Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(2) 2014 191-201

Statistical Challenges in Analysing Large Longitudinal Patient-level Data: The Danger of Misleading Clinical Inferences with Imputed Data

Gijo Thomas[1], Kerenaftali Klein[2] and Sanjoy K. Paul[1, 3]
[1]*School of Population Health, University of Queensland, Australia*
[2]*Statistics Unit, QIMR Berghofer Medical Research Institute, Australia*
[3]*Clinical Trials and Biostatistics Unit, QIMR Berghofer Medical Research Institute, Australia*

Received 30 July 2013; Revised 07 April 2014; Accepted 08 April 2014

SUMMARY

Large patient-level longitudinal databases play a crucial role in providing the evidence base for identifying pathways to optimal health outcomes, either informing effective prevention strategies or optimal clinical interventions. However, there are inherent complex challenges for valid statistical analyses of such data for robust assessment of risk factors and health outcomes. The longitudinal data often have a non-trivial amount of missing data with complex missing patterns. Many of the risk factors are also measured with errors. These crucial issues are often ignored in standard analyses, which often lead to biased estimates and misleading clinical or epidemiological inferences. These issues are addressed in this study, along with an empirical assessment of how different imputation techniques for missing data could affect the clinical inferences.

A simulated longitudinal data on systolic blood pressure (SBP) conditional upon the long-term macrovascular events (MVE) were generated following the risk factors' distributions observed in the BP arm of ADVANCE clinical trial. Missing data on longitudinal SBP measures were created following a random missing pattern. The effects of the dynamic changes in SBP over time on the risk of MVE were evaluated using complete as well as multiply imputed missing data sets. The performances of multiple imputations by Multivariate Normal Imputation and Fully Conditional Specification were compared with the analysis of complete data in relation to the consistency of clinical inferences.

The trajectories of longitudinal measures of BP appeared to be significantly different while compared between two sets of multiply imputed data and the original complete data. Although the clinical inferences in relation to the assessment of the effects of higher levels of BP over time on the risk of MVE were not contradictory between complete and imputed data sets, the multiple imputations of missing data could potentially mislead the true trajectory of SBP over time. This exploratory study clearly suggests the need for further methodological assessments of imputation techniques for missing data while dealing with large patient-level longitudinal data.

Keywords: Electronic clinical data, Longitudinal data, Missing data analysis, Multiple imputation, Survival analysis.

1. INTRODUCTION

The recent widespread implementation of electronic recording and nationwide linkage of patients' data from primary care practices and other sources in the developed countries has provided an enormous opportunity for clinical, epidemiological and health policy research. Many large longitudinal databases are currently available, some of which are multi-national, to provide a repository of information on long-term health outcomes and their antecedent factors. These databases create unique opportunities to increase understanding of the causes and mechanisms behind disease onset, disease progression, and mortality. Using

Corresponding author: Sanjoy K. Paul
E-mail address: Sanjoy.paul@qimrberghofer.edu.au

these databases, it is possible to plot the trajectory of disease related risk factors which can be non-linear, their possible interactions over time, and their effects on disease outcomes and mortality.

Unlike cross-sectional or one-off data collections, longitudinal studies are important for policy analysis not only because they document change over time but also because they enable the influence of policies and practice to be isolated from confounding influences such as social background and context. The potential to use such large databases for cohort studies for medical research is great and could unlock important clinical and health policy related answers for a number of conditions such as cardio vascular disease, asthma, diabetes, obesity and osteoporosis. The availability of longitudinal primary care data for millions of patients will also facilitate the investigation of trajectories of risk factors, and particularly modifiable risks, which can lead to improved outcomes. The emphasis on preventative medicine could also have enormous financial benefits with cheaper targeted early treatment avoiding more costly interventions in later life. Currently, these large primary care databases are used for a variety of purposes including clinical research planning, drug utilisation, studies of treatment patterns, clinical epidemiology, drug safety, health outcomes, pharmacoeconomics and health service planning.

Longitudinal primary care, long-term clinical trials and survey-based databases, where data collection takes place for various purposes including particular clinical investigations, have characteristics that require novel data quality management and analysis techniques. Longitudinal data have inherent problems of a large proportion of *random and non-random* missing and erroneous risk factor data (Delaney *et al*. 2008, Greenland and Finkle 1995). Furthermore, patients moving between practices (places) cannot be followed in the anonymised patient databases leading to curtailed observation time periods which may be detrimental to long term survival analyses. A high level of attrition threatens the validity of standard analyses and generalisability. An analysis which does not address these fundamental problems will generate misleading results. Another complex problem is the inappropriate alignment of medication and hospitalisation data, and contradictory information on prescription and other medication related data.

