Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose

Nicola Sartoria,*, Alberto Salvanc, Karl Thomasenb,c

aDipartimento di Statistica, Università “Ca’ Foscari” di Venezia, Campiello Sant’Agostin, San Polo 2347, 30125 Venezia, Italy
bIstituto di Analisi dei Sistemi ed Informatica “A. Ruberti”, Consiglio Nazionale delle Ricerche, viale Manzoni 30, 00185 Roma, Italy
cIstituto di Ingegneria Biomedica, Consiglio Nazionale delle Ricerche, Corso Stati Uniti 4, 35127 Padova, Italy

Received 3 May 2004; received in revised form 15 June 2004; accepted 15 June 2004
Available online 8 July 2004

Abstract

Imputation of missing values in a cancer mortality analysis in relation to estimated dose of dioxin for a cohort of chemical workers is considered. In particular, some subjects of the cohort have the body mass index (BMI) missing. This quantity is an essential ingredient for a toxicokinetic model that gives the estimated absorbed dose, which is then used for risk estimation in a proportional hazards model. Imputation of BMI allows to recover information and to use the entire cohort for risk estimation. Both conditional mean imputation and multiple imputation are used. The latter is a simulation-based approach to the analysis of missing data which takes into account the uncertainty of the imputation process using several imputations for each missing value. In the present context, the two imputation methods gave similar results, both correcting for bias (although with some questions) and leading to increased efficiency with respect to the complete-case analysis that simply discards the partially unobserved individuals.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Dioxin; Markov Chain Monte Carlo; Missing value; Multiple imputation; Proportional hazards model; Toxicokinetic model

* Corresponding author. Tel.: +39-0412347432; fax: +39-041710355. E-mail address: sartori@unive.it (N. Sartori).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved. doi:10.1016/j.csda.2004.06.013
1. Introduction

Missing data occur frequently in scientific investigations. There are three major problems that may arise when dealing with incomplete data (Barnard and Meng, 1999). First, there is a loss of information and, as a consequence, a loss of efficiency. Second, there are several complications related to data handling, computation and analysis, due to the irregularities in the data structure and the impossibility of using standard software. Third, and more important, there may be bias due to systematic differences between observed and unobserved data. One approach to incomplete data problems that potentially solves the above issues is multiple imputation (Rubin, 1987; Schafer, 1997). With multiple imputation, unobserved values are replaced by \( m > 1 \) independent draws from an imputation model. Then, each of these completed datasets is analyzed using standard methods for complete data. Finally, the results of the \( m \) analyses are combined to produce a single inferential result that takes into account the uncertainty in the imputed values.

In this paper, we consider an application of multiple imputation to an analysis of cancer mortality in relation to the absorbed dose of dioxin in an occupational cohort of US chemical workers, assembled by the National Institute for Occupational Safety and Health (NIOSH) (Fingerhut et al., 1991). Dioxins are chlorinated aromatic hydrocarbons of which the 2,3,7,8 isomer (tetrachlorodibenzo-p-dioxin or TCDD) is known for its multisite animal carcinogenicity. The International Agency for Research on Cancer (IARC) reviewed the available evidence for carcinogenicity and assigned TCDD to group 1 (human carcinogen) (IARC, 1997). The high exposure levels encountered in some occupational cohorts provide a basis to estimate cancer risk as a function of TCDD dose. The largest of these cohorts was assembled by the US National Institute for Occupational Safety and Health (NIOSH) and consisted of 5172 workers from 12 chemical plants in the United States. Cancer mortality in this cohort was analyzed by other authors in relation to duration of exposure (time spent in TCDD-exposed jobs) (Fingerhut et al., 1991), by quantitative exposure scores (Steenland et al., 1999) and by estimates of TCDD dose obtained by a statistical model of TCDD kinetics (Steenland et al., 2001). Mortality from all cancer types (an outcome predicted by animal experiments) was found to be consistently elevated in relation to TCDD exposure. Our dose–response analysis may be thought of as a two-stage analysis. In the first stage, we use a minimal physiological toxicokinetic model (MPTK) to obtain dose estimates (the time course of the absorbed dose of TCDD) for individual members of the cohort (Thomaseth and Salvan, 1998; Salvan et al., 1999; Bortot et al., 2002). Then, in the second stage, we use a proportional hazards model, treating TCDD dose as a time-dependent explanatory variable. The parameter of interest is the coefficient of TCDD dose, which yields measures of the relative risk associated with different doses (Salvan et al., 2001).

The MPTK model describes long-term kinetics of TCDD in human adults and it depends on several quantities such as anthropometric constants, estimated parameters and a detailed work history in TCDD exposed jobs. Moreover, a required quantity is the body mass index at hire, \( BMI(t_0) \) where \( BMI(t) = \text{bodyweight}(t)/(\text{body height})^2 \) at age \( t \) expressed in kg/m\(^2\) (\( t_0 \) is age at hire). Some subjects of the cohort have \( BMI(t_0) \) missing, while having all remaining information available. Hence it is not possible to estimate the associated TCDD dose for these subjects, and they cannot be used in the risk analysis.
A first solution to the missing-data problem, which is still quite common among practitioners, is to completely discard the subjects with missing values. This complete-case analysis may, however, lead to bias in risk estimates, in particular when the observed units behave differently from the missing ones. Even when this is not the case, there is a loss of information and, consequently, an inflation of standard errors of estimates.

