional information not available in HES such as the ‘underlying cause of death’, which could be used for a wide range of analysis, medical research and healthcare planning. Note that the linked data contains mortality information only on people who have attended

or have been admitted or treated in hospitals. Since our cohort of interest have all undergone a vascular procedure they should all have a corresponding HES record.

Since the consent given by the patients when they were included on the NVD did not allow for sharing of identifiable data with third parties, we had to link the databases using non-identifiable data on a probabilistic basis, matching the details of the operations rather than the patients themselves.

The HES extract was constructed by first filtering the inpatient dataset for any patient who underwent one of the procedures recorded in the NVD (apart from carotid endarterectomy). This created a cohort of patients who were eligible for inclusion into the NVD. This cohort was then used to return all inpatient episodes (1997 to 2010), outpatient episodes (2003 to 2010), mortality records (to date), Accident and Emergency attendances (2007 to 2010) and critical care episodes (2008 to 2010) for those patients. We also applied for and obtained an anonymised extract of the NVD (1995 to 2011) and received the AAA component of the database.

Probabilistic record linkage using the Fellegi-Sunter model (Fellegi & Sunter, 1969) was used to link the inpatient dataset to the NVD AAA database. In most typical record linkage exercises, the entity being linked is a person and we would use identifiers such as their date of birth and name. In this case however, since the data are anonymised, the entities we linked are the operations themselves. The variables used for the linkage were: where the operation took place, the month and year of birth of the patient operated on, sex of the patient operated on, the date of the operation, and the identifying code of the operation.

3.2 Linkage Methodology

For each piece of information, e.g. sex, we define two numbers:

- m : the chance that the sex agrees between the record in the NVD and the correct matching record in HES. This depends on how well sex has been recorded.

- u : the chance that the sex agrees between the record in the NVD and any other record in HES. This depends on how many operations were done on patients of a given sex.

From these we calculate an agreement weight and a disagreement weight for sex. The agreement weight is positive and the larger it is the stronger the certainty of the match. Similarly the disagreement weight is negative, with the larger the absolute value the stronger the certainty. In comparing records from the NVD and HES we assign either an agreement or a disagreement weight as the match weight for each variable we are comparing. If the variable matches between the two records being compared then the agreement weight is used, otherwise the disagreement weight is used. We then repeat this for each variable and sum the individual match weights to give an overall match weight.

The m and u values were defined either globally for a variable (for any value of the variable) or in an outcome-specific manner (different values of the variable had different m and u values).

3.2.1 Global Linkage Parameters

A flat error rate of 5% ($m = 0.95$) was assumed for the data quality and the u value calculated by assuming an even distribution of the range of values of the variable as shown in Table 1.

Table 1. Global linkage parameters for HES and NVD

Variable	m	u	agree	disagree
month of birth	0.95	1/12	3.51	-4.2
day of operation	0.95	1/31	4.88	-4.27
month of operation	0.95	1/12	3.51	-4.2
year of operation	0.95	1/14	3.73	-4.22

3.2.2 Outcome Specific Linkage Parameters

For the remaining variables we used $m = 0.95$ as before and calculated u directly from the distribution of the variable in the HES data. For example we show the values for the variable sex in Table 2.

The values of m and u calculated result in the desired effect—agreement on sex is not particularly discriminatory but disagreement is more so. If two records have the same sex that is just as likely to be random chance, but if the sex is different they are unlikely to represent the same person. Here we can see higher agreement and disagreement

when the sex is female due to the fact that there are twice as many males as females in the HES data. We calculated the m and u values in the same way for year of birth, operation code and place of operation.

Table 2. Outcome specific linkage parameters for HES and NVD

Sex	m	u	agree	disagree
Male	322601	0.644048	0.56	-2.83
Female	178295	0.355952	1.42	-3.69

A number of comapritors were also defined to refine the accuracy of the matching. A date comparator that returns a partial agreement weight was used for operation dates within one week of each other, and partial dates of birth within one month of each other. The date comparator also returns partial agreement weights for common date transcription errors such as getting the day and month back to front. These partial agreement weights are a penalised version of the full agreement weight. The operation codes and locations are also compared for any procedures in the patient records.

3.3 Linkage Results

Following the comparisons each NVD record was assigned the record identifier of the matching HES record: 14,580 of 22,145 NVD records (66%). Where only a probable match had been identified this was used: 2,685 (2,772 matches) of 22,145 records (12%). Where no match or probable match had been found, no HES record identifier could be linked: 4,880 of 22,145 records (22%).