Although primary care data are extensively used by clinical researchers and health policy makers, contradictory findings have been reported in various crucial clinical studies using the same database. For example, various studies have used the same patient cohort from the United Kingdom General Practice Research Database (GPRD) to evaluate the same side effects of specific drugs, often with opposite conclusions. Examples include third generation pills and venous thromboembolism (De Vries *et al*. 2011); proton pump inhibitors and fracture (De Vries *et al*. 2011); and oral bisphosphonates and gastrointestinal cancers (Green *et al*. 2010). These contradictions led to several law suits, in which judges requested reanalysis and urged the statistical community to develop better analytical methods to deal with these complex data (De Vries 2010). In the gastrointestinal cancer study (Green *et al*. 2010), more than 30% of data were missing on key risk factors including smoking status and body mass index (BMI). The researchers used longitudinal risk factors in their analyses without appropriate imputation of missing data and without recognising the issue of erroneous measurements in risk factors. Another study imputed for the risk factors, and reported completely opposite findings (New comb *et al*. 2010).

Although standard statistical techniques and software are available for the analysis of longitudinal data, the methodological challenges threaten the validity of findings from a simple application of existing methods. Recent controversies necessitate the urgent need for novel generalisation of existing statistical methods or development of new methodologies to appropriately analyse large longitudinal databases to provide valid and robust information. Firstly, these databases often have the problems of a non-trivial amount of missing data with complex missing patterns. Multivariate analyses are usually and necessarily applied to a 'complete data set', *i.e*. one which excludes any records with any missing data on covariates or risk factors. This can significantly reduce the power and substantially affect validity of findings. Correction for bias involves an understanding of the complex missing patterns, and incorporation of this in models to produce valid effect estimates. Secondly, many anthropometric, clinical and biochemical risk factors and predictors are measured with error (Tang and Tu 2013, Wulfsohn and Tsiatis 1997, Nahm *et al*. 2008) This crucial issue is ignored in standard analyses, which are well-known to then result in biased estimates and highly likely to generate completely misleading risk estimates based on survival analysis. Thirdly, attrition in longitudinal

studies leads to loss of information about health outcomes, usually on a biased set of individuals (Gustavson *et al*. 2012). Fourthly, these three factors will be expected to co-exist at an individual level. Finally, while some methods (*e.g*. multiple imputations of missing data) have been identified to deal with specific approaches to analysis (Barzi and Woodward 2004, Burton and Altman 2004, Durrant 2005, Engels and Diehr 2003), these have not been generalised to include more general models such as survival models with measurement errors in time-varying covariates, irregularly measured covariates, and non-linearity in trajectories.

In this paper we evaluate various aspects of missing longitudinal risk factor data and the statistical challenges to evaluate the dynamic effects of such risk factor data on long-term health risk. We introduce this in the context of our ongoing extensive clinical studies using large longitudinal patient-level follow-up data to evaluate the dynamics of blood pressure over time and its effects on vascular risks in patients with type 2 diabetes (T2DM). The methodological, clinical and epidemiological aspects discussed in this study are a part of this research programme.

2. CLINICAL CONTEXT

Diabetes is a serious chronic disease that is growing rapidly and that now affects more than 10% of adults in developed countries (Fox *et al*. 2007, Wild *et al*. 2004). Diabetes is associated with a reduced lifespan, mainly because of micro- and macrovascular complications of the disease (Fox *et al*. 2007, Stamler *et al*. 1993). Hypertension is a common comorbidity of diabetes, affecting a significant proportion of patients. The estimated prevalence of hypertension in adults with diabetes is 20–60%, which is 1.5–3 times higher than that in age-matched individuals without diabetes (Lloyd -Jones *et al*. 2010).

Guidelines for treatment of hypertension in patients with T2DM recommend to maintain systolic/ diastolic blood pressure below 140/80 mmHg (American Diabetes Association 2013). Anti-hypertensive medications along with life style modifications are advised in patients with blood pressure above 140/80 mmHg. Exploratory analyses of data from the Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation (ADVANCE) trial reported that additional systolic/ diastolic blood pressure lowering of 5.6/2.2 mmHg was associated with 18% and 14% reduction in cardiovascular death and all-cause mortality in patients with T2DM (Ninomiya *et al*. 2010). However, the analysis of the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial data showed no beneficial effect of tight blood pressure control in patients with T2DM (The ACCORD Study Group 2010). The meta-analysis conducted by McBrien *et al.* (2012) reported no significant association of intensive blood pressure lowering target with mortality in patients with T2DM. A recent systemic review conducted by Lv *et al*. (2012) reported no clear benefit of intensive blood pressure control on mortality. Also, a retrospective cohort study on about 126,000 newly diagnosed patients with T2DM reported that the systolic/diastolic blood pressure below 130/80 mmHg during one year of diagnosis of diabetes was not associated with improved survival (Vamos *et al.* 2012). However, information on the changes in blood pressure over time and its possible effect on mortality are very limited. Only a few studies have examined the blood pressure trajectories in people with diabetes. Post hoc analysis of UKPDS data reported significant risk reduction in mortality by 9–16% for every 10 mmHg decrement in systolic blood pressure level (Stratton *et al.* 2006). This study explored the "updated mean" of blood pressure observed over time, and not the changes in blood pressure over time. The VADT study has recently reported 54% increased risk of cardiovascular events associated with the longitudinal joint effects of high systolic and diastolic blood pressure over a period of 7 years (The VADT Study Group 2010).