As an alternative to the complete-case analysis, we may impute the missing values, thus recovering some of the remaining information for the partially observed subjects, such as the work history. A simple, and widely used, method is the conditional mean imputation (Little and Rubin, 2002, Chapter 4). Missing values are replaced by the predicted values of a regression model, estimated using the completely observed units. Imputing conditional means may correct for bias in the estimates, although it usually leads to underestimation of standard errors, since we are essentially substituting a missing value with a mean. Moreover, the analysis on the completed dataset does not distinguish between observed and imputed values. This is the main motivation for the use of multiple imputation. As previously described, using several imputed values for the missing data allows us to distinguish between observed and unobserved values and this is directly reflected in the final estimates of standard errors.

An overall outline of our dioxin project is reported in Fig. 1. The project consisted in the following phases: (1) Estimation of the basic kinetic parameters of the MPTK model (background input and liver elimination rate). In this phase we had to resort to using external data because the NIOSH data did not contain the necessary repeated measures of serum TCDD; (2) estimation of an additional parameter representing the occupational exposure rate. Here we used as data a subset of the NIOSH cohort with serum TCDD measurements taken at one point in time; (3) calculation of dioxin dose based on the kinetic model; and (4) analysis of the dependence of cancer mortality on dioxin dose (for all subjects of the NIOSH cohort with available work history). Missing values in BMI(t0) affected phases (2) through (4) of our analysis, and two different main approaches were required.
One task within phase (2) (NIOSH subcohort) was to estimate the function $BMI(t)$, that represents the evolution of $BMI$ over time, a required component of the MPTK model. In addition to $BMI(t_0)$ the subcohort data consisted of an additional $BMI$ value taken at the time of the serum TCDD measurement, usually several decades after hire. Missing values occurred only in $BMI(t_0)$. Because of the longitudinal nature of the $BMI$ data we used a linear mixed effects model (Schafer and Yucel, 2002) for the multiple imputation of $BMI(t_0)$.

Phases (3) and (4) can be viewed as a two-stage process regarding our missing-value problem. First (phase 3), we use the MPTK model to obtain the estimated absorbed dose of dioxin for the entire cohort. Then (phase 4), we use a proportional hazards model to compute risk estimates in relation to dose. Multiple imputation of missing $BMI(t_0)$ within the whole NIOSH cohort ($n = 4935$) is required before running the MPTK model. For each imputed data set we obtain a set of individual doses and one risk estimate. Risk estimates obtained over the imputed sets are finally summarized to give a single inferential statement.

The structure of the paper is as follows. In Section 2 we summarize the main features of multiple imputation. Then, after describing the dose-response analysis based on the MPTK model in Section 3, we apply conditional mean imputation and multiple imputation in Section 4. Results are compared with those obtained with the complete-case analysis and some final comments are given in Section 5.

2. Multiple imputation

Multiple imputation is a simulation-based approach to the analysis of incomplete data and in this section we summarize its main features. For detailed descriptions we refer to Rubin (1987) and Schafer (1997), while Rubin (1996) and Schafer (1999) are recent reviews. For the imputation process we used S-PLUS. However, there are many other software packages for multiple imputation (see Horton and Lipsitz, 2001, for a comparative description).

Multiple imputation is a three-step process. First, $m$ sets of plausible values for the missing observations are generated from an appropriate imputation model, thus giving $m$ completed datasets. In the second step, each completed dataset can be analyzed using standard software. Finally, in the third step, the results of the $m$ analyses are combined in a simple way that accounts for the uncertainty regarding the imputation process.

2.1. Generating the imputed values

Following the notation in Schafer (1997), we denote by $Y$ the $n \times p$ data matrix, which can be thought of as $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where $Y_{\text{obs}}$ and $Y_{\text{mis}}$ are, respectively, the observed and the missing parts. We consider a model $P(Y|\theta)$ for the data $Y$, where $\theta$ is a vector parameter.

In what follows we assume that the missing data are Missing At Random (MAR), which means that the probability that an observation is missing may depend on $Y_{\text{obs}}$, but conditional on $Y_{\text{obs}}$, not on $Y_{\text{mis}}$. This does not mean that the missing data values are a simple random sample of all data values. This latter condition is known as Missing Completely At Random (MCAR) and is a stronger requirement than the MAR assumption. More formally, we can define a matrix $R$ of indicator values, which are 1 if the corresponding element of $Y$ is observed.
and 0 otherwise. Then, the MAR assumption requires that $P(R | Y, \zeta) = P(R | Y_{\text{obs}}, \zeta)$, where $\zeta$ is a vector parameter, while the MCAR assumption has the condition $P(R | Y, \zeta) = P(R | \zeta)$. Since multiple imputation has a Bayesian nature, we also need to assume that the parameter $\theta$ of the model and the parameter $\zeta$ of the missingness mechanism are distinct, which means that they are a priori independent. When both MAR and distinctness hold, the missing-data mechanism is said to be ignorable (Schafer, 1997, Section 2.2). Usually, without any additional information, it is impossible to test the ignorability assumption against a nonignorable alternative. On the other hand, it is sufficient to find a completely observed variable, which behaves differently between observed and nonobserved units, to reject the MCAR assumption.