A total of 2,772 probable matches were identified covering 2,685 NVD records. For 82 NVD records (169 matches), more than one matching HES record was identified with equal certainty. No information was available to decide which was the correct record and it was decided to discard these matches. Additionally, 367 HES records (732 matches) matched to more than one NVD record. These are likely to be duplicates in the NVD records but we also decided cautiously to discard these matches. This left 16,451 of the total 24,227 NVD records matched to HES (68%, 74% of the 22,145 records we attempted to link).

3.4 Linkage Quality

This linkage exercise linked only 74% of the expected NVD records to HES. The resulting linked dataset is sufficient to show that, in principle, probalistic linkage of these (and other) datasets is both possible and useful. It shows that where privacy concerns outweigh the perceived benefit of any given research, some progress can still be made to link anonymised routine datasets that are of sufficient size to deliver results despite some unlinked records. Although there may be a systematic bias at work which determines which records cannot be linked (for example if the unlinked records all represent patients with poor outcomes), if sufficient care is taken in the analysis of the linked data, this can be identified and checked to determine if it affects the results.

4. Multiple Imputation and the National Vascular Database

4.1 Missingness Mechanisms

Missingness mechanisms are assumptions about the data describing how we believe the missing (unobserved) data are related to the observed data. The three main mechanisms are as follows:

1) *Missing completely at random* (MCAR): there are no systematic differences between the observed values and the missing values.

2) *Missing at random* (MAR): any systematic difference between the missing values and the observed values can be explained by differences in observed data.

3) *Missing not at random* (MNAR): even after the observed data are taken into account, systematic differences remain between the missing values and the observed values.

Multiple imputation generally assumes that data are MAR: that is the observed variables are predictive of the missing values. Analyses based on multiply imputed data avoid bias only if enough variables that predict the missing values are included in the imputation. Failure to do so may render the MAR assumption implausible and analyses based on the data may be biased.

It is straightforward to demonstrate that the NVD AAA data are not MCAR which precludes a valid complete case analysis. Table 3 summarises the discharge status (dead/alive) comparing missing and complete cases across vari-

ables. The proportion of deaths differs significantly between the missing and complete cases: there are systematic differences between missing and observed values.

Table 3. Discharge status comparing missing and complete data. p -value refers to two-sample test for equality of proportions with continuity correction and false discovery rate correction for multiple testing

Variable	Data complete		Data missing		p -value
	Dead (%)	Alive (%)	Dead (%)	Alive (%)	
Urea	1037 (8.9)	10666 (91.1)	347 (18.6)	1517 (81.4)	<0.001
Sodium	1093 (9.0)	11054 (91.0)	291 (20.5)	1129 (79.5)	<0.001
Potassium	839 (8.3)	9234 (91.7)	545 (15.6)	2949 (84.4)	<0.001
Haemoglobin	1126 (9.2)	11060 (90.8)	258 (18.7)	1123 (81.3)	<0.001
White Cell Count	1060 (8.9)	10893 (91.1)	324 (20.1)	1290 (79.9)	<0.001

4.2 Choice of Variables to Impute

It is vital that there is sufficient plausibility in the MAR assumption to justify the belief that the missing values can be predicted by the observed values. Including as many predictors as possible makes the missing at random assumption more plausible (Jacobusse, 2005) and yields imputations with minimal bias (Van Buuren et al., 1999; Collins et al., 2001), but including more than 15 to 25 predictors gives a negligible increase in the variance explained in the imputations (Schafer, 1997). 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 (Collins, 2001).

We included variables from the VBHOM model and also auxiliary variables to improve prediction of the missing values in the variables of the VBHOM model. When selecting auxiliary variables it's important to use clinical judgment on which variables might usefully predict those that are missing and also statistical judgment to avoid collinearities (i.e. variables contributing no additional information relative to the variables that are already selected) (van Buuren et al., 1999). The auxiliary variables themselves also have missing values that will be imputed.

