Multiple imputation analysis of case-cohort studies

Helena Marti, Michel Chavance

Multiple imputation analysis of case–cohort studies

Helena MARTI

Biostatistics, CESP Centre de recherche en Epidémiologie et Santé des Populations, U1018, Inserm, F-94807, Villejuif, France
Université Paris Sud, UMRS1018, Villejuif, F-94807, France
helena.marti-soler@inserm.fr

Michel CHAVANCE

Biostatistics, CESP Centre de recherche en Epidémiologie et Santé des Populations, U1018, Inserm, F-94807, Villejuif, France
Université Paris Sud, UMRS1018, Villejuif, F-94807, France
michel.chavance@inserm.fr

Contents

Abstract

The usual methods for analyzing case–cohort studies rely on sometimes not fully efficient weighted estimators. Multiple imputation might be a good alternative because it uses all the data available and approximates the maximum partial likelihood estimator. This method is based on the generation of several plausible complete data sets, taking into account uncertainty about missing values. When the imputation model is correctly defined, the multiple imputation estimator is asymptotically unbiased and its variance is correctly estimated. We show that a correct imputation model must be estimated from the fully observed data (cases and controls), using the case status among the explanatory variable. To validate the approach, we analyzed case–cohort studies first with completely simulated data and then with case–cohort data sampled from two real cohorts. The analyses of simulated data showed that, when the imputation model was correct, the multiple imputation estimator was unbiased and efficient. The observed gain in precision ranged from 8 to 37% for phase-1 variables and from 5 to 19% for the phase 2 variable. When the imputation model was misspecified, the multiple imputation estimator was still more efficient than the weighted estimators but it was also slightly biased. The analyses of case cohort data sampled from complete cohorts showed that even when no strong predictor of the phase-2 variable was available, the multiple imputation was unbiased, as precised as the weighted estimator for the phase2 variable and slightly more precise than the weighted estimators for the phase-1 variables. However the multiple imputation estimator was found to be biased when, because of interaction terms, some coefficients of the imputation model had to be estimated from small samples. Multiple imputation is an efficient technique for analyzing case-cohort data. Practically, we suggest building the analysis model using only the case cohort data and weighted estimators. Multiple imputation can eventually be used to reanalyze the data using the selected model in order to improve the precision of the results.

Keywords: Case–cohort design; Multiple imputation

1 Introduction

Cohort studies, which facilitate causal interpretations, are popular but expensive. Because precision is mainly limited by the number of cases, it is not essential to collect complete information for all the controls. Thus, case–cohort studies and nested case–control studies enable cost reduction with a minimal loss of efficiency [1]. Case–cohort studies were initially proposed by Prentice [2]. When using this approach, the information collected for incompletely observed controls is ignored and inefficient estimators for the effect of phase-1 variables are obtained. Also, the weighted estimators used to analyze case–cohort data are not fully efficient and this could affect the estimate of the effect of phase-2 variable(s). Alternatively, Breslow *et al.* [3] suggested optimizing the sampling weights using all the available data. But case–cohort studies can also be viewed as a particular example of incomplete data, in which the observation process is controlled by the study organizers. Paik and Tsai [4] proposed a simple imputation approach to

model censored observations with missing covariates. In the framework of case–cohort studies, it would imply the simple imputation of the expected value of the phase-2 variable(s) for incomplete controls. However, that approach ignores the uncertainty concerning imputation–model parameters and the values to impute according to a given model.

Multiple imputation is a simple and efficient method for analyzing incomplete observations, while taking into account all the levels of uncertainty regarding missing values. For case–cohort data analysis, the multiple imputation estimator may provide improved precision, compared to weighted estimators, because it integrally uses the available information and approximates the partial likelihood estimator, which can be more efficient than the weighted estimators.

The objective of this study was to establish multiple imputation as an alternative to weighted analysis of case–cohort data. Below, we present the multiple imputation analysis of case–cohort studies and validate this approach by comparing its results to those obtained with weighted estimators. First, we used entirely simulated data. Then, we simulated case–cohort surveys from two cohorts, the Prospective Epidemiological Study of Myocardial Infarction (PRIME) study, in which no strong surrogate of the chosen phase-2 variable was available, and the National Wilms' Tumor Study (NWTS) data, for which a surrogate was available. For simplicity, we only consider time-constant covariates. Extension to time-varying covariates is considered in the discussion.

2 Weighted analysis of case–cohort studies

Case–cohort surveys are examples of two-phase designs. First, the cohort is randomly selected from a general population and the phase-1 information is collected for all the subjects. A subcohort is randomly selected and the entire cohort is followed so as to identify the date of occurrence of the event(s) of interest. Then, the phase-2 information, more expensive, is collected for the subcohort subjects and for all the cases, whether or not they belong to the subcohort. Thus, the phase-2 information is not available for controls not belonging to the subcohort. In cohort surveys, where data are available for the whole cohort, the effect of risks factors on the occurrence of events is generally measured by fitting a proportional hazards model. In case–cohort surveys this model is based on phase-1 and phase-2 variables and the parameters must be estimated from the available incomplete data. In simulations we will consider two phase-1 variables, Z_1 and Z_3, and one phase-2 variable, Z_2.

What is lost in terms of efficiency, when using a case–cohort design rather than a full cohort analysis, can be quantified by the asymptotic relative efficiency (ARE). For a case–cohort design with simple random sampling it was shown to be [5]:

$$\begin{aligned} \text{ARE} &\approx \left\{1 + 2\frac{1-\alpha}{\alpha}\left[1 + \frac{1-d}{d}\log(1-d)\right]\right\}^{-1} \\ &\approx 1 - \gamma, \end{aligned} \tag{1}$$

where α is the proportion of the cohort in the subcohort sample, d is the probability of event occurrence and γ is the fraction of missing information. However, when a phase-1 variable is strongly predictive of the phase-2 variable, stratified sampling of the subcohort can improve efficiency as compared to simple random selection [6].