3. MISSING DATA ISSUES

Various approaches have been used to deal with missing data in large clinical databases (Barzi and Woodward 2004, Burton and Altman 2004, Durrant 2005, Engles and Diehr 2003). These include complete case analysis (including only patients with complete records), exclusion of variables with incomplete data from the analysis, and including patients with missing information but creating a separate category for missing values (EMA 2010). However, these approaches of dealing with missing data lead to *selection bias*, substantial reduction in the *power* of the study, and the very high potential for misleading conclusions when using these methods is well recognised (Greenland and Finkle 1995). Whichever technique is used, there is a

possibility that false associations may be shown. Several statistical and machine learning methods have been used to deal with the complex problem of missing data (Engels and Diehr 2003, Richman *et al.* 2009, Jerez *et al*. 2010, Farhangfar *et al*. 2008, Gheyas and Smith 2010, Herring and Ibrahim 2001, Kenward and Carpenter 2007, Little 1998, Little and Rubin 2001, Ma *et al*. 2012, Marshall *et al*. 2010). The statistical approaches for handling missing data include deletion, mean substitution, simple regression, regression with an error term, the hot-deck and cold-deck techniques, and the expectation maximization (EM) algorithm. However, most of these approaches of dealing with missing data have the inherent problems of selection bias, wrong or insufficient model specification, and poor theoretical background on their statistical properties – with the potential for drawing misleading conclusions . (Vries 2010, Rubin 1987). A review of the literature reveals that the efficacy of the proposed methods depends strongly on the problem domain (*e.g*., number of cases, number of variables, missingness patterns), and thus there is no clear indication that favours one method over the others (Jerez *et al*. 2010).

3.1 Multiple Imputation of Missing Data under Complex Missing Patterns

There is a strong body of literature on the methodological and application aspects of single as well as multiple imputations (MI) of missing values, covering both frequentist and Bayesian philosophy of statistics (Engels and Diehr 2003, Farhanfar *et al*. 2008, Gheyas and Smith 2010, Herring and Ibrahim 2001, Kenward and Carpenter 2007, Little 1998, Little and Rubin 2001, Ma *et al*. 2012, Marshall *et al*. 2010, Daniels and Hogan 2008). MI of missing data is one of the most advanced, widely applied and powerful techniques for handling missing data (Sterne *et al*. 2009, Klebanoff and Cole 2008), and several researchers have applied MI for missing data to analyse longitudinal data (Hippisley-Cox *et al*. 2007, Weiner *et al*. 2008). The methodological descriptions of techniques for MI are well presented in the existing literature, and are not repeated in this article. The process of MI can be summarised into steps, according to Rubin (2004) (Ma *et al*. 2012, Rubin 1996): (1) replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute; (2) analyse the multiple imputed datasets using complete-data methods; and (3) combine the results from the multiple analyses, which allows uncertainty regarding the imputation to be taken into account.

Despite the increasing use of MI, many aspects of its implementation vary and few publications provide sufficient details on the methods used, underlying assumptions or the extent to which the results can be regarded as more reliable than other approaches (Sterne *et al*. 2009, Klebanoff and Cole 2008). The main obstacle for appropriate implementation of MI in longitudinal data is the limited information available on the extent and mechanisms giving rise to missing data. The data could be both *missing at random* (MAR) as well as *missing not at random* (MNAR). The data are MAR if the probability of the observed missing pattern, given the observed and unobserved data, does not depend on the values of the unobserved data. MNAR occurs if missingness depends not only on the observed data but also on the unobserved (missing) values. In health care data, the MNAR could be because of the health states of the patients and residential changes. For example, mental health data are particularly prone to MNAR - people who have been diagnosed as depressed are less likely than others to report their mental status. Clearly the mean mental status score for the available data will be a biased estimate of the mean that we would have obtained with complete data. One study suggested that blood pressure data were not recorded randomly in the GPRD as subjects with more blood pressure readings tended to have higher recorded values (Delaney *et al*. 2008).

The analysis of missing data under MNAR is a very complex statistical challenge. The only way to obtain an unbiased estimate of parameters of interest is to model the missingness (Dunning and Freedman 2008). Although MI of missing data under MNAR scenario can be theoretically dealt with, this is rarely discussed in the literature and available MI software almost uniformly assumes MAR (Nevalainen *et al*. 2009, White and Carlin 2010). The complications with MNAR occur because of the need to extend the MAR imputation model to include an informative model for dropouts. Kenward *et al*. (2007) first addressed this issue through simple sensitivity analyses (Kenward and Carpenter 2007). They proposed first imputing the missing data under MAR and obtaining parameter estimates for each imputed data set. Then the overall MNAR parameter estimate was a weighted average of these parameter estimates, where the weights depend

on the assumed degree of departure from MAR. Daniels *et al*. (2008) discussed strategies for Bayesian modelling and sensitivity analysis for drawing inference from incomplete data under MNAR (Daniels and Hogan 2008). In some settings (based on simulation data), this approach gave results that closely agree with joint modelling as the number of imputations increases. However, these studies did not address the core issues of developing a theoretical framework and application protocols under MI setup.

