A comparison of different methods for including ‘age at menopause’ in analyses of the association between hormone replacement therapy use and breast cancer

Julie A Simpson, Dallas R English, Robert J MacInnis, Dorota M Gertig, John L Hopper, Graham G Giles

Abstract

Background and methodology Late ‘age at menopause’ is a recognised risk factor for postmenopausal breast cancer and is also associated with decreased use of hormone replacement therapy (HRT). When investigating the association between HRT use and breast cancer risk it is therefore necessary to adjust for the potential confounder, ‘age at menopause’. ‘Age at menopause’, however, cannot be determined for women with a hysterectomy and ovarian conservation. Using data on 13 357 postmenopausal women in whom 396 cases of invasive breast cancer were diagnosed during 9 years of follow-up from the Melbourne Collaborative Cohort Study, we compared the estimates of relative risk of HRT use for breast cancer for three different methods of dealing with missing data: complete-case analysis, single imputation and multiple imputation.

Results ‘Age at menopause’ was missing for 17% of the data. Both HRT use and ‘age at menopause’ were significant risk factors for breast cancer, although ‘age at menopause’ only marginally confounded the estimates of risk for HRT. Women with ‘age at menopause’ missing did not represent a random sample of the population. Complete-case analyses resulted in higher estimates of the risk associated with HRT use compared with the different methods of imputation.

Discussion and conclusions We recommend that analyses investigating the association between HRT and breast cancer should present the results in two ways: excluding women with ‘age at menopause’ missing and including the women using multiple imputation. For both methods, estimates of risk, with and without the adjustment of ‘age at menopause’, should be given.

Keywords age at menopause, hormone replacement therapy, imputation, missing data

Introduction

When investigating the association between use of hormone replacement therapy (HRT) and breast cancer risk, the analysis should include adjustment for the potential confounder ‘age at menopause’. Historically, ‘age at menopause’ is the age when a woman no longer menstruates, and this is often termed the age at natural menopause. Women who have undergone a hysterectomy (a common surgical procedure) prior to when their periods would have stopped naturally have their uterus removed together with both, one or none of the ovaries. For women with the uterus and both ovaries removed (i.e. a bilateral oophorectomy), the ‘age at menopause’ is the age at surgery. For women with the uterus and one or none of the ovaries removed, which we will term a ‘simple hysterectomy’, the ‘age at menopause’ cannot be defined. This is because the women are no longer menstruating but still have functioning ovaries after surgery.

There are different approaches adopted for the above problem of missing data in analyses of observational studies investigating the association between HRT and breast cancer. In a pooled analysis of more than 50 epidemiological studies all postmenopausal women with an unknown ‘age at menopause’ were excluded, whereas in the Million Women Study (MWS) women with an unknown ‘age at menopause’ were included by assigning a fixed ‘age at menopause’.

We recommend that the risk of breast cancer associated with HRT be presented in two ways: excluding women with a hysterectomy and ovarian conservation, and including the women using multiple imputation. For both methods, the estimates of risk, with and without the adjustment of ‘age at menopause’, should be presented to show the extent of confounding.
sensitivity analysis was performed that excluded women with ‘age at menopause’ missing.

In the papers by Pike et al. and Rockhill et al., it was assumed that the so-called ‘complete-case analysis’ (i.e. exclusion of women missing ‘age at menopause’) gives unbiased estimates of the true risk of HRT for breast cancer, and hence the lower risk from assigning a fixed age at menopause was considered to be an underestimate. However, excluding these women may also have produced biased estimates if women with simple hysterectomy did not represent a random subsample of the data. In addition, the statistical power of such analyses will be reduced by excluding subjects.

An alternative to excluding the women with ‘age at menopause’ missing would be to use the methodology called multiple imputation. Multiple imputation has advantages over single imputation methods because: it allows variation in the ‘age at menopause’, any type and any number of variables can be used in the imputation procedure, and it gives the most robust variance estimates.

In this paper, we have applied multiple imputation to a prospective study of breast cancer and compared the estimates of relative risk of HRT use for breast cancer between the multiple imputation method, other single imputation techniques and complete-case analysis.

**Methods**

**Study subjects**

The data used in this paper came from 13 447 postmenopausal women who were participants of the Melbourne Collaborative Cohort Study (MCCS). The MCCS is a prospective cohort study comprising 41 528 women aged 25 and 75 years at baseline, 99.3% of whom were aged 40–69 years. Men and women volunteers were recruited from the Melbourne metropolitan area in the early 1990s. Details of the MCCS have been published elsewhere.