Table 4. Key variables and summary of missing data. See also <http://www.vascularsociety.org.uk/library/audit.html>

Variable	Type	% missing	Imputation method	Clinically plausible limits
AAA Surgery	Categorical	23.2	Polytomous regression	
Stroke	Categorical	66.7	Polytomous regression	
Mode of admission	Binary	0.8	Polytomous regression	
Gender	Binary	0.0	Polytomous regression	
Diabetes	Binary	5.9	Polytomous regression	
Current Smoker	Binary	21.5	Polytomous regression	
Renal Dialysis	Binary	13.7	Polytomous regression	
Renal Transplant	Binary	15.5	Polytomous regression	
Previous Aortic Stent Surgery	Binary	11.2	Polytomous regression	
Haemorrhage	Binary	10.5	Polytomous regression	
Myocardial Infarction	Binary	62.6	Polytomous regression	
Cardiac Failure	Binary	65.4	Polytomous regression	
Hypotension	Binary	65.4	Polytomous regression	
Discharge Status	Binary	3.2	Predictor only	
Age	Continuous	1.3	Predictive mean matching	(18,100)
Haemoglobin	Continuous	10.5	Predictive mean matching	(2,20)
White Cell Count	Continuous	12.3	Predictive mean matching	(2,50)
Urea	Continuous	14.1	Predictive mean matching	(0.1,800)
Sodium	Continuous	10.9	Predictive mean matching	(105,165)
Potassium	Continuous	26.3	Predictive mean matching	(2,40)
Lowest Systolic BP	Continuous	17.4	Predictive mean matching	(20,250)
Highest Pulse	Continuous	17.4	Predictive mean matching	(20,200)

The subset of variables to be imputed and their missingness characteristics are summarised in Table 4. Note that for continuous variables any data values that lie outside clinically plausible limits have been declared as missing data.

4.3 Choice of Imputation Method

We use predictive mean matching (PMM) to impute continuous variables and polytomous regression for categorical variables. PMM is a general-purpose imputation method (Meng, 1994) that confines the imputations to the observed distribution. PMM preserves non-linear relations between predictors such as that between age and haemoglobin (Figure 2). We used generalised additive models (GAMs) (Little, 1988; Hastie et al., 2008; Wood, 2004) to explore non-linearities. GAMs are a sum of smooth functions able to characterize non-linear regression effects and an advantage is that there is some automation to the fitting of the smooth functions (Hastie et al., 2008; Wood, 2004).

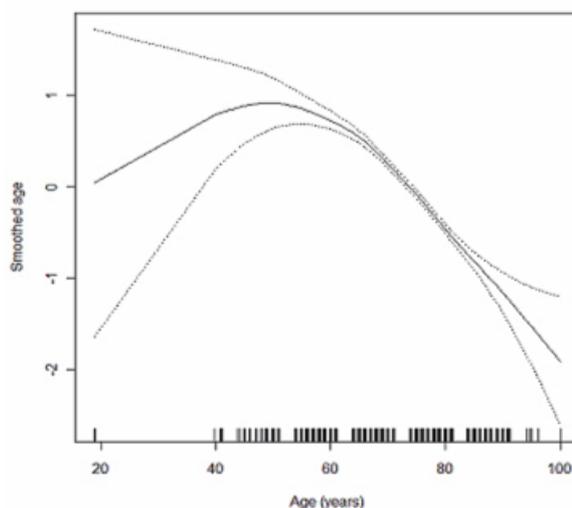


Figure 2. Generalised additive model showing that age has a non-linear association with haemoglobin

A disadvantage of PMM is it may not give sufficient between-imputation variation with only a few predictors (van Buuren & Groothuis-Oudshoorn, 2011). The sample sizes of the NVD AAA data are large with many predictors, so we believe that PMM offers a viable method of imputing continuous variables. To mitigate concerns about insufficient between-imputation variation, we followed the suggestions in Sterne et al. (2009) by using 20 imputations rather than the 'standard' five imputations suggested (Sterne et al., 2009; Jacobusse, 2005). This increases the computational burden for subsequent analyses but represents a balance between computational effort and variation.

This study is proof of concept, but in a definitive study it would be possible to determine more precisely the number of imputations that would be needed. For instance, for each variable included in VBHOM model we might plot coefficients, standardized coefficients (coefficient divided by its standard deviation) and corresponding p -values as function of number of imputations. The stability can be studied and then the choice of the number of imputations discussed, for example by including a threshold for the difference between consecutive p -values in order to conclude in favour of stability.

The imputation scheme assumes normality of the variables being imputed and so we checked that this assumption is satisfied. For variables with a non-normal distribution transformations to approximate normality (e.g. logarithmic transformations) were used. We used a logarithmic transformation for the variables White Cell Count, Urea, Sodium and Potassium, which are sufficiently non-normal to cause concern about the normality assumption.

We must include the outcome variable (e.g. mortality status at discharge or mortality status at one year) as a predictor in the imputations. Not including the outcome dilutes associations between the outcome and the other variables (Wood, 2006).