Assuming ignorability, the $m$ sets of imputed values $Y^{(1)}_{\text{mis}}, Y^{(2)}_{\text{mis}}, \ldots, Y^{(m)}_{\text{mis}}$ are obtained by means of a Bayesian procedure, in the sense that, given a prior distribution for $\theta$, they are independent draws from the posterior predictive density

$$P(Y_{\text{mis}} | Y_{\text{obs}}) = \int P(Y_{\text{mis}} | Y_{\text{obs}}, \theta) P(\theta | Y_{\text{obs}}) \, d\theta.$$ 

From a computational point of view, to obtain the imputed values we use data augmentation (Tanner and Wong, 1987). In practice, it is a Markov Chain Monte Carlo (MCMC) procedure in which, given the values $\theta^{(t)}$ and $Y^{(t)}_{\text{mis}}$ at the $t$th iteration, we update these values by drawing random values from the conditional distributions as follows:

$$Y^{(t+1)}_{\text{mis}} \sim P(Y_{\text{mis}} | Y_{\text{obs}}, \theta^{(t)}),$$  \hspace{1cm} (1)

$$\theta^{(t+1)} \sim P(\theta | Y_{\text{obs}}, Y^{(t+1)}_{\text{mis}}).$$  \hspace{1cm} (2)

Step (1) is said the Imputation step, while (2) is called the Parameter step. Under reasonable conditions, as $t \to \infty$, the sequence $(\theta^{(t)}, Y^{(t)}_{\text{mis}})$, generated by iterations (1)–(2), has a stationary distribution whose marginals are, respectively, $P(\theta | Y_{\text{obs}})$ and $P(Y_{\text{mis}} | Y_{\text{obs}})$. Hence, after convergence, the imputations are obtained from (1).

There are some issues related to a sensible application of data augmentation, which are mainly those related to the general application of MCMC methods. These are not discussed here (see for instance Robert and Casella, 1999). However, it is important to note that a high fraction and a sparse pattern of missing values might slow down the convergence to the stationary distribution. Hence, particular care should be used in such situations. For other practical guidelines, see Section 4.5 of Schafer (1997).

Finally, we note that the choice of the imputation model $P(Y | \theta)$ may affect the results of subsequent analyses. In particular, it is important that the structure of the imputation model encompass that in the analysis model, otherwise the effect of imputation will be to dilute the associations in the analysis model. For a discussion on the role of the imputation model, see Meng (1994), Rubin (1996), Schafer (1997, 1999) and Barnard and Meng (1999).

2.2. Analysis of the completed dataset

Once $m$ imputed sets of values have been obtained, we have $m$ completed dataset $(Y_{\text{obs}}, Y^{(i)}_{\text{mis}}), i = 1, \ldots, m$. We can analyze each of them with a suitably chosen model.
using standard software, as if there were no missing values. For simplicity, we assume that
the interest focuses on a scalar quantity \( Q \) and we denote by \( \hat{Q}(i) \) and \( U(i) \), respectively,
an estimate of \( Q \) and its associated estimated variance, based on the dataset \((Y_{\text{obs}}, Y_{\text{mis}}(i))\), \( i = 1, \ldots, m \).

2.3. Inference based on multiple imputation

The final step consists in combining the results \( \hat{Q}(i) \) and \( U(i) \), \( i = 1, \ldots, m \). Under general conditions (Schafer, 1997, Section 4.3), we have that the multiple imputation estimate of \( Q \) is simply the average of the single estimates

\[
\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}(i).
\]

(3)

The estimate of the variance is

\[
T = \bar{U} + \left(1 + \frac{1}{m}\right)B,
\]

(4)

where \( \bar{U} = m^{-1} \sum_{i=1}^{m} U(i) \) and \( B = (m - 1)^{-1} \sum_{i=1}^{m} (\hat{Q}(i) - \bar{Q})^2 \) are, respectively, the within-imputation and between-imputation variances.

Then, inference about \( Q \) may be based on the approximation

\[
T^{-1/2} (\bar{Q} - Q) \sim t_v, \quad \text{with } v = (m - 1) \left[ 1 + \frac{\bar{U}}{(1 + m^{-1})B} \right]^2.
\]

(5)

Rubin (1987) defines the quantity \( r = (1 + m^{-1})B/\bar{U} \) as the relative increase in variance due to nonresponse. When \( r \) is small and/or \( m \) is large, \( v \) will be large and \( T^{-1/2} (\bar{Q} - Q) \) will be approximately normal. Another useful quantity is \( \hat{\lambda} = [r + 2/(v + 3)]/(r + 1) \), which represents an estimate of the fraction of missing information about \( Q \). In applications, \( r \) and \( \hat{\lambda} \) are useful diagnostics for assessing how the missing data contribute to inferential uncertainty about \( Q \).

The main feature of multiple imputation is that a very small value of \( m \) will usually suffice \((2 \leq m \leq 5)\). There are two fundamental reasons for this. First, simulation is used to solve only the missing-data aspect of the problem. Hence, choosing a very large \( m \) would result in an unimportant gain in efficiency. In particular, the relative efficiency of an estimate based on \( m \) imputations to one based on an infinite number of them is approximately \((1 + \hat{\lambda}/m)^{-1} \), where \( \hat{\lambda} \) is the rate of missing information (Rubin, 1987, p. 114). Hence, unless rates of missingness are very high, there tend to be no real benefit in using more than 5–10 imputations. The second reason why valid inferences can be obtained with very small \( m \) is that the rules for inference from the \( m \) completed data analyses explicitly conditions on the number of imputations, \( m \) (Schafer, 1997, p. 107).
### 3. Interfacing MPTK and proportional hazards models

In this section, we briefly summarize our work on the MPTK model for TCDD and on the cancer mortality analysis for All cancer in the NIOSH cohort, based on TCDD dose estimates obtained from the kinetic model. For more detailed accounts we refer to Thomaseth and Salvan (1998); Salvan et al. (1999, 2001); Bortot et al. (2002).