3.2 Multiple Imputation by Multivariate Normal Imputation and Fully Conditional Specification

There are a variety of MI models that have been used. Multivariate Normal Imputation (MVNI) assumes that all variables in the imputation model jointly follow a multivariate normal distribution. Implementation uses a Bayesian approach with a Markov chain Monte Carlo (MCMC) algorithm to obtain imputed values from the estimated multivariate normal distribution, allowing appropriately for uncertainty in the estimated model parameters (Schafer and Graham 2002).

The fully conditional approach to imputation (FCS) is a more flexible method that does not rely on the assumption of multivariate normality (Van Buuren *et al*. 1999). Conditional distributions in the regression models are specified for each variable with missing values, conditional on all of the other variables in the imputation model. However, it is possible for some of the conditional distributions to be incompatible with each other, potentially leading to unsound imputations (Van Buuren 2007). Within this setup the predictive mean matching method (PMM) can be used to impute a value randomly from a set of observed values whose predicted values are closest to the predicted value from a specified regression model. This process is straightforward when imputing a continuous random variable. However, PMM approaches led to biased results when applied to missing predictor models (Allison 2000).

4. THE MEASUREMENT ERROR PROBLEM

Longitudinal data, especially the primary care databases, have the inherent problem of erroneous measurements. Measurement errors in the regression covariates bias the estimates of regression slope coefficients towards the null, and can make a true association statistically non-significant (Chesher 1991). Although there are many methods for measurement-error correction, these methods remain rarely used in analysing longitudinal data despite the ubiquity of measurement error. Most clinical and epidemiological studies, using generalised linear models to analyse the longitudinal health care data, fail to address this crucial issue.

4.1 Survival Analysis with Measurement Errors in Risk Factors (Covariates)

Clinical and epidemiological studies with longitudinal data seek to explore the effects of disease risk factors on disease outcome(s). The application of the Cox regression model in these studies is very common. However, while fitting the Cox model with *time-independent* or *time-varying* covariates, the considerations for the adjustments of measurement errors in risk factors are rare. Failure time regression analyses subject to covariate measurement errors or missing covariates has aroused much interest over the past two decades. Several studies have demonstrated the impact of measurement errors by deriving the induced hazard function in the presence of covariate measurement error and advocated several methods to draw inferences (Zhou and Pepe 1995, Zhou and Wang 2000, Liao *et al*. 2011). The most popular approach in this context is the ordinary regression calibration (ORC) approach. Although ORC approach is approximately valid and efficient for measurement error correction of relative risk estimates from the Cox model with *time-independent* risk factors when the disease is rare, it is not adaptable for use with *time-varying* risk factors. As researchers are more interested in exploring risks associated with time-varying risk factors, it is very important to develop an appropriate methodological framework to conduct survival analysis with *time-varying* covariates with measurement errors. There is very limited methodological literature on this challenging issue (Liao *et al*. 2011). Clinical and behavioural risk factor variables are both continuous and semi-continuous. Correction for measurement errors in semi-continuous data is a non-trivial statistical problem. The need to explore the time-varying effects of various risk factors on disease outcomes warrants development of appropriate joint models for survival time and longitudinally observed risk factors with measurement errors. Validation and sensitivity analysis of models with continuous and semi-continuous risk factor data is very important, and will be a novel contribution to the biostatistical literature.

4.2 The Problem of Irregular Risk Factor Measurements

In most longitudinal health care data, clinical and biochemical measurements are not available at equally-spaced periods of time for obvious reasons. The data are recorded when a particular patient actually gets his/her clinical assessment done or blood tested for specific reasons. This creates a potential problem in terms of aligning the data by a pre-specified timeline and then conducting appropriate analysis. For example, the glycated haemoglobin (HbA1c) measure in diabetes patients reflects the accumulation of blood glucose over a period of three months. However, HbA1c measures of individual patients are unlikely to be available every three months or so, introducing complexities in terms of creating specific time-aligned data to explore the trajectory of this crucial risk factor over time and to assess the hazard associated with this factor in terms of cardiovascular disease or mortality.

4.3 Survival Analysis with Irregularly Observed Time-dependent Risk Factors

Most methods described in the literature for joint modelling of irregularly measured longitudinal and event time data are quite complex and do not belong to the standard statistical tools (Griffin *et al*. 2006, Pawitan and Self 1993). It is very important to account for the differences in observation frequency between individual patients, so that the time elapsed since last observation is added to the model. It is highly likely that the interaction effects of this *time elapsed* with time-varying risk factors will have a strong and significant effect on the hazard. Unfortunately clinical and epidemiological studies dealing with this aspect are rare. Such methodological generalisation in the context of Cox regression model will be a novel contribution to the biostatistical literature.