4.4 Checking and Validation of Imputations

There is no definitive method for checking the imputations or the within imputation iterations of the chained

equations. At each iteration the chain mean and standard deviation can be plotted and on convergence the different streams should freely intermingle with no definite trends. Having checked this for each of our imputations we are satisfied that convergence of the MICE algorithm is satisfactory.

A good imputed value is one that could have been observed had it not been missing. It is desirable that the imputed values be plausible and we checked that there were no implausible values in our imputations. We verified the fit of our imputation models to the observed data as a precursor to imputation and refined our models where non-linearity or non-normality were observed. After transformation we did not observe heteroscedasticity in our imputation models.

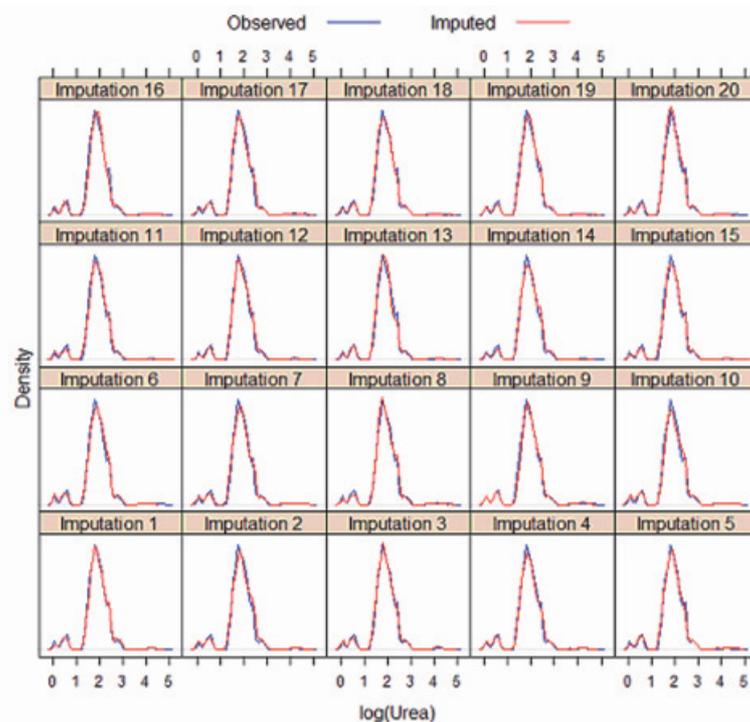


Figure 3. Examples of observed and imputed distributions for log Potassium concentrations

Figure 3 compares the distributions of the observed and imputed values for each completed dataset of log Potassium concentration. Large differences between the observed and imputed distributions could be genuine if the characteristics of patients with missing data were substantially different from those for whom it was recorded, but such differences can indicate problems requiring attention.

The observed and imputed distributions are similar suggesting that the imputations are plausible. The observed and imputed distributions for log Urea shown in Figure 3 demonstrates that the imputations captured the behaviour of the observed distribution in each of the 20 imputed datasets, adding credibility to the MAR assumption. For continuous variables we compared the observed and imputed distributions conditioned on the propensity score (probability of missingness) (Raghunathan & Bondartenko, 2011). We found that the distributions of observed and imputed values were similar, which provides more evidence that the imputations are reasonable. We looked at the residuals of regression of the continuous variables on their propensity score and found that the observed and imputed residual distributions have large overlap giving credibility that the spread of the imputations is appropriate.

5. Multiple Imputation Results for Linked Data

5.1 VBHOM Model

Table 5 shows the performance of the VBHOM model for predicting status at discharge (dead/alive) using only data with complete cases and a full data set with missing values imputed and pooled using the MICE scheme explained above. The magnitudes of the coefficients in the model are broadly similar both with and without imputation. However, notice that, for all of the variables in the VBHOM model, the confidence intervals are narrower for the MICE imputed data. A narrower confidence interval represents reduced uncertainty in the model coefficients and

hence greater confidence in the validity of the model.

There is a large difference in odds ratio between the original and completed data for the variable admission mode. This is because in the original data the number of patients having the surgery as non-elective is very small, meaning that in the completed data, a slight increase in the number of non-elective patients produces a large effect and consequent change in odds ratio.