#### 3.1. MPTK Modelling of TCDD

Our estimates of TCDD dose for individual members of the cohort were based on a minimal physiological toxicokinetic model (MPTK) that describes long-term kinetics of TCDD in human adults. The model is a simplified physiological model that accounts for variations in lipid volume (TCDD binds to lipids) over time (due to individual changes in the lipid mass) and how they affect the predicted serum concentration of TCDD in blood lipids. The model assumes a dynamic equilibrium of TCDD concentration between various body lipid compartments (blood, liver and adipose tissue), with TCDD being eliminated by the liver according to a fixed fractional clearance rate.

The TCDD kinetic model is described by the following dynamic equations:

\[
\frac{dx(t)}{dt} = - \left( k_f \frac{lv_{\text{liver}}(t)}{tlv(t)} + \frac{(dBMI(t)/dt)}{BMI(t)} \right) x(t) + \text{intake}(t), \quad (6)
\]

\[
x(t_0) = C(t_0) \frac{tlv(t_0)}{tlv(t)}, \quad (7)
\]

\[
C(t) = x(t)/tlv(t), \quad (8)
\]

where the time scale \(t\) is the age of the subject, \(x(t)\) (pg/kg) is the average TCDD amount per kg of body weight (BW), \(k_f\) (days\(^{-1}\)) is the liver elimination constant, \(lv_{\text{liver}}(t)\) (g/kg) is the liver lipid volume per kg BW at time \(t\), \(tlv(t)\) (g/kg) is the total lipid volume per kg BW at time \(t\), (7) represents the initial condition at time of hire \(t_0\) with \(C(t_0)\) (ppt) the lipid-adjusted serum TCDD concentration, and (8) representing the predicted lipid-adjusted serum TCDD concentration, \(C(t)\) (ppt), at time \(t\).

Daily intake per kg BW of TCDD is the sum of background exposure, characterized by an input parameter (pg/kg/day), and by a term proportional to the exposure time curve obtained from the individual work history:

\[
\text{intake}(t) = \text{input} + \text{exposure} \ u_{\text{exp}}(t, p_w), \quad (9)
\]

where exposure (pg/kg/day) is the occupational exposure level, which is assumed to be the same for all exposed jobs (for lack of better information at the time of this analysis), and \(u_{\text{exp}}(t, p_w)\) is a piecewise constant exposure function, with values of 1 if the job at time \(t\) was exposed to TCDD and 0 otherwise. The vector \(p_w\) represents the information regarding individual work history.

The last component of the TCDD kinetic model describes the time course of BMI as a function of age

\[
\frac{dBMI(t)}{dt} = \alpha_{BMI} \ t + \beta_{BMI}, \quad (10)
\]
which yields the time course

\[ \text{BMI}(t) = \text{BMI}(t_0) + \frac{\alpha_{\text{BMI}}}{2}(t^2 - t_0^2) + \beta_{\text{BMI}}(t - t_0), \]  

(11)

where \( \text{BMI}(t_0) \) is the individual body mass index measured at the time of hire, and \( \alpha_{\text{BMI}} \) and \( \beta_{\text{BMI}} \) are estimated.

The complete list of model parameters is then: \( \omega = [k_f, \text{input}, \text{exposure}, C(t_0), \text{BMI}(t_0), \alpha_{\text{BMI}}, \beta_{\text{BMI}}] \). For details of estimation procedures we refer to Thomaseth and Salvan (1998); Salvan et al. (1999); Bortot et al. (2002). Here, we discuss only parameters and model-derived quantities that were affected by missing value problems. Parameters \( \alpha_{\text{BMI}} \) and \( \beta_{\text{BMI}} \) were estimated from the subcohort of the NIOSH data consisting of a subgroup of 253 chemical plant workers with two \( \text{BMI} \) measures taken several years apart. The occupational exposure parameter was estimated from the same NIOSH subcohort, for whom a single measure of TCDD was available, usually taken long after termination of employment; \( \text{BMI}(t_0) \) was measured in each individual with known work history \( p_w \). The time course of TCDD concentration in serum lipids, \( C(t|\omega^i) \), for the \( i \)th subject in the cohort was obtained from the MPTK model on the basis of the individual work history. Note that \( \omega^i \) is the value of \( \omega \) for the \( i \)th subject (\( \text{BMI}(t_0) \) is subject dependent). A cumulative exposure index for the \( i \)th subject computed at \( t \) (time at risk), \( D_i(t) \), is computed as a weighted integral of the TCDD plasma concentration profile

\[ D_i(t) = D_i(t; \omega^i) = \int_{t_0}^{t} f(t - u, \pi) C(u|\omega^i) \, du, \]  

(12)

where \( f(t - u, \pi) \) is a weighting function parameterized by \( \pi \), e.g. (i) the unweighted cumulative exposure with \( f(\tau, \pi) = 1 \), and (ii) the lagged cumulative exposure with \( f(\tau, \pi) = 1 \) if \( \tau \geq \pi \), and \( f(\tau, \pi) = 0 \) if \( \tau < \pi \).

3.2. Cancer mortality analysis

The outcome of primary interest for the analysis of cancer mortality in relation to TCDD dose is \textit{All cancer}. We use a proportional hazards regression model with log-transformed TCDD dose

\[ \text{ldose}(t) = \log\{bkgD(t) + D(t)\}, \]  

(13)

where \( bkgD(t) \) is the background dose (the dose that determines a level of 7 ppt observed in the average among unexposed workers) and \( D(t) \) is given by (12). In the model, age is the time scale. Time of entry (left truncation) is age at first exposure. Failure time is age at death for \textit{All cancer}. Censoring time is age at last observation. Risk sets are stratified on birth cohort (with cut points at 1910, 1915, 1920, 1925, 1930). We considered also an exposure lag of 0 and 10 years for the dose metrics. We considered models with just \( \text{ldose}(t) \) as explanatory variable and models with \( \text{ldose}(t) \) and additional categorical covariates: Age at entry (20, 30, 40 years as cut points), Year of entry (1948, 1951, 1954, 1961 as cut points), and Duration of employment (1, 5, 10 as cut points).