Weighted estimators of the log-relative risks maximize a weighted pseudo-likelihood ($\widetilde{L}(\beta)$):

$$\widetilde{L}(\beta) = \prod_j \left(\frac{\exp\{\beta' Z_i\} w_{i_j}}{\sum_{k \epsilon \widetilde{C} \cup D} Y_k(t_j) \exp\{\beta' Z_k\} w_{k_j}} \right) \tag{2}$$

where event j occurs at time t_j, $\widetilde{C}$ is the subcohort of size n_{sc}, D is the set of cases, $Y_k(t_j)$ indicates whether subject k is at risk at time t_j, β is the vector of log relative risks, Z_k the vector of covariates for subject k, w_{k_j} the weight of subject k at time t_j, i_j the index of the subject whose event occurs at t_j, and the symbol $'$ denotes transposition. Barlow [7] proposed weighting each complete observation by the inverse of its probability of being included (1 for the cases). Other authors have proposed variable weights, as a function of time, to slightly improve efficiency [6].

The variance of this estimator must take into account the increased uncertainty associated with the randomized selection of the subcohort. This requirement can be achieved by using the sandwich variance, which can be estimated as:

$$Var(\hat{\beta}) = \widehat{I}^{-1} + \frac{n_{sc}(n - n_{sc})}{n} CovD_C \tag{3}$$

where I is the Fisher information matrix, n and n_{sc} are the respective sizes of the cohort and subcohort, and $CovD_C$ is the empirical covariance matrix of *dfbeta* residuals from subcohort members defined as [8]:

$$dfbeta_{ji} = \frac{\beta_j - \beta_{j(i)}}{s_{(i)} \sqrt{(Z'Z)^{-1}_{jj}}} \tag{4}$$

with $\beta_{j(i)}$ the parameter j estimate obtained after deletion of subject i and $s_{(i)}$ the standard error of this estimate.

Stratified sampling of the subcohort considers some information obtained during phase-1, but the information provided by the phase-1 variables is generally ignored in the analysis. Kulich and Lin [9] proposed a family of doubly weighted estimators intended to more efficiently account for the information provided by the initial variables. Qi *et al.* [10] developed nonparametric methods to estimate selection probabilities and nonparametric kernel-smoothing techniques to estimate conditional expectation in fully augmented weighted estimating functions. Breslow *et al.* [3] suggested calibrating or estimating the weights using all the phase-1 information in order to improve precision: 1) with calibration, the weights are

subjected to the constraint that the cohort totals of some auxiliary variables are equal to their weighted sum among all phase-2 subjects. Practically, one builds a prediction model for the phase-2 variable to perform a simple imputation of the predicted values among the controls not belonging to the subcohort, fits the model of interest to the completed data set and uses the influence function from the model of interest to calibrate. Eventually, the model of interest is fitted to the calibrated case-cohort data; 2) with estimation, the weights are the reciprocals of the inclusion probabilities, as estimated from a logistic model fitted to the full cohort.

3 Incomplete observations and multiple imputation

Little and Rubin [11] distinguished three observation processes: data missing completely at random (MCAR), when the probability of incomplete observation is constant; data missing at random (MAR), when this probability depends only on observed values; and data missing not at random (MNAR), when this probability depends on unobserved values. The distinction between MAR and MNAR is of utmost importance, because with the former, unbiased estimators of the parameters of interest are available. Case–cohort data are MAR, because the probability of being completely observed depends only on case status and, under stratified sampling, on some phase-1 variables.

The multiple imputation method developed by Little and Rubin [11] provides an approximation of the maximum likelihood estimator and thus enables the potential selection bias to be corrected. This method relies on the generation of several plausibly completed data sets ($M \geq 2$), accounting for all the levels of uncertainty concerning the missing values. A prediction model must be built, taking into consideration the relationships between the incomplete variable and the other variables, as observed in the complete part of the data. The missing data are not replaced by their expectation but by a value drawn from the distribution posited by the model. To take into account the uncertainty concerning the parameters of the imputation model, several imputations are performed with parameters drawn from the asymptotic distribution of their estimator. An estimate of the parameter of interest, $\hat{\theta}_m$, $m = \{1, \ldots, M\}$, and an estimate of the variance of the estimator, $\widehat{V}(\hat{\theta}_m)$, are obtained from each completed data set. If the imputation model is correct, these estimators are not biased. The multiple imputation estimate, also unbiased, is the mean of these M estimates:

$$\hat{\theta}_{IM} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{5}$$

The multiplicity of imputations enables a correct estimation of the variance of this single estimator. The variance is the sum of two components: the *within-imputations* component (W_{IM}), estimated as the mean of the M asymptotic variances, $\widehat{W}_{IM}$, and the *between-imputations* component (B_{IM}), estimated from the

observed variance of the M estimates, $\widehat{B}_{IM}$:

$$\begin{aligned}\widehat{V}(\hat{\theta}_{IM}) &= \widehat{W}_{IM} + \widehat{B}_{IM} \\ &= \frac{1}{M}\sum_{m=1}^{M}\widehat{V}(\hat{\theta}_m) + (1+M^{-1})\frac{\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{IM})(\hat{\theta}_m - \hat{\theta}_{IM})'}{M-1} \end{aligned} \quad (6)$$

where the factor $1 + M^{-1}$ is an adjustment for using a finite number of imputations [12].

Rubin [13] showed that the relative efficiency of multiple imputations with a finite number M, as compared to an infinite number of imputations, is:

$$\text{ARE} \approx \sqrt{1 + \frac{\gamma}{M}} \quad (7)$$

where γ is the fraction of missing information:

$$\gamma \approx \frac{B_{IM}}{B_{IM} + W_{IM}} \quad (8)$$

A nonexhaustive review of published case–cohort studies showed that, among 25 studies, the fraction of missing information ranged from 0.05 to 0.5, with a median around 0.3. With as much as 40% information missing, M=5 imputations provides an ARE=0.97, and, with 50% missing information, M=10 provides an ARE=0.98. Thus, in most instances, there is not much to gain by using more than 5 or 10 imputations to analyze case–cohort data.

Multiple imputation requests a correct model of the relationships between the incomplete variable(s) and the variables that are linked to the former. When the statistician doing the multiple imputation is independent of the statistician conducting the analyses: "...it is important to include as predictors as many of the variables that are likely to be used in subsequent analyses as possible. Leaving out such variables, even when they are weak predictors, implies that it is known with certainty that they have no relation with the missing values. The result is that correct uncertainty is not reflected [12]." In case–cohort studies, the imputation model and the analysis are in the hands of the same statistician, thus only variables useful for the analysis of interest have to be included.