The Tang analysis is complete case for in hospital death and thus can only be compared with our complete case for in hospital death. They cover a different period of the NVD from 1 January 2002 and 31 January 2004 we cover from 1 Jan 1995 to 31 Sep 2009. In terms of comparison, the point odds ratios reported in the Tang paper are not in the CI reported for out complete case analysis (Table 5) for potassium, sex, urea and white cell count. For potassium and sex the direction of the point estimates differ (i.e. $OR < 1$ or $OR > 1$). For urea and white cell count the direction of the effects agrees. The Tang paper does not report confidence intervals for the point estimates and hence this comparison cannot be made robust.

Table 5. Comparison of VBHOM model complete case analysis and MICE imputation

Variable	Estimate	Unimputed		Estimate	Imputed	
		95% CI	CI width		95% CI	CI width
Gender: Male	0.818	(0.664, 1.008)	0.344	0.914	(0.787, 1.069)	0.282
Admission mode: non-elective	1.693	(0.184, 15.530)	15.345	0.978	(0.184, 4.450)	4.266
Age	1.049	(1.036, 1.061)	0.025	1.051	(1.042, 1.060)	0.018
Urea	1.016	(1.006, 1.026)	0.020	1.010	(1.002, 1.018)	0.016
Sodium	1.001	(0.994, 1.007)	0.013	0.999	(0.994, 1.004)	0.010
Potassium	1.061	(0.992, 1.135)	0.142	1.050	(0.986, 1.117)	0.131
Haemoglobin	0.774	(0.743, 0.806)	0.063	0.740	(0.723, 0.771)	0.047
White Cell Count	1.145	(1.126, 1.165)	0.038	1.119	(1.102, 1.131)	0.030

5.2 One Year Mortality VBHOM Model

The simplest extension to the VBHOM model afforded by the linked data involves replacing the status on discharge (dead/alive) outcome variable with a longer term status variable (e.g. status at one year post surgery: dead/alive). To demonstrate proof-of-principle for mitigating the problems of missing data with a longer term status variable we modify the imputation scheme outlined in Section 4 by replacing status on discharge with status at one year post surgery, identified from the death records provided by the linked HES-NVD data.

Studying Table 6 and comparing complete case analysis with imputed data we observe imputation has reduced the variability of the coefficient estimates and the estimates themselves remain largely unchanged. This is in agreement with the analysis of Section 4 and the literature (Sterne et al., 2009; Cattle et al., 2011), thus giving confidence in the validity of the multiple imputation scheme.

Table 6. Comparison of VBHOM model complete case analysis versus MICE imputation with status at one year outcome

Variable	Estimate	Unimputed		Estimate	Imputed	
		95% CI	CI width		95% CI	CI width
Gender: Male	0.912	(0.762, 1.094)	0.332	0.936	(0.801, 1.094)	0.293
Admission mode: non-elective	3.125	(2.395, 4.053)	1.658	3.888	(3.146, 4.806)	1.66
Age	1.067	(1.056, 1.078)	0.023	1.063	(1.053, 1.072)	0.019
Urea	1.018	(1.008, 1.027)	0.018	1.013	(1.005, 1.021)	0.016
Sodium	0.983	(0.965, 1.001)	0.036	0.981	(0.964, 0.998)	0.034
Potassium	0.996	(0.919, 1.066)	0.147	0.997	(0.931, 1.068)	0.137
Haemoglobin	0.848	(0.819, 0.878)	0.059	0.838	(0.813, 0.864)	0.051
White Cell Count	1.042	(1.025, 1.059)	0.034	1.035	(1.019, 1.051)	0.032

Some changes are observed in the coefficient for admission mode (in line with the discussion of Subsection 5.1). Comparing this long term mortality model with the status on discharge model we observe the direction of effects are in broad agreement with age, urea and white cell count (significant at $p < 0.05$) all being positively associated with odds of death and haemoglobin negatively associated with odds of death ($p < 0.05$).

5.3 Survival Analysis Outcome Model

The availability of linked data (as described in Section 3) not only allows the simple extension to longer term mortality models (such as that in Section 5.2), it also allows full survival models based on data of death (censored to reflect length of follow up). It is therefore important to demonstrate that multiple imputation can successfully be used for survival modelling.

We modify the imputation scheme by including an event indicator variable (whether the patient died in the period of follow up or not) and the log of the survival time instead of the status at discharge (dead/alive) variable. This approach is informed by the literature (Clark & Altman, 2003), though there remains some controversy over whether survival time or the log of survival time should be included in the imputation scheme (van Buuren et al., 1999). We favour inclusion of log survival time to mitigate problems with normality assumptions.