The hazard function at time \( t \) in the birth cohort \( s \) is modeled as

\[ \lambda_s(t; \beta) = \lambda_{0s}(t) \exp\{z(t)^T \beta\}, \]
where \( \lambda_{0s}(t) \) is a baseline hazard function for birth cohort \( s \) and \( z(t)^T \) is the transpose of the vector of explanatory variables. Maximum likelihood estimates of the parameter \( \beta \) are obtained maximizing the partial loglikelihood, using the EPICURE\textsuperscript{TM} software package (Preston et al., 1996). The interest is in estimating the coefficient \( \beta \) of \( ldose(T, \omega) \), since the relative risk associated with a dose \( D \), relative to a reference dose \( D_0 \), is given by \( RR = \exp\{\beta \log(D/D_0)\} \). For a detailed report of the results, see Salvan et al. (2001).

4. Imputation of missing values in the NIOSH cohort study

In this section we apply conditional mean imputation and multiple imputation to the two different missing-value problems, both related to missing \( BMI(t_0) \), found in the NIOSH cohort study. First, in Section 4.1, we deal with a subgroup of 253 individuals with repeated measures of \( BMI \) over time (42 with missing \( BMI(t_0) \)). This subcohort is used to estimate some parameters of the MPTK model. In the second part, in Section 4.2, we impute \( BMI(t_0) \) in the whole NIOSH cohort. The NIOSH cohort consists of 5172 subjects. Some observations without a detailed work history were excluded from the analysis, thus leaving 4935 individuals. Furthermore, 886 units have missing \( BMI(t_0) \). For these subjects it is not possible to compute (13) and therefore they cannot be used in the cancer risk analysis. We first impute \( BMI(t_0) \) for these units. Then we compute TCDD dose using the MPTK model and we estimate the associated risk with the proportional hazards model.

4.1. Estimation of MPTK parameters

A subset of 253 individuals of the NIOSH cohort has two measures of \( BMI \) over time, \( BMI(t_j) \), where we denote by \( t_j \) the age of the subject at time \( j \) (\( j = 0, 1 \)). In particular, \( j = 0 \) corresponds to time at hire and \( j = 1 \) corresponds to the moment when the second measure of \( BMI \) was taken (at the time of the serum TCDD measurement for these workers). These subjects were used to estimate function (11), which is a component of the MPTK model. Parameter estimates were obtained through nonlinear least squares, comparing the log-transformed values of \( BMI \) at \( j = 1 \) with those predicted by Eq. (11).

Among the 253 subjects of the subcohort, 42 had missing \( BMI(t_0) \). Hence, they cannot be used to estimate \( z_{BMI} \) and \( \beta_{BMI} \). The estimates of the complete case analysis, i.e. discarding the partially missing individuals, are given in Table 1. However, from Fig. 2 we note that the completely observed subjects \( (R = 1) \) behave differently from those with missing \( BMI(t_0) \) \( (R = 0) \), particularly with age (both \( t_0 \) and \( t_1 \)). Hence, the MCAR assumption does not hold in this case and there could be a potential bias with the complete-case analysis.

As an alternative, we may impute \( BMI(t_0) \) for these 42 subjects, thus recovering their information about \( BMI(t_1) \). A simple solution is the conditional mean imputation. In particular, we estimated a linear model for \( \log\{BMI(t_0)\} \), with predictors \( \log\{BMI(t_1)\} \) and \( t_1 - t_0 \), using the 211 complete observations. Then, the predicted values from this model were taken as the imputed values for the 42 missing subjects. Finally, we estimated the parameters in (11) using the completed dataset. The results are given in Table 1.

As a second choice for the imputation model we used a linear mixed effects model for the logarithm of the body mass index. This formulation takes into account the longitudinal
Table 1
Model for $BMI(t)$: Parameter estimates and standard errors with complete-case (CC), conditional mean with linear model (CML), conditional mean with mixed model (CMM), and multiple imputation (MI) methods, with several values of $m$

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>s.e.</th>
<th>d.f.</th>
<th>$r$</th>
<th>$\hat{\lambda}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{BMI}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>$-4.445 \times 10^{-3}$</td>
<td>$1.06 \times 10^{-3}$</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CML</td>
<td>$-3.745 \times 10^{-3}$</td>
<td>$0.90 \times 10^{-3}$</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CMM</td>
<td>$-3.853 \times 10^{-3}$</td>
<td>$0.88 \times 10^{-3}$</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MI $m = 2$</td>
<td>$-3.699 \times 10^{-3}$</td>
<td>$1.01 \times 10^{-3}$</td>
<td>45</td>
<td>0.1757</td>
<td>0.1851</td>
</tr>
<tr>
<td>MI $m = 3$</td>
<td>$-3.719 \times 10^{-3}$</td>
<td>$0.97 \times 10^{-3}$</td>
<td>369</td>
<td>0.0795</td>
<td>0.0786</td>
</tr>
<tr>
<td>MI $m = 5$</td>
<td>$-3.699 \times 10^{-3}$</td>
<td>$0.99 \times 10^{-3}$</td>
<td>505</td>
<td>0.0977</td>
<td>0.0926</td>
</tr>
<tr>
<td>MI $m = 10$</td>
<td>$-3.688 \times 10^{-3}$</td>
<td>$1.00 \times 10^{-3}$</td>
<td>642</td>
<td>0.1342</td>
<td>0.1211</td>
</tr>
<tr>
<td>$\beta_{BMI}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.30443</td>
<td>0.046</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CML</td>
<td>0.26941</td>
<td>0.040</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CMM</td>
<td>0.27545</td>
<td>0.039</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MI $m = 2$</td>
<td>0.27110</td>
<td>0.046</td>
<td>25</td>
<td>0.2471</td>
<td>0.2545</td>
</tr>
<tr>
<td>MI $m = 3$</td>
<td>0.27303</td>
<td>0.044</td>
<td>180</td>
<td>0.1177</td>
<td>0.1150</td>
</tr>
<tr>
<td>MI $m = 5$</td>
<td>0.26963</td>
<td>0.044</td>
<td>477</td>
<td>0.1008</td>
<td>0.0954</td>
</tr>
<tr>
<td>MI $m = 10$</td>
<td>0.27176</td>
<td>0.044</td>
<td>865</td>
<td>0.1136</td>
<td>0.1041</td>
</tr>
</tbody>
</table>