We need to impute missing phase-2 variable values for the controls who do not belong to the subcohort. This requires an imputation model taking into account the differences between cases and controls: otherwise, the multiple imputation estimator of the effect of the phase-2 variable would be biased. Under the rare disease assumption, it can be shown that a simple generalized linear model using all the case–cohort data and including the status indicator among the explanatory variables has to be considered.

Let us assume that the distribution of Z_2 belongs to the exponential family and depends on a phase-1 variable, possibly multidimensional $\tilde{z}_2$ through:

$$f(z_2 \mid \tilde{z}_2) = \exp\left(\frac{\theta z_2 - b(\theta) + c(z_2)}{a(\phi)}\right), \quad (9)$$

where ϕ is the dispersion parameter and the canonical parameter θ is a linear function of unknown parameters:

$$\theta = \alpha_0 + \alpha_1' \tilde{z}_2. \tag{10}$$

Under the rare disease assumption, the distribution of Z_2 is approximately the same for the whole population and among controls:

$$f(z_2 \mid \tilde{z}_2, \Delta = 0) \simeq f(z_2 \mid \tilde{z}_2), \tag{11}$$

where Δ is the case indicator. Let $\pi(\tilde{z}_2, \mu_2, t_c)$ be the probability of being a case at the end of the observation time, for a subject with $\tilde{Z}_2 = \tilde{z}_2$, $Z_2 = \mu_2 = E[Z_2 \mid \tilde{Z}_2 = \tilde{z}_2]$ and censoring time $T_C = t_c$; let $\pi(\tilde{z}_2, z_2, t_c)$ be the probability of being a case at the end of the observation time, for a subject with $\tilde{Z}_2 = \tilde{z}_2$, $Z_2 = z_2$ and $T_C = t_c$. According to the proportional hazards model and using the rare disease assumption:

$$\pi(\tilde{z}_2, z_2, t_c) = \pi(\tilde{z}_2, \mu_2, t_c)\exp[\beta(z_2 - \mu_2)] \tag{12}$$

We also assume that the distribution of the censoring time is independent of Z_2, so integrating over the censoring time:

$$\pi(\tilde{z}_2, z_2) = \pi(\tilde{z}_2, \mu_2)\exp[\beta(z_2 - \mu_2)] \tag{13}$$

and the distribution of Z_2 conditionally on being a case and on $\tilde{Z}_2$, can be obtained as:

$$\begin{aligned} f(z_2 \mid \Delta = 1, \tilde{z}_2) &= \frac{P[\Delta = 1 \mid \tilde{z}_2, z_2]}{P[\Delta = 1 \mid \tilde{z}_2]} f(z_2 \mid \tilde{z}_2) \\ &= \frac{\pi(\tilde{z}_2, \mu_2)\exp[\beta(z_2 - \mu_2)]}{\int \pi(\tilde{z}_2, \mu_2)\exp[\beta(z - \mu_2)] f(z \mid \tilde{z}_2) dz} f(z_2 \mid \tilde{z}_2) \\ &= \frac{\exp(\beta z_2)}{\int \exp(\beta z) f(z \mid \tilde{z}_2) dz} f(z_2 \mid \tilde{z}_2) \end{aligned} \tag{14}$$

The denominator can be developed as:

$$\begin{aligned} \int \exp(\beta z) f(z \mid \tilde{z}_2) dz &= \int \exp\left(\beta z + \frac{\theta z_2 - b(\theta) + c(z_2)}{a(\phi)}\right) dz \\ &= \int \exp\left(\frac{(\theta + a(\phi)\beta)z - b(\theta + a(\phi)\beta) + b(\theta + a(\phi)\beta) - b(\theta) + c(z)}{a(\phi)}\right) dz \\ &= \exp\left[\frac{b(\theta + a(\phi)\beta) - b(\theta)}{a(\phi)}\right] \end{aligned} \tag{15}$$

with θ given by (10). The results of (15) and (11) lead to:

$$\begin{aligned} f(z_2 \mid \Delta = 1, \tilde{z}_2) &\simeq \exp\left(\frac{a(\phi)\beta z_2 - b(\theta + a(\phi)\beta) + b(\theta)}{a(\phi)}\right) f(z_2 \mid \Delta = 0, \tilde{z}_2) \\ &= \exp\left(\frac{[\theta + a(\phi)\beta]z_2 - b[\theta + a(\phi)\beta] + c(z_2)}{a(\phi)}\right) \end{aligned} \tag{16}$$

Thus, under the rare disease assumption, the distributions of Z_2, given $\Delta = 0$ or $= 1$, differ by a shift of $a(\phi)\beta$ on the scale of the canonical link, i.e., a shift of $\sigma^2\beta$ for a linear model, a shift of β on the logit scale for a logistic model, or a shift of β on the log scale for a log-linear model, where β is the coefficient associated with Z_2 in the proportional hazards model of interest.

The building of the prediction model, particularly the choice of the variable(s) $\tilde{z}_2$, is crucial to perform multiple imputation. Practically, in addition to the status indicator and stratification variables, it is necessary to adjust for the confounding variables included in the Cox model and for other predictive variables, which could be available.

The analyses were performed with R software (version 2.9.0, The R Foundation for Statistical Computing), using the mice (Multivariate Imputation by Chained Equations) package http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm), which generates multiple imputations, and the *survival* package, which enables case–cohort designs to be carried out and analyzed by means of weighted estimators.

4 Validation of the method

To validate the method, first, we used entirely simulated data and compared the estimates and their standard errors to the true values. Then, using the PRIME cohort data, for which no strong surrogate for the chosen phase-2 variable was available, and the NWTS cohort data, which had an available surrogate, we compared the multiple imputation estimator to three weighted estimators: inverse probability weights (the most popular one), calibrated weights and re-estimated weights.

4.1 Simulations

4.1.1 Completely simulated data

Two phase-1 variables were simulated: a binary variable, Z_1, and a Gaussian variable, Z_3, observed in the entire cohort. Also simulated was a phase-2 standard Gaussian variable, Z_2, which was independent of Z_1, but had a correlation of 0.2 with Z_3. The time to the event of interest had an exponential distribution, with $\lambda = \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$. β_1, β_2 and β_3 were fixed at the same value of 0 or $\log(2)$. The censoring time followed a uniform distribution over the interval $[0, \tau]$, where τ was defined so that the probability of the event was approximately 0.01 ($\tau = 0.008$). The cohort size was 25,000 subjects. We also simulated a phase-1 variable predictive of the variable Z_2, $\widetilde{Z}_2 = Z_2 + \varepsilon$ with $\varepsilon \sim N(0, 1)$ independent of Z_2 (the correlation between Z_2 and $\widetilde{Z}_2$ was $\sqrt{2}/2 \simeq 0.7$).