5. DATA

We have generated a simulated data set following the distributions of cardiovascular risk factors and macrovascular event rates as observed in the blood pressure arm of the ADVANCE trial (Ninomiya *et al*. 2010). The risk factors included age and duration of diabetes of patients along with their longitudinal SBP measurements.

Simulation Protocol

Algorithms to generate time-to-event data have been discussed in the literature (Burton *et al*. 2006, Sylvestres and Abrahamowicz 2008). However, in most cases, these algorithms cannot be used to generate events conditional on time varying covariates because this would require inverting the survival function. The algorithm proposed by Sparling *et al*. (2006) generates time-to-event data with time varying covariates for interval censored datasets only (Sparling *et al*. 2006). This poses a great challenge since the methods to analyse such datasets with time varying covariates are not yet generalised and such methods are often computationally intensive or of high dimension due to many nuisance parameters. We have generalised the existing methods to generate continuous time-to-event data along with time-varying covariates to generate 6-monthly blood pressure data conditional upon the observed macrovascular events and the time to macrovascular events.

The continuous time-to-event data for 1000 trials of 20000 patients per simulation were generated based on the event rate, the hazard ratio of treatment and the hazard ratio of individual risk factors, closely matching the data from ADVANCE trial (Ninomiya *et al*. 2010). The event rate for the macrovascular disease was 9.3 % (1860 out of 20000 patients). The events were simulated using exponential distribution. The hazard ratio (HR) for the intensive treatment group versus the standard treatment group was 0.92 for the macrovascular disease. The seven mean 6-monthly longitudinal measures (including data at randomisation) of SBP (mmHg) for standard and intensive treatment arms of the trial were (142, 136, 135, 136, 135, 136, 135) and (145, 137, 133, 134, 134, 135, 134) respectively. The correlation matrices for these variables were generated based on our previous experiences in the same field of study.

Missing Data Creation

In the simulated datasets, we randomly deleted some of the blood pressure measures longitudinally to create an artificial longitudinal dataset with random missing patterns. The baseline data was kept none-

Table 1. The missing proportions and basic statistics of longitudinal SBP (mmHg) for complete and imputed data

SBP	Proportion of missing (%)	Original data	Imputed data	
			FCS	MVNI
At diagnosis#	0	142.13 (141.85, 142.41)	-	-
At 6 month#	5.61	135.60 (135.32, 135.88)	137.64 (137.33, 137.95)	139.25 (138.97, 139.52)
At 12 month#	5.52	134.69 (134.42, 134.97)	136.69 (136.39, 137.00)	138.44 (138.17, 138.71)
At 18 month#	6.01	135.99 (135.70, 136.29)	138.05 (137.72, 138.38)	139.61 (139.31, 139.90)
At 24 month#	5.79	135.39 (135.10, 135.67)	137.44 (137.12, 137.76)	139.02 (138.75, 139.31)
At 30 month#	5.49	136.17 (135.89, 136.46)	138.28 (137.96, 138.60)	139.77 (139.49, 140.06)
At 36 month#	6.07	134.87 (134.57, 135.16)	136.77 (136.44, 137.10)	138.49 (138.20, 138.78)

Mean (95% CI)

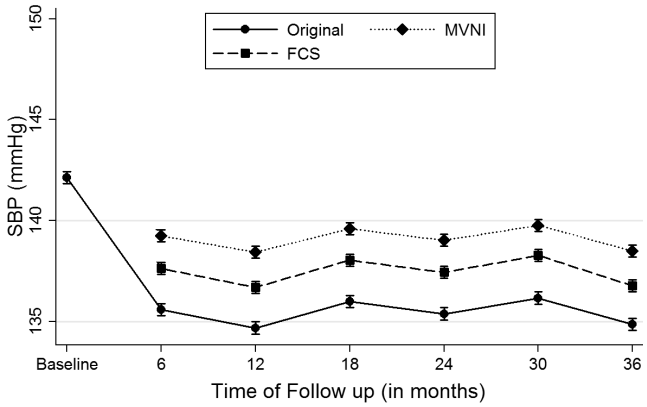

Fig. 1. Trajectories of SBP for the study period from original and imputed datasets (LR, PMM)

missing. We created two sets of multiply imputed data following MVNI and FCS approaches. In both cases 25 multiple imputations under fixed seed number were conducted. The consistency of imputations in terms of the longitudinal distributions of SBP for imputed data and complete data for both approaches were checked.

6. DATA ANALYSIS RESULTS

The primary aims were: (1) to explore the trajectories of SBP with complete data and the imputed data by MVNI and FCS, and (2) to evaluate if the inference related to the association of SBP with the risk of MVE differ between complete and imputed data sets.

Table 2. Hazard ratios and 95% CI associated with continuous and categorised measures of SBP for complete and imputed data. The Akaike Information Cretia (AIC) and the Bayesian Information Criteria (BIC) estimates provides information for comparison of model fits.