For multiple imputation, the degrees of freedom (d.f.), the relative increase in variance due to nonresponse ($r$), and the estimate of the fraction of missing information ($\hat{\lambda}$) are also reported.

nature of the data, since we have 253 subjects with two repeated measures. Moreover, this model has also the advantage of using the available observation of the 42 incomplete subjects. The model has the following form, where we denote by $BMI(t_{ij})$ the measure for the $i$th subject at time $j$ ($j = 0, 1$),

$$
\log \{BMI(t_{ij})\} = \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{ij},
$$

(14)

where $b_i \sim N(0, \psi)$ is a random effect and $\epsilon_{ij} \sim N(0, \sigma^2)$. This corresponds to normally distributed $\log\{BMI(t_{ij})\}$, with $E[\log\{BMI(t_{ij})\}] = \beta_0 + \beta_1 t_{ij}$, $\text{Var}[\log\{BMI(t_{ij})\}] = \psi + \sigma^2$ and covariance between $\log\{BMI(t_{i0})\}$ and $\log\{BMI(t_{i1})\}$ equal to $\psi$. Hence, we are assuming independence between subjects and correlation among observations on the same subjects. The estimation method used was restricted maximum likelihood (although maximum likelihood gave essentially the same results) and the predicted values from the model were used as conditional imputed values. With the completed dataset we estimated the parameters in (11) which are given in Table 1.

Both imputation models led to similar results, that differ from those of the complete-case analysis. The inclusion of the 42 partially observed subjects should correct for bias in $z_{BMI}$ and $\beta_{BMI}$. However, mean imputation will usually underestimate standard errors. With multiple imputation we should be able to correct for bias and to avoid underestimation of standard errors as well. Below, we apply multiple imputation in the linear mixed effects model.

Following Schafer and Yucel (2002), multiple imputation for model (14) is performed by means of a Gibbs sampling. In particular, denoting by $\theta^{(t)} = (\beta_0^{(t)}, \beta_1^{(t)}, \sigma^2(t), \psi^{(t)})$ and
Fig. 2. NIOSH subcohort: comparison between completely observed ($R = 1$) and partially observed ($R = 0$) individuals: (a) $t_0$ (Age at first employment); (b) $t_1$ (Age at time of second measurement of TCDD); (c) $BMI(t_1)$.

$Y_{mis}^{(t)}$, respectively, the parameters and the imputed missing values at iteration $t$, steps (1)–(2) becomes

$$b_i^{(t+1)} \sim P(b_i | Y_{obs}, Y_{mis}^{(t)}, \theta^{(t)}), \quad i = 1, \ldots, 253,$$

$$\theta^{(t+1)} \sim P(\theta | Y_{obs}, Y_{mis}^{(t)}, B^{(t+1)}),$$

$$Y_{mis}^{(t+1)} \sim P(Y_{mis} | Y_{obs}, B^{(t+1)}, \theta^{(t+1)}),$$
where \( B = (b_1, \ldots, b_{253}) \) is the vector of random effects. We note that at each iteration it is necessary to obtain also draws of the random effects. This usually slows down the convergence of the algorithm, although this fact depends also on the structure and number of missing values. In the present context, this was not an issue since we had a small number of missing values (from a computational point of view) and only in one variable.

As priors for the parameters, we used improper uniform for \( \beta_1 \) and \( \beta_2 \), while we chose independent inverted gamma for the variance parameters. The latter choice is the one suggested by Schafer and Yucel (2002) to avoid problems with convergence. However, we used small values for the hyperparameters, in order to make the prior densities relatively diffuse. The conditional distributions (15)–(17) are easily obtained and are given in detail in Schafer and Yucel (2002). The imputations have been generated using the S-PLUS library PAN, which is also available for the statistical package R (Ihaka and Gentleman, 1996) and as a stand-alone program (Schafer and Yucel, 2001). Plots of the components of \( \theta^{(t)} \) and of the autocorrelation functions were used to assess convergence. Convergence was achieved after few iterations, and with almost no dependence on starting values. However, since computational time was not an issue, for safety we took each imputed value after 1000 iterations and from parallel chains.

Table 1 show the estimates of \( \alpha_{BMI} \) and \( \beta_{BMI} \) obtained with various numbers of imputations. The results indicate that there is little difference between conditional means imputation and multiple imputation although, as expected, standard errors estimates are larger for multiple imputation. In the present context, \( m = 5 \) is probably large enough, since with \( m \geq 5 \) the results were quite stable. Degrees of freedom of (5) and diagnostic quantities (see Section 2) are also reported.