We wanted to estimate the effect of Z_2 on the occurrence of the event, adjusting for Z_3 within the framework of stratified sampling of the subcohort. The cohort was divided into 9 strata based on $\widetilde{Z}_2$ and Z_3 tertiles, and the controls were chosen by stratified sampling. Case–cohort sampling was simulated with 1,000 subjects

in each subcohort (Table 1).

We built a linear prediction model for Z_2 based on the stratum indicators (phase-1 variables) and the status indicator. Z_3 was not directly included in the imputation model to predict Z_2, because it was used to define the strata, included in the model, and weakly correlated with Z_2.

The imputation model was:

$Z_2 = \alpha_0 + \alpha_1 status + \alpha_2' I_{strata} + e$

where I_{strata} is the vector of stratum indicators. The mean multiple R^2 was: 0.40.

Five plausible sets of complete data were generated for each cohort. Using weighted analysis, the variance estimator II proposed by Borgan *et al.* [6] was used. One thousand cohorts were simulated.

To assess the consequences of a misspecification of the imputation model, we compared the multiple imputation estimates, obtained when a predictive variable was omitted or included, to weighted estimates. Two scenarios were considered. In the first, the omitted variable was a confounder. Data were simulated according to the previously used conditions. The cohort was stratified only according to $\widetilde{Z}_2$ tertiles, ignoring Z_3. We used two imputation models, including the status and strata indicators with or without Z_3. The same two models were used in the simple imputation stage of the calibrated weighted analysis. For the second scenario, the omitted variable was not a confounder. Data were still simulated according to the previous conditions. The subcohort was selected by simple random sampling so that the imputation model included no strata indicator; $\widetilde{Z}_2$, which was related to the phase-2 variable (correlation $\rho = 0.7$), was omitted or included in the model.

4.1.2 Results

In Table 2, for phase-1 (Z_1 and Z_3) and phase-2 (Z_2) variables, all estimates of the log relative risks were unbiased. Likewise, the multiple imputation variance and Borgan's variance estimator II (BII) agreed with the observed dispersion of estimates. For phase-1 variables, this observed dispersion was close to those of the entire cohort and multiple imputation but larger with the weighted estimator. For the estimated effect of the phase-2 variable, the observed standard errors were obviously larger for case–cohort than full cohort analyses, but they were slightly smaller with multiple imputation than with weighted estimators. For phase-1 variables, the relative increases of standard errors for the weighted analysis estimators compared to those of multiple imputation, ranged from 8 to 37%, whereas they were slightly smaller (5 or 19%) for phase-2 variables.

Table 3, gives the results obtained with the calibrated weighted estimator and the multiple imputation estimator using a correct imputation model and a misspecified imputation model omitting a confounder variable. All calibrated weighted estimates were unbiased, for phase-1 and phase-2 variables, using a correct or a misspecified simple imputation model. All multiple imputation estimates were unbiased when a correct imputation model was used. However, the misspecified model yielded biased multiple imputation estimates for the phase-2 variable and

the omitted phase-1 variable. All multiple imputation estimates were more precise than calibrated weighted estimates, for phase-1 and phase-2 variables effects. The relative increase of the standard error of the calibrated weighted estimate, as compared to multiple imputation estimate, exceeded 10%.

Table 4 reports the results obtained with the weighted estimator, a correct imputation model and a misspecified imputation model omitting a predictive variable that was not a confounder. All the estimates of phase-1 variable effects were unbiased. The precision of the two multiple imputation estimates, obtained with the correct or the misspecified models, were similar. The weighted estimate was less precise than multiple imputation estimates, more specifically, the relative increase of the standard error for the weighted estimator compared to multiple imputation estimators was 20% for Z_1 and over 46% for Z_3.

For the phase-2 variable, the multiple imputation estimate with the correct imputation model was unbiased, while the misspecified imputation model and the weighted analysis led to estimates slightly biased in opposite directions, respectively, 2.3% and 1.9%. As expected, ignoring Z_3 in the imputation model decreased the precision of the multiple imputation estimator. The relative increase of the standard error with the misspecified model compared to that with the correct model was 7.6%. The standard error of the weighted estimator was larger than that of the multiple imputation estimator. The relative increase of the standard error of the weighted estimator, compared to multiple imputation estimator, was 30% with the correct model and 21% with the misspecified model.

4.2 PRIME data

4.2.1 Description of the data

The PRIME survey [14] was a multicenter cohort investigation studying risk factors of ischemic heart disease (IHD) and other cardiovascular end points. Among the 9,520 male subjects, 642 (6.7%) experienced a cardiovascular event. The median follow-up time was 10 years and for those who suffered an event, the median time to its occurrence was 5 years.

We chose to estimate the effect of fibrinogen (phase-2 variable) on the occurrence of the event, adjusting for age, center, total cholesterol (CH), high-density lipoprotein cholesterol (HDL), systolic blood pressure (SBP) and tobacco use. Case–cohort data were simulated based on the entire PRIME cohort, and the results obtained with the different estimators were compared to each other and to the results obtained from the full cohort. The validation procedure used 1,000 simulated subcohorts of size 2,100.

4.2.2 Sampling of the subcohort

Fibrinogen is a protein whose circulating concentration increases during inflammatory conditions. It also plays an important role in normal and pathological blood coagulation, and elevated fibrinogen levels are associated with a higher risk of cardiovascular events. Because smoking is one of the main factors determining

the fibrinogen level [15], the cohort was stratified according to tobacco use, treated as a phase-1 variable available for the entire cohort. The strata were created as follows: stratum 1: non-smokers, stratum 2: ex-smokers, stratum 3: smokers [1–9] g/day, stratum 4: smokers [10–19] g/day, stratum 5: smokers $\geq$ 20 g/day. Stratified sampling was used to select subcohorts and the number of events in each stratum is given in Table 5. The sampling probabilities were approximately 0.16, 0.20, 0.21, 0.25 and 0.46, respectively, for strata 1–5.