Table 7. Extension of VBHOM model complete case analysis versus MICE imputation to full survival analysis

Variable	Unimputed			Imputed		
	Estimate	SE	<i>p</i> -value	Estimate	SE	<i>p</i> -value
Gender: Male	0.999	0.06	0.991	1.011	0.052	0.839
Admission mode: non-elective	1.709	0.088	0	2.091	0.069	0
Age	1.061	0.003	0	1.059	0.003	0
Urea	1.012	0.002	0	1.007	0.002	0.002
Sodium	0.985	0.006	0.012	0.985	0.006	0.013
Potassium	1.002	0.019	0.902	1.002	0.018	0.915
Haemoglobin	0.889	0.012	0	0.876	0.01	0
White Cell Count	1.028	0.005	0	1.022	0.005	0

Table 7 shows the hazard ratios for a Cox proportional hazards model of survival using VBHOM predictors. Comparing multiple imputation with complete case analysis we note that standard errors have been reduced by imputation and estimates remain largely unchanged (except admission mode, see Section 5.1). This gives confidence in the validity of multiple imputation for full survival modelling. We note that age, urea and white cell count have hazard ratios significantly ($p < 0.05$) in excess of one implying positive association with hazard of death, in agreement with the direction of effects in the status on discharge model. Similarly, haemoglobin has a hazard ratio significantly less than one. However, in addition non-elective admissions have a significant ($p < 0.05$) increased hazard of death and both sodium is negatively associated with hazard of death.

5.4 Discussion of Linked Data Modelling

We have demonstrated the utility of missing data methods in the context of linked data that provide a broader range of potential outcome variables for case-mix adjustment of vascular procedures. We have however, not yet explored the full potential of these additional variables to improve the predictive power of case-mix adjustment models in vascular surgery. For this reason we do not critically interpret the models in terms of effects on survival and mortality as considerable bias may remain in these analyses. The methods evaluated in this report can only mitigate the third scope of missingness (Subsection 2.1). To properly evaluate the potential predictive power afforded by data linkage would require addressing both the first and second scope of missingness, i.e. ensuring all patient eligible to contribute to the database are included and all important predictor variables are measured (where importance is related e.g. to the literature and clinical consensus).

6. Conclusions

There are many ways of handling missing data (Moons et al., 2006; Arnold & Kronmal, 2003; Donders, 2006) although the best solution is to prevent its occurrence. Data imputation techniques allow missing data to be imputed by a value that is predicted using the patient's other known characteristics and have been validated for up to 40% missingness of data. Such techniques have been widely used to handle missingness problems in large-scale censuses and social surveys (Schafer, 2002; McCleary, 2002; Nur et al., 2005), especially in the US, though little in health research (Rubin, 1996). Imputation is generally beneficial because it allows use of information from the incomplete cases that would otherwise have been lost and this is reflected in greater precision in estimation. Therefore an important benefit of our multiply imputed data is an appreciable increase in the number of cases available compared with complete cases analysis.

We sought to demonstrate proof-of-principle that missing data imputation methods can be used (in the context of

the National Vascular Database) to resolve the situation where measurements on some patients are missing for some variables of interest. We further sought to demonstrate that additional variables of interest can be provided by linking the NVD to Hospital Episode Statistics data. We used probabilistic methods to match anonymised HES and NVD records. We showed that a reasonable match rate (74%) was achieved.

We focused on the VBHOM risk model, but we also extended this model to long-term outcomes and survival outcomes in order to further demonstrate the usefulness of multiple imputation for the National Vascular Database.

This work is an important first step towards valid case-mix adjustment models for vascular procedures that can inform service configuration, evaluation and patient choice. However, in interrogating the NVD it has become clear that problems of missingness go beyond those that can be addressed by multiple imputation.

Predictor variables that could inform case-mix models (identified, e.g., from the international literature) are missing from the NVD. Data from some hospitals and some patients within hospitals are also likely missing. To address these issues and make further progress towards valid case-mix modelling requires full re-evaluation of the fields collected in the NVD. This re-evaluation can take advantage of data linkage (demonstrated to be feasible by this work) to reduce the burden on clinical staff in entering data available elsewhere. By reducing the burden on clinical staff it is likely that data quality of the entered variables will improve and any further issues of data missingness can be addressed using multiple imputation as validated by this work.

Although imputation methods are necessarily complex, once implemented on an agreed minimum dataset the imputed (anonymised) data can be made available for analysis by interested parties in the vascular community.