We note that the estimated function (11) was then used in the MPTK model to estimate the parameter exposure in (9), as explained in detail in Thomaseth and Salvan (1998). Only minor differences in the estimates of exposure were found between the results of the complete-case, the conditional mean imputation and the multiple imputation analyses (not shown). Hence, the recovery of the missing information for this particular task did not appear to be scientifically important. Indeed, also the differences in the estimates of \( \alpha_{BMI} \) and \( \beta_{BMI} \) in Table 1 are not very large, because sensitivity analyses showed that these parameters are not particularly influential for both the long-term time course of TCDD (Salvan et al., 1999) and the final risk estimates (Sartori et al., 2004).

4.2. Cancer risk analysis

We estimate the risk for All cancer associated with TCDD dose using a proportional hazards model. As described in Section 3.2, we used several models. However, since all models showed a similar behaviour, as an illustration we report the results of the model with an exposure lag of 10 years, and TCDD dose with additional covariates as explanatory variables, which are listed in Section 3.2.

If we discard the 886 observations with missing \( BMI(t_0) \) and TCDD dose from the analysis, we obtain the estimate in the first row of Table 2.

Apart from the loss of efficiency due to the elimination of about 18% of the subjects, the results of the complete-case analysis could be biased if the observed individuals
Table 2
Proportional hazards model for the *All Cancer* outcome: Parameter estimates and standard errors for log-dose.
Model with exposure lag of 10 years and additional covariates (see text): complete-case (CC), conditional mean (CM) and multiple imputation (MI) methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Imp. model</th>
<th>log-dose</th>
<th>s.e.</th>
<th>d.f</th>
<th>r</th>
<th>(\hat{\lambda})</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>-</td>
<td>0.1539</td>
<td>0.05879</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CM</td>
<td>(I)</td>
<td>0.1736</td>
<td>0.05421</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MI (m = 2)</td>
<td>(I)</td>
<td>0.1740</td>
<td>0.05424</td>
<td>790924</td>
<td>0.0011</td>
<td>0.0011</td>
</tr>
<tr>
<td>MI (m = 3)</td>
<td>(I)</td>
<td>0.1727</td>
<td>0.05416</td>
<td>322234086</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>MI (m = 5)</td>
<td>(I)</td>
<td>0.1729</td>
<td>0.05416</td>
<td>376277493</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>MI (m = 10)</td>
<td>(I)</td>
<td>0.1731</td>
<td>0.05417</td>
<td>144085176</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td>CM</td>
<td>(II)</td>
<td>0.1751</td>
<td>0.05422</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MI (m = 2)</td>
<td>(II)</td>
<td>0.1736</td>
<td>0.05420</td>
<td>3030595</td>
<td>0.0006</td>
<td>0.0006</td>
</tr>
<tr>
<td>MI (m = 3)</td>
<td>(II)</td>
<td>0.1756</td>
<td>0.05426</td>
<td>674133</td>
<td>0.0017</td>
<td>0.0017</td>
</tr>
<tr>
<td>MI (m = 5)</td>
<td>(II)</td>
<td>0.1755</td>
<td>0.05422</td>
<td>6841569</td>
<td>0.0008</td>
<td>0.0008</td>
</tr>
<tr>
<td>MI (m = 10)</td>
<td>(II)</td>
<td>0.1752</td>
<td>0.05422</td>
<td>10597810</td>
<td>0.0009</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

We used two imputation models: (I) *Age at first employment*, (II) *Age at first employment* and additional variables. For multiple imputation, the degrees of freedom (d.f.), the relative increase in variance due to nonresponse (r), and the estimate of the fraction of missing information (\(\hat{\lambda}\)) are also reported.

Systematically differ from those with missing BMI\((t_0)\). Fig. 3(a) shows that *Age at first employment* \((t_0)\) has a similar distribution among observed and nonobserved individuals. On the other hand, *Duration of employment* (in years) (Fig. 3(b)) has a quite different distribution in the two groups. Moreover, the variable *Duration of exposure*, which is the total amount of time (in years) spent in exposed jobs, also has a behaviour which depends on \(R\). In particular, the mean is 3.032 for the totally observed subjects, while it is 1.989 for those with missing BMI\((t_0)\). The latter comment seems to suggest that the partially unobserved subjects have a different work history than the completely observed ones. Since the work history at the investigated plants is known for all subjects, the imputation of BMI\((t_0)\) would allow us to recover this information and to produce dose estimates for the entire cohort through the MPTK model.

In both conditional mean imputation and multiple imputation we use two imputation models. In the first model (I), we consider only \(\log\) BMI\((t_0)\) and *Age at first employment*, which is the age at which the measure of BMI was (or should have been) taken. The reason for this choice is that it is likely that a physical quantity such as the body mass index should be related to age. Moreover, even if other variables may help to discriminate between observed and nonobserved subjects, such as *Age at last employment* or *Duration of exposure*, it may not make much sense to think that these variables should affect the value of BMI\((t_0)\). However, following statistical considerations (Meng, 1994; Schafer, 1997), in the second model (II) we add other variables, such as those related with the missingness mechanism and also the response variable of the proportional hazards model (*All cancer*). To find the variables related to missingness of BMI\((t_0)\) we used a logistic model for the observed/nonobserved indicator \(R\), which gave the following significant predictors: *Age at first employment*, *Age at last employment*, *Duration of exposure*, *Race* and *Plant*. In particular, *Race* is an indicator variable equal to 1 for white and 0 otherwise, while *Plant* is
Fig. 3. NIOSH cohort: comparison between completely observed \((R=1)\) and partially observed \((R=0)\) individuals: (a) \(t_0\) (Age at first employment); (b) Duration of employment.

an indicator of two groups of plants with different fractions of missing values \((Plant = 1, \text{ for plants } 3 \text{ and } 6; Plant = 0, \text{ for the remaining } 10 \text{ plants})\).