The linear model for fibrinogen imputation used as explanatory variables the status indicator and the variables included in the Cox model: tobacco, age, center, CH, HDL and SBP (mean multiple R^2: 0.07). Five imputations were performed for each subcohort.

4.2.3 Results

The multiple imputation estimator was compared to the standard weighted estimator [6], calibrated weights estimator [3] and re-estimated weights estimator [3]. The mean of the 1,000 log relative risk estimates, corresponding to the 1,000 subcohorts, are given in Table 6. The median fraction of missing information about the fibrinogen effect was 0.22, so five imputations can be considered sufficient. For the phase-2 variable, the estimates were similar but, as expected, the standard error was larger in the case–cohort analysis than the full cohort analysis. Similar mean standard errors were obtained with multiple imputation and standard weighted estimation (Borgan's variance estimator II), they were very close to those obtained with calibrated and re-estimated weights. For phase-1 variables, the multiple imputation estimates were more precise than those obtained by weighted, calibrated and re-estimated weights analyses, and were nearly the same as those obtained from the entire cohort.

4.3 NWTS data

The NWTS cohort [16] consisted of 3,915 patients with Wilms' tumor diagnosed during 1989–1994 and followed until the earliest sign of disease progression or death for event-free survival. Baseline covariates included stage (I–IV), age at diagnosis, tumor diameter and two binary histological evaluations (favorable vs. unfavorable): the local hospital histology and central histology evaluated in a centralized reference laboratory. The former was strongly predictive of the latter (specificity 98%, sensitivity 74%) and both were available for all the patients. However, like Breslow *et al.* [3], we simulated case–cohort studies using central histology as the phase-2 variable. For the NWTS analysis, the Cox model included central histology (phase-2 variable), age as a piecewise linear variable with change point at 1 year, stage III/IV versus I/II, tumor diameter and the interactions local histology*age and stage*diameter. Breslow *et al.* [3] defined 16 strata according to event-free survival, stage (I/II or III/IV), favorable local histology (Yes or No) and age $<$ 1 year (Yes or No). They sampled controls from only three strata, defined by favorable local histology, and "age $<$ 1 year + stage I/II" (n = 120), "age $\geq$ 1 + stage I/II" (n = 160) and "age $\geq$ 1 + stage III/IV" (n = 120), while including all

the subjects in the 13 other strata. They predicted unfavorable central histology, according to local histology, stage, age, tumor diameter and the interaction local histology*stage.

These data present two specific features: first, a phase-1 variable, strongly predictive of the phase-2 variable, is available and, second, an interaction exists between the phase-2 variable and a phase-1 variable. Among these real data, in the sampled strata (all with favorable local histology) only a few controls had unfavorable central histology, especially in stratum 1 (n = 2) and stratum 4 (n = 17) (Table 7). Thus, for some subcohorts, the specific imputation model for these strata could not be estimated because of observed infinite odds ratio. Moreover, even when the subcohorts included some controls with unfavorable central histology in these strata, their number was necessarily small, and the estimator of the imputation model parameters might not have been distributed as assumed according to the asymptotic results.

Because of both particularities of these data, we considered two imputation models. The first, based on the proportional hazards model of interest and local histology, included the status indicator, local histology, stage, age, tumor diameter and the interaction local histology*stage; the second was limited to the status indicator and local histology. When the imputation model included the interaction local histology*age, biased estimates and large standard errors were observed due to the small number of unfavorable central histologies in some subgroups used for the estimation of the interaction terms (Table 8). For the imputation model including only the status indicator and the surrogate variable, the estimates were slightly biased. The simple effect of central histology, which represented the expected difference at age 1 between children with favorable or unfavorable central histology, differed slightly from the effect estimated for the full cohort (respectively, 4.11 vs 4.04); the same held true for the age effect after age 1 for the children with favorable central histology (respectively, 0.10 vs 0.11), and for the age effect after age 1 for the children with unfavorable histology (respectively, –3.63 vs –3.30). With the second imputation model, including only local histology and status indicator, we observed a small bias and good precision: although the multiple imputation was slightly biased, in particular for the age*histology interaction, it was always more precise than the standard weighted estimator and slightly more precise than the calibrated and re-estimated weighted estimators.

5 Discussion

The aim of the weighted analysis and of multiple imputation is to reconstitute the whole cohort. The former, weights the subjects in the case–cohort sample by the inverse of the probability of being observed during phase-2, but the phase-1 data observed for the other subjects are generally ignored. Alternatively, Breslow *et al.* (2009), proposed two approaches using all the phase-1 information. Multiple imputation uses all the available data supplementing them with plausible values agreeing with what is observed in the complete data.

A key aspect of multiple imputation is the construction of the prediction model.

It is necessary to reproduce correctly the relationship between the outcome and the incomplete variable, adjusting for the confounders included in the Cox model. With case–cohort data, the problem is complicated by the censoring process. One might think that it useful to include censoring time in the imputation model because, when the phase-2 variable is predictive of the event, its distribution among controls might not be the same at the beginning and the end of follow-up. However, in section 3, we demonstrated that, for a phase-2 variable with an exponential family distribution, a generalized linear model, including the case-status indicator as an exploratory variable, provides an approximately unbiased multiple imputation estimate. The proof relies on several assumptions: first, the studied event has to be rare; and second, the phase-2 variable has to be independent of the censoring time. The first assumption is precisely what justifies the use of a case–cohort design, while the second is required by the proportional hazards model. Thus, they do not represent new limitations. Our simulations showed that using all the complete subjects, cases as well as controls, and with a correct imputation model, including the case-status indicator, the multiple imputation estimator was unbiased. We also performed some simulations that confirmed no improvement of the multiple imputation estimator by adding the censoring time to the imputation model (data not shown). Confounding variables appearing in the analysis model also have to be included in the imputation model, if they remain predictive of the phase-2 variable, adjusting for the other predictors. On the other hand, misspecification of the imputation model can affect the phase-2 variable estimates but also the estimate of the omitted phase-1 variable. Omitting variables that improve the prediction can yield biased estimates and increase uncertainty about the parameter. Including variables that do not improve the prediction increases the uncertainty of the model coefficients and thus the between-imputation variance. Calibrated weighted estimates were found to be less sensitive to a misspecification of the imputation model than multiple imputation estimates concerning bias. However, multiple imputation estimates were more precise.