	Original data	Imputed data	
		MVNI	FCS
SBP (5 mmHg)	1.034 (1.027, 1.037)	1.032 (1.021, 1.034)	1.028 (1.025, 1.032)
130 ≤ SBP<140	Reference		
140 ≤ SBP<150	1.09 (1.04, 1.17)	1.11 (1.05, 1.16)	1.09 (1.03, 1.16)
SBP ≥ 150	1.18 (1.12, 1.24)	1.23 (1.18, 1.28)	1.23 (1.17, 1.30)
120 ≤ SBP<130	0.94 (0.88, 1.01)	0.93 (0.87, 1.01)	0.99 (0.92, 1.06)
AIC	17142.83	17089.94	17087.47
BIC	17191.20	17138.31	17135.84

The proportions of missing SBP data ranged between 5.6% to 6.1% over 36 months (Table 1). The avaerage level of SBP at the time of randomisation was 142.13 (141.85, 142.41) mmHg. A comparison of the mean (95% CI) of SBP for the complete and imputed datasets are presented in Table 1 and Fig. 1. Clearly, the distributions of SBP were significantly higher longitudinally for both imputation methods, compared to the complete data. This clearly suggests that both methods of imputations failed to capture the true longitudinal distributions of the observed SBP.

The association of different levels of SBP with the macrovascular risk was evaluated using Cox regression model with timevarying risk factors in all the three datasets, (complete, and imputed data sets from MVNI and FCS methods). The SBP was considered as a continuous measure, and the effect of 5mmHg higher SBP over 36 months on the risk of MVE was evaluated (Table 2). For all of the three datasets, a 5 mmHg higher SBP trjectory would increase the risk of vascular event by 3% significantly.

The SBP data was also categorized to draw inferences on the effects of higher and lower levels of SBP on vascular risks. Compared to SBP range of 130-140 mmHg (reference), the risks associated with higher levels of SBP of 140-150 mmHg and >150 mmHg, and lower level of SBP ranging between 120 to 130 mmHg were evaluated. The point estimate of HRs from the imputed data are 5% higher than the 18% observed increased risk in the analysis with complete data. However, the confidence intervals of the risk estimates are overlaping suggesting no statistically significant difference (Table 2).

7. DISCUSSION

Our empirical analysis clearly suggests that both standard methods of multiple imputation, under the assumption of "missing at random", fail to capture the true distribution of the longitudinal measures of risk factors. The distributions of SBP at each time point were approximately normal, and the proportions of missing data were also within 6% only.

Although the inferences related to the effect of high blood pressure level over time on the risk of vascular event was not statistically significantly different for three data sets, a 5% higher risk estimate (HR = 1.23) was observed in the imputed data sets for patients with SBP > 150 mmHg (reference SBP: 130-140 mmHg), compared to those in the complete data. However, clinically, a 5% higher risk estimate over a follow-up period of only 3 to 5 years would be considered as an alarming increased risk. Also, the trajectories of SBP for the imputed data sets were on average 2 to 3 mmHg higher throughout the follow-up, compared to the complete data. In an observational study evaluating the efficacy of antihypertensive medications in relation to tight blood pressure control, such differences could be clinically misleading.

Although MVNI assumes normality of the distributions of study variable, the FCS is more relaxed in terms of distributional assumptions. Under MAR, bias in analyses based on MI may be as big as or bigger than the bias in analyses of complete cases. Unfortunately, it is impossible to determine from the data how large a problem this may be. Currently there is no study available, to the best of our knowledge, which addresses this complex issue. This necessitates a thorough evaluation of the patterns of missingness at least in the key data, and the evaluation of the extent of biases generated by the implementation of existing MI methods to impute for the missing data. This should be followed by novel generalisations of existing MI procedures to deal with missing data under both MAR and MNAR setup.

Another important aspect is the comparison of performances of multiple imputation techniques with other imputation techniques, especially with the established hot-deck and cold-deck methods. Although there are several advantages of the hot-deck or cold-deck method over other statistical approaches for missing data imputations, these methods have not been applied extensively in clinical longitudinal data analysis. In contrast to many parametric statistical approaches, this approach does not rely on model fitting for the variable to be imputed, and hence is potentially less sensitive to model misspecification. The hot-deck imputation can be useful to deal with logically inconsistent data also. For example, in a clinical trial evaluating the efficacy of an anti-hypertensive medication, the diastolic blood pressure of a patient at a particular visit is reported as 200 mmHg; or a 50-year old father reported to have a 45-year old son in a longitudinal survey study – the edit-imputation techniques with the hot-deck method can be used to correct the inconsistent or contradictory values by

deleting these values and imputing these values. Thus, the hot-deck approach can be used simultaneously to take care of measurement error and imputation of missing data. Future studies will include comparisons of MI and machine learning techniques, while evaluating the statistical properties of parameter estimates from the hot-deck approaches.