In the conditional mean imputation we used linear regression with response \((\log) BMI(t_0)\) and explanatory variables: Age at first employment (I); Age at first employment, Age at last employment, Duration of exposure, Race, Plant, All cancer (II). We used the predicted values as imputations for missing \(BMI(t_0)\). Then, we ran the MPTK model and obtained the estimated TCCD dose for the entire cohort. Finally, risk estimates were obtained fitting the proportional hazards model. Results are given in Table 2.

Multiple imputations for models (I) and (II) have been computed using a multivariate normal model (Schafer, 1997, Chapters 5–6). We note that the assumption of normality is clearly wrong for the dummy variables Race and Plant. However, this fact has no consequences because these variables are completely observed and they do not need to be imputed.
Fig. 4. Distribution of $BMI(t_0)$ for: the 4049 observed units (left), the 886 CI imputed values (center) and one set of 886 MI imputed values (right).

(see for instance Schafer, 1997, Section 6.3). Steps (1)–(2) become a simple Gibbs sampling, due to the normality assumption. This has been implemented in the library NORM for S-PLUS (Schafer, 1997, Appendix C), which is available also for R. The structure of our problem is rather simple, since we have missing values only in one variable. Using noninformative priors, plots of the components of the parameter and of the autocorrelation functions suggested a very quick convergence of the chains for both imputation models; hence the imputed values were taken after 500 iterations and from parallel chains. Then, for each completed dataset, we obtained estimated TCDD dose with the MPTK model and risk estimates with the proportional hazards model. Finally, the results are summarized using (3) and (4) and are reported in Table 2 for both models and various values of $m$.

Fig. 4 compares the distribution of $BMI(t_0)$ between the observed subjects and those imputed (model (II)), with both conditional mean imputation and multiple imputation (only one representative set is shown). It is clear the difference in variability between the two methods of imputation. However, after the next two phases of the analysis (MPTK and proportional hazards models), this difference largely disappears in risk estimates and standard errors, as indicated by the results in Table 2. This may be explained by the fact that $BMI(t_0)$ displays a decreasing influence on TCDD dose as time (age) increases (Salvan et al., 1999). We also note that there is almost no difference between the results of the two imputation models. This confirms the idea that $BMI(t_0)$ should depend mainly on $t_0$ (age), rather than on the other set of variables, even if these variables are related to the missingness mechanism. Regarding multiple imputation, the degrees of freedom $v$ in (5) were very large in all models considered, thus giving a normal approximation for the standardized estimate. This is a consequence of the homogeneity among imputations. Indeed, $r$ and $\hat{r}$ (see p. 6) were in all cases very small. Hence, as is also clear from the results in Table 2, a very small number of imputations ($3 \leq m \leq 5$) is enough in the present context.
5. Discussion

In the dose–response analysis, we note that the conditional mean and the multiple imputation methods gave almost coincident results, but these are different from those of the complete-case analysis. These results were found in all models considered. The reduction in the standard errors was expected, due to the recovery of the 886 discarded subjects. However, there is an increase in the risk estimate in the analyses including imputed observations. This may suggest that the inclusion of the partially observed subjects has corrected for bias, which is likely to occur when the missingness mechanism is not MCAR (see also Demissie et al., 2003). It is obvious that this assumption does not hold in our problem, as is clear from Fig. 3(b). In particular, the individuals with missing \( BM(t_0) \) have a shorter employment at the study plants and also a shorter work history in exposed jobs. Hence, these subjects have lower estimated doses. On the other hand, a preliminary standardized mortality ratio analysis (Salvan et al., 2001, Table 7) showed no evidence of different mortality between the entire cohort and the partially unobserved cohort. Hence, a similar mortality together with a lower exposure for the partially unobserved individuals might explain the overall increase in risk in the entire cohort. On the other hand, in a risk analysis context, one should consider that workers with a short duration of employment are affected by an additional missing-value problem regarding their work history outside the plants under study. These workers may have had a higher chance than the long-term workers of being exposed to unaccounted agents affecting the outcome under study. Therefore, even if the use of imputation methods allowed us to gain further insight into the data and to recover the work histories of units with missing \( BM(t_0) \), in this particular case there may be reasons to prefer the risk estimates based on the complete data (Salvan et al., 2001). Alternatively, one should extend the missing-value problem to account for missing exposure over the work life, i.e. to include employment outside the plants under study.

As a final remark, we note that other authors have considered multiple imputation in the proportional hazards model (see for instance Paik, 1997; Chen and Little, 1999). However, our approach is different since we are imputing the missing values before applying the first stage model, which is the model that gives the estimated dose, and then we are summarizing the results at the end of the second stage model, which is the proportional hazards models that gives the risk estimate. This is motivated by two reasons. First, \( BM(t_0) \) is only one of the ingredients of the MPTK model. The imputation at this stage allows us to recover other available information about the partially unobserved subjects, such as the complete work history, that would be otherwise discarded. Second, we summarize the results of multiple imputation at the end of the analysis because the quantity of interest is the risk associated with TCDD dose and not the dose itself. Hence, it is the estimate of the risk (and its standard error) which should take into account the uncertainty of the imputation process.

Acknowledgements

This work was partially supported by the National Institute for Occupational Safety and Health, Cincinnati, USA and by a grant from the Italian Ministero del Lavoro e della Previdenza Sociale.
References