The completely simulated data showed that, when the imputation model was correct, the multiple imputation approach provided unbiased and efficient estimators. Both the weighted and multiple imputation estimators were centered on the true values. An important result of this simulation study was that multiple imputation correctly estimated the variance of its estimators (just as BII correctly estimated the variance of the standard weighted estimator). This finding allowed us to compare the variance estimators of case–cohort data simulated from cohort surveys, for which the variance of the estimators cannot be compared to the true value because it is unknown. As expected, the multiple imputation approach was more precise than the usual weighted estimators for the parameters associated with phase-1 variables. The former also was slightly more precise than the latter for the phase-2 variable, despite the fact that it only used a categorized transformation of the explanatory variables (the stratum indicators used for stratified sampling). One explanation might be that multiple imputation approximates the maximum partial likelihood estimate, which is more efficient than weighted estimators. The simulations implying misspecified imputation models revealed, as

expected, that the omission of a confounding variable from the imputation model had consequences in terms of bias and precision. We insist that is essential to include the stratification variable(s) and the variables included in the analysis model in the imputation model. The consequences of omitting variable(s) related to the phase-2 variable, but not to the event of interest, mainly concern precision. In these simulations, the weighted estimators did not suffer from serious bias problems, not even the calibrated weighted estimator using a misspecified model in its imputation phase. However, they suffered from appreciable losses of precision. It should be underlined that most case–cohort surveys are generally performed to answer several scientific goals dealing with different diseases and exposures. Thus, it can be difficult to define a stratified sampling efficient for all the analyses and to optimize the weighted estimators. By contrast, multiple imputation can be adapted to each phase-2 variable and each sub-study to improve the precision of the estimates of interest. When a misspecified imputation model was used, multiple imputation estimators for the phase-1 variable effects were not biased and their precision was similar to that obtained with a correct imputation model. For the phase-2 variable effect, slightly biased estimates were observed with the misspecified imputation model and weighted analyses. The effect of the misspecification on the imputation model was more noticeable in terms of precision. The loss of precision using a misspecified model as compared to a correct model was 7.6%. This loss was greater for the weighted estimate than the multiple imputation estimate, and exceeded 21%. We did not include the Z_3 variable, which was weakly correlated to the phase-2 variable. Results were less satisfactory when the correlation between the two variables was stronger, in particular concerning the precision of the weighted estimators (data not shown).

The case–cohort data simulated from the PRIME cohort study showed that multiple imputation can be used, even when no strong predictor of the phase-2 variable is available. Using the variables of the analysis model plus the case indicator in the imputation model, the multiple imputation estimator was more precise than the weighted estimators, particularly the standard estimator, for the effects of phase-1 variables. For the phase-2 variable, the multiple imputation estimator had the same precision as the standard weighted estimators and it was slightly less precise than calibrated and re-estimated weighted estimators.

The NWTS cohort data represent one of the worst possible situations for using multiple imputation. Inclusion of an interaction term between the phase-2 variable and the two age-effect components required an imputation model including similar interactions between the indicator status and the age effects. The corresponding coefficients had to be estimated in separate strata, with very low numbers of patients presenting unfavorable central histology or even no such patient at all. As a consequence, the maximum likelihood estimator of the imputation model could be expected to be biased [17, chap. 8.4] and the multiple imputation estimator reflected this bias. Although local histology was strongly predictive of central histology, imputation model 2, using only the former and the case indicator, also yielded biased estimates of the Cox model parameters. The imputed values correctly reflected the global proportion of patients with unfavorable central histology,

but not always the proportional difference between cases and controls, as was the case for the interaction terms age*local histology or stage*diameter.

It is reasonable to wonder how many imputations are needed. The number of requested imputations increases with the proportion of missing information. However, in case–cohort studies, the proportion of missing information is considerably smaller than the percentage of incompletely observed subjects, and a small number of imputations, 5–10, should suffice. With the PRIME data and using a small sampling rate (700/25,000), the results obtained with 5 imputations did not differ appreciably from those obtained with 10 (data not shown).

Herein, we presented simulations with only one phase-2 variable. However, the approach can easily be extended to several phase-2 variables. If several covariates are incomplete, we suggest imputing them using a multivariate distribution, which takes into account the correlation structure between the covariates, for instance with the mice package, which generates multivariate imputations via Markov Chain Monte Carlo (MCMC) algorithm according to their joint distribution.

We focused on situations in which covariates were time-fixed. However, the multiple imputation approach can be extended to time-varying covariates by, using mixed models to account for the repeated measurement, and the within-subject correlation structure [18].

When the phase-2 data allow consistent estimation of the imputation model, multiple imputation is an efficient technique to analyze the data from case–cohort surveys. For phase-1 variables, multiple imputation has better efficiency than weighted estimators, a valuable improvement for prediction studies or when the effect of some phase-1 variables is a focus of interest. A large number of imputations is not required to obtain good quality estimates. Software that simply implements this procedure is available: under R, the mice library; under Stata, the Imputation by Chained Equations library, or under SAS, the PROC MI and PROC MIANALYZE.

To gain time and determine whether multiple imputation provides estimates similar to weighted analysis, we suggest building the analysis model using only the case–cohort data and weighted estimators. Multiple imputation can eventually be used to reanalyze the data with the selected final model to improve the precision of the results.

Acknowledgments

This study was supported by a grant from the Région Île-de-France. The authors are grateful to Pierre-Yves Scarabin and Pierre Ducimetière for providing the PRIME data.

References

1. Langholz B, Thomas DC. Nested case–control and case–cohort methods of

sampling from a cohort: a critical comparison. *American Journal of Epidemiology* 1990; **131**:169–176.

2. Prentice RL. A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**:1–11.

3. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case–cohort data. *American Journal of Epidemiology* 2009; **169**:1398–1405. DOI: 10.1093/aje/kwp055

4. Paik MC, Tsai WY. On using the Cox proportional hazards model with missing covariates. *Biometrika* 1997; **84**:579–593. DOI: 10.1093/biomet/84.3.579

5. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case–cohort studies. *The Annals of Statistics* 1988; **16**:64–81. DOI: 10.1214/aos/1176350691

6. Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case–cohort designs. *Lifetime Data Analysis* 2000; **6**:39–58. DOI: 10.1023/A:1009661900674

7. Barlow WE. Analysis of case–cohort designs. *Journal of Clinical Epidemiology* 1999; **52**:1165–1172. DOI: 10.1016/S0895-4356(99)00102-X

8. Langholz B, Jiao J. Computational methods for case–cohort studies. *Computational Statistics and Data Analysis* 2007; **51**:3737–3748.