Longitudinal studies suffer from the problems of attrition, missing data, and irregular and erroneous measurements. The available standard statistical tools are not suitable to address these complex issues simultaneously while exploring the trajectories of key risk data and their association with events or outcomes. Current longitudinal studies often ignore these issues, with the potential of producing misleading results and the subsequent implementation of poorly evidence-based practices. Future research in this line should concentrate on how the current methodologies can be generalised to improve the accuracy and reliability of the analysis of outcomes in longitudinal studies.

REFERENCES

Allison, P.D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociol. Methods Res.,* **28(3)**, 301-309.

American Diabetes Association. (2013). Standards of Medical Care in Diabetes—2013. *Diabetes Care,* **36** (Supplement 1), S11-S66.

Barzi, F. and Woodward, M. (2004). Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *Am. J. Epidemiol.,* **160(1)**, 34-45.

Burton, A. and Altman, D. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br. J. Cancer,* **91(1)**, 4-8.

Burton, A., Altman, D.G., Royston, P. and Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statist. Medicine*, **25(24)**, 4279-4292.

Chesher, A. (1991) The effect of measurement error. *Biometrika*, **78(3)**, 451-462.

Daniels, M.J. and Hogan, J.W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis.* Chapman and Hall, Boca Raton.

De Vries, F. (2010). Two studies, same data source, two answers. *BMJ,* **341**, c5980.

De Vries, F., van Staa, T.P. and Leufkens, H.G. (2011). Proton pump inhibitors, fracture risk and selection bias: three studies, same database, two answers. *Osteoporos Int.,* **22(5)**, 1641-1642.

Delaney, J.A., Moodie, E.E. and Suissa, S. (2008). Validating the effects of drug treatment on blood pressure in the General Practice Research Database. *Pharmacoepidemiol Drug Saf*., **17(6)**, 535-545.

Dunning, T. and Freedman, D. (2008). *Modeling Selection Effects*. Social Science Methodology Sage, London.

Durrant, G. (2005). Imputation methods for handling item-nonresponse in the social sciences: a methodological review, National Centre for Research Methods Working Paper Series.

EMA (2010). Guideline on missing data in confirmatory clinical trials. In: Committee for Medicinal Products for Human Use (CHMP), (ed. Agency EM)

Engels, J.M. and Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *J. Clinical Epidemiol.,* **56(10)**, 968-976.

Farhangfar, A., Kurgan, L. and Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition,* **41(12)**, 3692-3705.

Fox, C.S., Coady, S., Sorlie, P.D., D'Agostino Sr, R.B., Pencina, M.J. and Vasan, R.S., *et al*. (2007). Increasing cardiovascular disease burden due to diabetes mellitus: The Framingham Heart Study. *Circulation*, **115(12)**, 1544-1550.

Gheyas, I.A. and Smith, L.S. (2010). A neural network-based framework for the reconstruction of incomplete data sets. *Neurocomputing*, **73(16-18),** 3039-3065.

Green, J., Czanner, G., Reeves, G., Watson, J., Wise, L. and Beral, V. (2010). Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. *BMJ,* **341**, c444.

Greenland, S. and Finkle, W.D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Amer. J. Epidemiology*, **142(12)**, 1255-1264.

Griffin, J.T., Fraser, C., Gras, L., de Wolf, F. and Ghani, A.C. (2006). The effect on treatment comparisons of different measurement frequencies in human immunodeficiency virus observational databases. *Amer. J. Epidemiol.,* **163(7)**, 676-683.

Gustavson, K., von Soest, T., Karevold, E. and Røysamb, E. (2012). Attrition and generalizability in longitudinal studies: findings from a 15-year population-based study

and a Monte Carlo simulation study. *BMC Public Health,* **12(1)**, 1-11.

Herring, A. and Ibrahim, J. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *J. Amer. Statist. Assoc.*, **96(453)**, 292-302.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M. and Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ*, **335(7611),** 136.

Jerez, J.M., Molina, I., Garcia-Laencina, P.J., Alba, E., Ribelles, N., Martin, M. *et al*. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intell. Medicine*, **50(2)**, 105-115.

Kenward, M. and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statist. Methods Medical Res.*, **16(3)**, 199-218.

Klebanoff, M.A. and Cole, S.R. (2008). Use of multiple imputation in the epidemiologic literature. *Am. J. Epidemiol.*, **168(4)**, 355-357.

Liao, X., Zucker, D.M., Li, Y. and Spiegelman, D. (2011). Survival analysis with error-prone time-varying covariates: A risk set calibration approach. *Biometrics*, **67(1)**, 50-58.

Little, R. (1998). Missing data. *Encyclop. Biostatist.*, 2622-2635.

Little, R.J. and Rubin, D.B. (2001). Statistical data, missing. In: *International Encyclopedia of the Social and Behavioral Sciences*. (eds. Neil JS, Paul BB) Pergamon, Oxford, 15019-15025.

Lloyd-Jones, D., Adams, R.J., Brown, T.M., Carnethon, M. and Dai, S., De Simone G., *et al*. (2010). Executive summary: Heart disease and stroke statistics-2010 update: A report from the american heart association. *Circulation*, **121(7)**, e46-e215.