**Outcome of interest**

Diagnosis of invasive breast cancer was the outcome of interest. Breast cancers were identified from the population-complete Victorian Cancer Registry. Follow-up began at the age at baseline (attendance for interview) and ended (censoring) at the age of breast cancer diagnosis, age of death, the age left Victoria, or the age at 30 June 2002, whichever came first. The average follow-up was approximately 9 years. Women with in situ breast cancers were not included as cases.

**Predictor variables**

The potential predictors of breast cancer risk that we investigated were HRT use (never, past, current users), age at menopause (years), age at baseline (years), education level (none/primary, secondary/trade, tertiary), age at menarche (<13, 13, 13+ years), age at first live birth and parity combined (nulliparous, age at first live birth <25 years and parity equals one, age at first live birth <25 years and parity >1, age at first live birth ≥25 years and parity equals one, age at first live birth ≥25 years and parity >1), smoking status (never, past, current), alcohol intake (0, 1–19, 20–39, 40+ g/day), body mass index (BMI; kg/m²) and country of birth (Australia, UK, Greece/Italy). All of the above variables were recorded at baseline. No other additional variables, that were not potential predictors of breast cancer, were considered a priori to be associated with missing ‘age at menopause’.

‘Age at menopause’ was determined as the age at which a woman’s periods had ceased naturally for at least the past 12 months (age at natural menopause) or, if she had a hysterectomy before her periods stopped naturally, the age at which she had a bilateral oophorectomy (age at surgical menopause). The remaining women, with simple hysterectomy, were considered postmenopausal with unknown ‘age at menopause’ if their plasma concentration of estradiol was measured (~10% of the women) and was less than 109 pmol/l (the cut-off level at which 90% of women were correctly classified as premenopausal or naturally postmenopausal in the cohort) or their age at baseline was greater than 55 years (the age at which natural menopause had occurred in 90% of the total cohort).

**Investigation of the missing data**

The relationship between missing ‘age at menopause’ and breast cancer was examined using Kaplan Meier curves stratified by a missing value indicator. The log rank test was used to test for a difference in the probability of breast cancer at any time of follow-up for the missing and non-missing data. The relationship between ‘missingness’ (yes vs no) and the predictor variables was assessed using logistic regression.

**Multiple imputation**

In brief, five sets of data were imputed for the women with ‘age at menopause’ missing from a model that included all of the predictor variables mentioned above and the outcome variable, diagnosis of breast cancer (yes/no). For a more detailed explanation of multiple imputation refer to the Appendix.

**Statistical software**

Multiple imputation was carried out using SPLUS 6.2 (Insightful Corporation, Seattle, WA, USA). Combining the results of the analyses of the five complete datasets (imputed plus observed data) using Rubin’s rules was performed in Stata/SE 8.0 (Stata Corporation, College Station, TX, USA) using the programmes written by Carlin et al. All other statistical analyses were performed using Stata/SE 8.0.

**Results**

In the MCCS there were 13 447 postmenopausal women in whom 398 invasive breast cancer cases were diagnosed during follow-up. The incidence rate of breast cancer was 3.28 per 1000 woman-years. Approximately 6% of the women had either died or moved within the time of follow-up.

**Missing data**

The proportion of missing data ranged from 0% to 0.3% for all variables except ‘age at menopause’ where 17.5% of data were missing. For the purpose of this paper, therefore, data were imputed only for ‘age at menopause’ since the amount of missing data in the other variables was negligible. The final dataset comprised 13 357 women (396 breast cancer cases) with complete data for all variables in the model except ‘age at menopause’, which was unknown for 2322 (17.4%) women. Ninety-nine percent of the 2322 women with ‘age at menopause’ missing had a simple hysterectomy before menopause. The mean ± standard deviation (SD) observed ‘age at menopause’ was: 49.6 ± 4.5 years for never users (n = 8227), 49.1 ± 5.4 for past users (n = 1143), and 48.1 ± 5.9 for current users (n = 1665) of HRT.

The probability of breast cancer at any time point during follow-up was similar for women with ‘age at menopause’ missing and not missing (log rank test, p = 0.64). Additionally, the type of menopause (natural,
bilateral oophorectomy and simple hysterectomy before menopause) did not modify the risk of breast cancer associated with HRT (likelihood ratio test, \( p = 0.21 \)). The distribution of the predictor variables by the dichotomous variable, ‘age at menopause’ missing (yes/no), are presented in Table 1. This shows that current and past users of HRT compared with never users (\( p < 0.001 \)), an age at menarche of less than 13 years compared with 13+ years (\( p = 0.03 \)), less than 25 years of age for birth of first child compared with nulliparous (\( p = 0.002 \)), Australian born compared with Greek/Italian born (\( p = 0.002 \)) and an older age at baseline were associated with missing ‘age at menopause’ (\( p < 0.001 \)). These differences indicate that a complete-case analysis may generate biased results.