9. Kulich K, Lin DY. Improving de efficiency of relative-risk estimation in case–cohort studies. *Journal of the American Statistical Association* 2004; **99**:832–844. DOI: 10.1198/016214504000000584

10. Qi L, Wang CY, Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 2005; **100**:1250–1263. DOI: 10.1198/016214505000000295

11. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons: New York, 1987.

12. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* 1986; **81**:366–374.

13. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons: New York, 1987.

14. Yarnell JWG. The PRIME study: classical risk factors do not explain the severalfold differences in risk of coronary heart disease between France and Northern Ireland. *The Quarterly journal of medicine* 1998; **91**:667–676.

15. Scarabin PY, Jiao J. Plasma fibrinogen explains much of the difference in risk of coronary heart disease between France and Northern Ireland. The PRIME study. *Atherosclerosis* 2003; **166**:103–109.

16. d'Angio GJ, Breslow N, Beckwith JB, Evans A, Baum H, deLorimier A, Fernbach D, Hrabovsky E, Jones B, Kelalis P, *et al.* Treatment of Wilms' tumor. Results of the Third National Wilms' Tumor Study. *Cancer* 1989; **64**:349–360.

17. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman and Hall: London, 1974.

18. Yucel RM. Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A* 2008; **366**:2389–2403. DOI: 10.1098/rsta.2008.0038

Table 1: Distribution of subcohort by stratum.

		Correlation $(Z_2, \widetilde{Z}_2)$=0.7	
Stratum	Subcohort size	Cohort size[1]	Cases[1]
Tertile 1 $\widetilde{Z}_2$, tertile 1 Z_3	80	3,241	7
Tertile 1 $\widetilde{Z}_2$, tertile 2 Z_3	80	2,768	14
Tertile 2 $\widetilde{Z}_2$, tertile 1 Z_3	80	2,765	10
Tertile 3 $\widetilde{Z}_2$, tertile 1 Z_3	80	2,327	15
Tertile 1 $\widetilde{Z}_2$, tertile 3 Z_3	100	2,323	27
Tertile 2 $\widetilde{Z}_2$, tertile 2 Z_3	100	2,802	23
Tertile 2 $\widetilde{Z}_2$, tertile 3 Z_3	150	2,766	55
Tertile 3 $\widetilde{Z}_2$, tertile 2 Z_3	150	2,762	38
Tertile 3 $\widetilde{Z}_2$, tertile 3 Z_3	180	3,244	117
Total	1,000	25,000	308

[1]Rounded mean of the 1,000 replications.

Table 2: Parameter estimates, stratified sampling of the subcohort (mean results from 1,000 simulations).

	Correlation $(Z_2, \widetilde{Z}_2)$=0.7				
Parameter	Est	$\widehat{SE}$	SE	PC	Ratio
$\beta_1 = 0$					
Cohort	0.0143	0.1553	0.1568	95.0	
IM	0.0144	0.1553	0.1568	95.0	
BII	0.0151	0.1713	0.1763	95.3	1.10
$\beta_2 = 0$					
Cohort	0.0002	0.0618	0.0613	95.4	
IM	0.0004	0.0667	0.0681	94.2	
BII	0.0014	0.0703	0.0701	95.1	1.05
$\beta_3 = 0$					
Cohort	0.0022	0.0606	0.0624	93.9	
IM	0.0021	0.0609	0.0627	93.4	
BII	0.0015	0.0658	0.0682	94.4	1.08
$\beta_1 = 0.6931$					
Cohort	0.7133	0.1737	0.1744	95.2	
IM	0.7016	0.1742	0.1747	95.1	
BII	0.7177	0.2011	0.2011	95.7	1.15
$\beta_2 = 0.6931$					
Cohort	0.6940	0.0588	0.0589	95.0	
IM	0.6890	0.0707	0.0718	94.6	
BII	0.7040	0.0844	0.0835	95.8	1.19
$\beta_3 = 0.6931$					
Cohort	0.6955	0.0576	0.0621	92.7	
IM	0.6831	0.0601	0.0656	91.9	
BII	0.7069	0.0824	0.0894	94.1	1.37

Est, mean of the estimates; $\widehat{SE}$, mean of the standard error estimates; SE, standard error of the estimates; PC, % coverage; Ratio, weighted; SE to multiple imputation estimator; BII, Borgan's variance estimator II
IM, imputation model: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas\_i} + \alpha_2 Strata + e_i$.

Table 3: Consequences of a misspecification of the imputation model. Stratified sampling of the subcohort. Results from 1,000 simulations.

	Cohort	IM1	IM2	Calibrated1	Calibrated2
	β	β	β	β	β
Parameter	$(\widehat{SE})$	$(\widehat{SE})$	$(\widehat{SE})$	$(\widehat{SE})$	$(\widehat{SE})$
$\beta_1 = 0.6931$	0,7133	0,7017	0,6947	0,7126	0,7123
	(0,1737)	(0,1742)	(0,1743)	(0,1901)	(0,1933)
$\beta_2 = 0.6931$	0,6940	0,6867	0,7554	0,6921	0,6902
	(0,0588)	(0,0718)	(0,0694)	(0,0853)	(0,0865)
$\beta_3 = 0.6931$	0,6955	0,6911	0,7543	0,6975	0,6970
	(0,0576)	(0,0596)	(0,0573)	(0,0702)	(0,0772)

Mean of the 1,000 estimates (mean of the 1,000 standard error estimates)
IM1: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 Strata + \alpha_3 Z_3 + e_i$ (correct model)
IM2: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 Strata + e_i$ (misspecified model)
Calibrated1: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 Strata + \alpha_3 Z_3 + e_i$ (correct model)
Calibrated2: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas} + \alpha_2 Strata + e_i$ (misspecified model).