Acknowledgment

This work has been supported by the National Institute for Health Research under Programme Development Grant RP-PG-1209-10059. The extract of the NVD was provided through the Audit and Quality Improvement Committee from the Vascular Society of Great Britain & Ireland. We acknowledge hospitals across the UK for their contribution of data to the NVD. The authors declare that they have no conflict of interest. The study was approved by the UK National Health Service Research Ethics Service. The data supplied is covered by copyright. Copyright (C) 2011, Re-used with the permission of The Health and Social Care Information Centre. All rights reserved. This work remains the sole and exclusive property of The NHS Information Centre for health and social care, and may only be reproduced where there is explicit reference to the ownership of Health and Social Care Information Centre.

References

- Arnold, A. M., & Kronmal, R. A. (2003). Multiple imputation of baseline data in the cardiovascular health study. *American Journal of Epidemiology*, *157*(1), 74-84. <http://dx.doi.org/10.1093/aje/kwf156>
- Aylin, P., Bottle, A., & Majeed, A. (2007). Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. *British Medical Journal*, *334*(7602), 1044-1047. <http://dx.doi.org/10.1136/bmj.39168>
- Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, *160*, 34-45. <http://dx.doi.org/10.1093/aje/kwh175>
- Baxter, P. D., Cattle, B. A., Gale, C. P., Gilthorpe, M. S., & Scott, D. J. A. (2011). Missing data, multiple imputation and the UK National Vascular Database. In *Proceedings of the 26th International Workshop on Statistical Modelling*.
- Carpenter, J. R., & Kenward, M. G. (2008). A critique of common approaches to missing data. In *Missing data in randomised controlled trials: a practical guide*. Birmingham: National Institute for Health Research.
- Cattle, B. A., Baxter, P. D., Greenwood, D. C., Gale, C. P., & West, R. M. (2011). Multiple Imputation for Completion of a National Clinical Audit Dataset. *Statistics In Medicine*, *30*(22), 2736-2753. <http://dx.doi.org/10.1002/sim.4314>
- Clark, T. G., & Altman, D. G. (2003). Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology*, *56*, 28-37.
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(3), 330-351.

<http://dx.doi.org/10.1037//1082-989X.6.4.330>

- Dimick, J. B., & Upchurch, Jr. G. R. (2008). Endovascular technology, hospital volume, and mortality with abdominal aortic aneurysm surgery. *Journal of Vascular Surgery*, 47(6), 1150-1154. <http://dx.doi.org/10.1016/j.jvs.2008.01.054>
- Donders, A. R., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10), 1087-1091. <http://dx.doi.org/10.1016/j.jclinepi.2006.01.014>
- Dueck, A. D., Kucey, D. S., Johnston, K. W., Alter, D., & Laupacis, A. (2004). Long-term survival and temporal trends in patient and surgeon factors after elective and ruptured abdominal aortic aneurysm surgery. *Journal of Vascular Surgery*, 39(6), 1261-1267. <http://dx.doi.org/10.1016/j.jvs.2004.02.021>
- Fellegi, I. P., & Sunter, A. B. (2012). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.
- Fleming, T. J., Scott, D. J. A., Murray, C., Mitchell, D. C., McCabe, C., & Baxter, P. D. (2011). Linking the National Vascular Database (NVD) to Hospital Episodes Statistics (HES) for long term mortality follow up. In *Exploiting Existing Data for Health Research, University of St Andrews, September 2011*.
- Goldstein, H., Carpenter, J. R., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3), 173-197.
- Hadjianastassiou, V. G., Tekkis, P. P., Athanasiou, T., Muktedir, A., Young, J. D., & Hands, L. J. (2007). External validity of a mortality prediction model in patients after open abdominal aortic aneurysm repair using multi-level methodology. *European Journal of Vascular and Endovascular Surgery*, 34(5), 514-521. <http://dx.doi.org/10.1016/j.ejvs.2007.06.017>
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: data mining, inference and prediction*. New York: Springer.
- Jacobusse, G. W. (2005). *WinMICE user's manual*. TNO Quality of Life, Leiden.
- Jibawi, A., Hanafy, M., & Guy, A. (2006). Is there a minimum caseload that achieves acceptable operative mortality in abdominal aortic aneurysm operations? *European Journal of Vascular and Endovascular Surgery*, 32(3), 273-276.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analyses with Missing Data*. New York: Wiley.
- Little, R. J. A. (1988). Missing data adjustments in large surveys (with discussion). *Journal of Business Economics and Statistics*, 6, 287-301.
- McCleary, L. (2002). Using multiple imputation for analysis of incomplete data in clinical research. *Nursing Research*, 51(5), 339-343.
- McCullum, P. T., da Silva, A., Ridler, B. D. M., & de Cossart, L. (1997). Carotid endarterectomy in the UK and Ireland: Audit of 30-day outcome. *European Journal of Vascular and Endovascular Surgery*, 14(5), 386-391.
- Meng, X. L. (1994). Multiple imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538-558.
- Mohammed, M. A., & Stevens, A. (2007). The value of administrative databases. *British Medical Journal*, 334(7602), 1014-1015.
- Moons, K. G., Donders, R. A., Stijnen, T., & Harrel, Jr. F. E. (2006). Using the Outcome for Imputation of Missing predictor Values was Preferred. *Journal of Clinical Epidemiology*, 59, 1092-1101. <http://dx.doi.org/10.1016/j.jclinepi.2006.01.009>
- Nur, U. A. M., Cade, J. E., & Greenwood, D. C. (2005). Dealing with incomplete data in questionnaires of food and alcohol consumption. *Statistics in Transition*, 7(1), 11-34. <http://dx.doi.org/10.1007/s10654-009-9384-1>
- Osborne, N. H., Ko, C. Y., Upchurch, Jr. G. R., & Dimick, J. B. (2010). Evaluating parsimonious risk-adjustment models for comparing hospital outcomes with vascular surgery. *Journal of Vascular Surgery*, 52(2), 400-405. *Paediatric Intensive Care Audit Network National Report 2008-2010*. (published September 2011): Universities of Leeds and Leicester.