Lv, J., Neal, B., Ehteshami, P., Ninomiya, T., Woodward, M. and Rodgers, A., *et al*. (2012). Effects of intensive blood pressure lowering on cardiovascular and renal outcomes: A systematic review and meta-analysis. *PLoS Med.*, **9(8)**, e1001293.

Ma, J., Raina, P., Beyene, J. and Thabane, L. (2012). Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *Open Access Medical Statistics*, **2012(2),** 93-103.

Marshall, A., Altman, D. and Holder, R. (2010). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Medical Res. Methodology*, **10(1)**, 112.

McBrien, K., Rabi, D.M. and Campbell, N., Barnieh, L., Clement, F., Hemmelgarn, B.R., Tonelli, M., Leiter, L.A., Klarenbach, S.W., Manns, B.J. (2012). Intensive and standard blood pressure targets in patients with type 2 diabetes mellitus: Systematic review and meta-analysis. *Arch. Intern. Med.,* **172(17),** 1296-1303.

Nahm, M.L., Pieper, C.F. and Cunningham, M.M. (2008). Quantifying data quality for clinical trials using electronic data capture. *PLoS ONE,* **3(8)**, 3049.

Nevalainen, J., Kenward, M.G. and Virtanen, S.M. (2009). Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Statist. Medicine*, **28(29)**, 3657-3669.

Newcomb, P.A., Trentham-Dietz, A. and Hampton, J.M. (2010). Bisphosphonates for osteoporosis treatment are associated with reduced breast cancer risk. *Br. J. Cancer,* **102(5)**, 799-802.

Ninomiya, T., Zoungas, S., Neal, B., Woodward, M., Patel, A. and Perkovic, V., *et al*. (2010). Efficacy and safety of routine blood pressure lowering in older patients with diabetes: results from the ADVANCE trial. *J. Hypertension,* **28(6)**, 1141-1149.

Pawitan, Y. and Self, S. (1993). Modeling disease market processes in AIDS. *J. Amer. Statist. Assoc.*, **88(423)**, 719-726.

Richman, M., Trafalis, T. and Adrianto, I. (2009). Missing data imputation through machine learning algorithms. In: *Artificial Intelligence Methods in the Environmental Sciences*, (eds. Haupt S., Pasini A., Marzban C.), Springer, Netherlands, 153-169.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.*, **91(434)**, 473-489.

Schafer, J. and Graham, J. (2002). Missing data: our view of the state of the art. *Psychological Methods*, **7(2)**, 147-177.

Sparling, Y.H., Younes, N., Lachin, J.M. and Bautista, O.M. (2006). Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics,* Oxford, England, **7(4)**, 599-614.

Stamler, J., Vaccaro, O., Neaton, J.D. and Wentworth, D. (1993). Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the multiple risk factor intervention trial. *Diabetes Care.* **16(2)**, 434-444.

Sterne, J.A., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G. *et al*. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ,* **338**, b2393.

Stratton, I., Cull, C., Adler, A., Matthews, D., Neil, H. and Holman, R. (2006). Additive effects of glycaemia and blood pressure exposure on risk of complications in type 2 diabetes: a prospective observational study (UKPDS 75). *Diabetologia*, **49(8)**, 1761-1769.

Sylvestre, M.P. and Abrahamowicz, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statist. Medicine*, **27(14)**, 2618-2634.

Tang, W. and Tu, X.M. (2013). Modern clinical trial analysis. Springer, New York.

The ACCORD Study Group (2010). Effects of intensive blood-pressure control in type 2 diabetes mellitus. *New England J. Medicine*, **362(17)**, 1575-1585.

The VADT Study Group (2010). Blood pressure and cardiovascular disease risk in the veterans affairs diabetes trial (VADT). *Diabetes Care*, **34(1)**, 34-38.

Vamos, E.P., Harris, M., Millett, C., Pape, U.J., Khunti, K., Curcin, V., *et al*. (2012). Association of systolic and diastolic blood pressure and all cause mortality in people with newly diagnosed type 2 diabetes: retrospective cohort study. *Br. Medical J.,* **345**, e5567.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statist. Methods Medical Res.,* **16(3)**, 219-242.

van Buuren, S., Boshuizen, H.C. and Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statist. Medicine,* **18(6)**, 681-694.

Weiner, M.G., Barnhart, K., Xie, D. and Tannen, R.L. (2008). Hormone therapy and coronary heart disease in young women. *Menopause*, **15(1)**, 86-93.

White, I.R. and Carlin, J.B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statist. Medicine*, **29(28)**, 2920-2931.

Wild, S., Roglic, G., Green, A., Sicree, R. and King, H. (2004). Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care*. **27(5)**, 1047-1053.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53(1)**, 330-339.

Zhou, H. and Pepe, M.S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, **82(1)**, 139-149.

Zhou, H. and Wang, C.Y. (2000). Failure time regression with continuous covariates measured with error. *J. Roy. Statist. Soc., Series B* (Statistical Methodology), **62(4)**, 657-665.