Table 4: Simple random sampling of the subcohort. Parameter estimates with correct and misspecified imputation model. Mean results from 1,000 simulations

	Correlation $(Z_2, \widetilde{Z}_2)$=0.7		
Parameter	Est	$\widehat{SE}$	SE
$\beta_1 = 0.6931$			
Cohort	0.7133	0.1737	0.1744
IM1	0.7045	0.1741	0.1745
IM2	0.6969	0.1744	0.1742
Weighted	0.7120	0.2088	0.2032
$\beta_2 = 0.6931$			
Cohort	0.6940	0.0588	0.0589
IM1	0.6853	0.0697	0.0706
IM2	0.6770	0.0750	0.0764
Weighted	0.7066	0.0909	0.0928
$\beta_3 = 0.6931$			
Cohort	0.6955	0.0576	0.0621
IM1	0.6934	0.0592	0.0632
IM2	0.6911	0.0603	0.0642
Weighted	0.7113	0.0881	0.1011

Est, mean of the estimates; $\widehat{SE}$, mean of the standard error estimates; SE, standard error of the estimates;
IM1: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas\_i} + \alpha_2 \widetilde{Z}_2 + \alpha_3 Z_3 + e_i$ (correct model)
IM2: $Z_{2i} = \alpha_0 + \alpha_1 Ind_{cas\_i} + \alpha_3 Z_3 + e_i$ (misspecified model).

Table 5: Distribution of cases by stratum.

Stratum	Cases (%)	Stratum size	Subcohort size
1 Non-smokers	153 (5.3)	2,890	475
2 Ex-smokers	261 (6.4)	4,078	800
3 Smokers 1-9 g/day	51 (6.3)	816	175
4 Smokers 10-19 g/day	55 (7.8)	707	175
5 Smokers $\geq$ 20 g/day	122 (11.9)	1,029	475
Total	642 (6.7)	9,520	2,100

Table 6: Estimates of the log relative risks.

	Whole cohort[1]	Multiple imputation[2]	Weighted estimator		
			Standard [2]	Calibrated[2]	Re-estimated[2]
	β	β	β	β	β
Variable	(SE)	(SE[3])	(SE[3])	(SE[3])	(SE[3])
Fibrinogen	0.1312	0.1381	0.1346	0.1338	0.1337
	(0.0346)	(0.0421)	(0.0425)	(0.0414)	(0.0413)
Tobacco	0.0219	0.0218	0.0220	0.0214	0.0217
	(0.0036)	(0.0037)	(0.0045)	(0.0040)	(0.0039)
Age	0.0573	0.0566	0.0573	0.0566	0.0570
	(0.0138)	(0.0138)	(0.0160)	(0.0162)	(0.0160)
Center	0.2788	0.2764	0.2795	0.2749	0.2770
	(0.0857)	(0.0863)	(0.1006)	(0.1001)	(0.0990)
CH	0.0078	0.0077	0.0078	0.0078	0.0078
	(0.0010)	(0.0010)	(0.0012)	(0.0012)	(0.0012)
HDL	-0.0526	-0.0528	-0.0528	-0.0529	-0.0529
	(0.0070)	(0.0070)	(0.0080)	(0.0084)	(0.0082)
SBP	0.0104	0.0104	0.0105	0.0105	0.0105
	(0.0016)	(0.0016)	(0.0020)	(0.0019)	(0.0019)

[1] Unique estimates provided by the PRIME cohort.
[2] Mean estimations of the 1,000 subcohorts.
[3] Asymptotic standard error (SE) of the estimate.
FN, fibrinogen; CH, cholesterol; HDL, high-density lipoprotein; SBP, systolic blood pressure.
Imputation model: $FN = \beta_0 + \beta_1 status + \beta_2 tobacco + \beta_3 age + \beta_4 center + \beta_5 CH + \beta_6 HDL + \beta_7 SBP + \epsilon$

Table 7: Distribution of central histology among cases and controls in the sampled strata.

	Controls			Cases		
	Central histology		Subcohort	Central histology		Subcohort
Stratum	Favorable	Unfavorable	fraction	Favorable	Unfavorable	fraction
1	450	2	0.27	53	4	1
2	1,569	51	0.10	216	16	1
4	897	17	0.13	188	20	1

Table 8: Mean results from 1,000 simultated phase-2 samples based on the NWTS data.

	Whole cohort[1]	Multiple imputation		Weighted estimator		
		Model 1[2]	Model 2[3]	Standard	Calibrated	Re-estimated
	β	β	β	β	β	β
Parameter	(SE)	(SE[4])	(SE[4])	(SE[4])	(SE[4])	(SE[4])
UCH	4.042	3.596	4.106	4.046	4.046	4.050
	(0.413)	(0.555)	(0.469)	(0.537)	(0.520)	(0.518)
Age0	-0.661	-0.702	-0.537	-0.669	-0.663	-0.676
	(0.326)	(0.329)	(0.329)	(0.359)	(0.324)	(0.324)
Age1	0.104	0.102	0.106	0.106	0.104	0.107
	(0.017)	(0.017)	(0.016)	(0.026)	(0.017)	(0.017)
Stage	1.346	1.441	1.353	1.344	1.345	1.344
	(0.244)	(0.257)	(0.251)	(0.346)	(0.273)	(0.272)
Diameter	0.069	0.073	0.072	0.070	0.070	0.070
	(0.014)	(0.014)	(0.014)	(0.021)	(0.015)	(0.015)
Stage*diameter	-0.076	-0.082	-0.083	-0.076	-0.076	-0.076
	(0.019)	(0.020)	(0.020)	(0.029)	(0.021)	(0.021)
UCH*age0	-2.635	-2.239	-3.097	-2.648	-2.655	-2.651
	(0.464)	(0.611)	(0.525)	(0.612)	(0.592)	(0.590)
UCH*age1	-0.058	-0.041	-0.065	-0.051	-0.050	-0.052
	(0.034)	(0.041)	(0.040)	(0.051)	(0.050)	(0.048)

[1] Unique estimates provided by the NWTS cohort.

[2] Imputation model for unfavorable central histology: status indicator, local histology, stage, age, tumor, diameter and the interaction local histology*stage.

[3] Imputation model for unfavorable central histology: status indicator and local histology.

[4] Asymptotic standard error (SE) of the estimate.

UCH, binary indicator of unfavorable central histology; Age0 and Age1, piecewise linear terms for age at diagnosis (years) before and after 1 year; stage, binary indicator of stage III/IV disease; diameter, diameter (cm) of excised tumor; SE, standard error.