- Prytherch, D. R., Ridler, B. M. F., Beard, J. D., & Earnshaw, J. J. (2001). A model for national outcome audit in vascular surgery. *European Journal of Vascular and Endovascular Surgery*, 21(6), 477-483. <http://dx.doi.org/10.1053/ejvs.2001.1369>
- Prytherch, D. R., Ridler, B. M., & Ashley, S. (2005). Audit Research Committee of the Vascular Society of Great Britain and Ireland, Risk-adjusted predictive models of mortality after index arterial operations using a minimal data set. *British Journal of Surgery*, 92(6), 714-718. <http://dx.doi.org/10.1002/bjs.4965>
- Raghunathan, T., & Bondartenko, I. (2011). Diagnostics for Multiple Imputations. Retrieved from <http://ssrn.com/abstract=1031750>
- Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, 4(3), 227-241.
- Rubin, D. (1987). *Multiple imputation for non-response in surveys*. New York: Wiley.
- Rubin, D. B., & Schenker, N. (1989). Multiple Imputation in health-care databases - an overview and some applications. In *3rd Biennial Conference on Methods for Using Large Databases in Health Care Research : Problems and Promises*. John Wiley & Sons Ltd.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall.
- Scafer, J. L., & Yucel, R. M. (2002). Computational Strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11, 437-457.
- Sobocinski, J., Maurel, B., Delsart, P., d'Elia, P., Guillou, M., Maioli, F., ... Haulon, S. (2006). Should we modify our indications after the EVAR-2 trial conclusions? *Annals of Vascular Surgery*, 25(5), 590-597.
- Sterne, J. A. C., White I. R., Barlin J. B., Spratt M., Royston P., Kenward M. G., ... Carpenter J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 338, b2393. <http://dx.doi.org/10.1136/bmj.b2393>
- Tang, T., Walsh, S. R., Prytherch D. R., Lees, T., Varty, K., & Boyle, J. R. (2007). VBHOM, a data economic model for predicting the outcome after open abdominal aortic aneurysm surgery. *British Journal of Surgery*, 94, 717-721. <http://dx.doi.org/10.1002/bjs.5808>
- Vach, W., & Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal Epidemiology*, 134, 895-907.
- van Buuren S., Boshuzien, H. C., & Knook D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6), 681-694.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2012). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1-67.
- Vogel, T. R., Dombrovskiy, V. Y., Graham, A. M., & Lowry, S. F. (2011). The impact of hospital volume on the development of infectious complications after elective abdominal aortic surgery in the Medicare population. *Vascular and Endovascular Surgery*, 45(4), 317-324.
- Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673-686. <http://dx.doi.org/10.1198/016214504000000980>
- Wood S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
- Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A*, 366, 2389-2403. <http://dx.doi.org/10.1214/ss/1177010269>