Imputed data

Data from the 840 (6.3%) women who had undergone a bilateral oophorectomy were excluded from the imputation model. Although the means of the imputed values for ‘age at menopause’ were consistent with the observed data for each of the five imputed datasets (Table 2), the proportions of subjects within specific age categories and the minimum ages were different. Further investigation showed that the observed values for ‘age at natural menopause’ were not Normally distributed but slightly negatively skewed with evidence of digit preference (Table 2, Figure 1).

Estimates of breast cancer risk associated with HRT use and ‘age at menopause’

Table 3 presents the estimated risk for breast cancer associated with HRT use and ‘age at menopause’. The results of six different Cox regression models are presented. The first two models exclude women with ‘age at menopause’ missing and so are complete-case analyses (\( n = 11 035 \)). The third model includes data from all of the women (\( n = 13 1357 \)) but does not adjust for ‘age at menopause’. Models 4 to 6 include data from all of the women (\( n = 131357 \)) and adjust for ‘age at menopause’. Model 4 uses multiple imputation, Model 5 uses single imputation (each woman assigned 55 years) for those women with ‘age at menopause’ missing, and Model 6 uses single imputation by assigning the age at 13.
hysterectomy as the ‘age at menopause’ for those women where the ‘age at menopause’ is missing. For 255 of the women with simple hysterectomy the age at hysterectomy was unknown, and therefore these women were excluded from the analysis.

In all models, current use of HRT and a later ‘age at menopause’ were statistically significant risk factors for invasive breast cancer. For the complete-case analysis (11 035 women), the estimates of risk associated with both current and past HRT use were only slightly increased when ‘age at menopause’ was included as a confounder but were much higher compared with the estimates generated from data of all 13 557 women, irrespective of the inclusion or exclusion of ‘age at menopause’ in the model. Assuming the estimate obtained from multiple imputation was unbiased, the percentage relative bias for the complete-case analyses was 14% for the estimate of current HRT use and 9% for past HRT use. A single imputation of ‘age at menopause’ as 55 years gave only slightly attenuated estimates of risk for HRT use compared with the multiple imputation of ‘age at menopause’ (percentage relative bias of –3% for current and –4% for past users).

For ‘age at menopause’, the risks associated with an increase of 1 SD were similar for the complete-case analysis and the analysis with multiple imputation of ‘age at menopause’. Both methods of single imputation of ‘age at menopause’ resulted in attenuated estimates of the risk for breast cancer (percentage relative bias was –7%).

### Discussion

We used data from the MCCS to compare different methods for dealing with missing ‘age at menopause’ in analyses investigating the risk of HRT for invasive breast cancer in postmenopausal women. Our analyses showed that excluding women with ‘age at menopause’ missing from the analysis, (i.e. a complete-case analysis) resulted in higher estimates of the risk of HRT use for breast cancer compared with the estimates obtained from all postmenopausal women, irrespective of the inclusion or exclusion of ‘age at menopause’ in the model. The estimates of risk for HRT from single imputation of 55 years were slightly attenuated compared with the multiple imputation estimates. The differences between the single imputation of 55 years’ estimates and the multiple imputation estimates were small compared with the higher estimates observed for the complete-case analysis. Although we agree with Pike et al.3 and Rockhill et al.4 that the methods of single imputation of ‘age at menopause’ give attenuated estimates of the risk of HRT, for our data the magnitude of this was small compared with

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Age at menopause (years)</th>
<th>Mean ± SD (range)</th>
<th>&lt;45</th>
<th>45–49</th>
<th>50–54</th>
<th>55+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed values for women with a natural menopause (n = 10 195)</td>
<td>49.8 ± 4.3 (21–63)</td>
<td>11%</td>
<td>27%</td>
<td>51%</td>
<td>12%</td>
<td></td>
</tr>
<tr>
<td>imputed values (n = 2322)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset 1</td>
<td>50.5 ± 4.7 (24–66)</td>
<td>10%</td>
<td>28%</td>
<td>43%</td>
<td>19%</td>
<td></td>
</tr>
<tr>
<td>Dataset 2</td>
<td>50.4 ± 4.6 (24–65)</td>
<td>10%</td>
<td>30%</td>
<td>42%</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>Dataset 3</td>
<td>50.4 ± 4.6 (24–64)</td>
<td>10%</td>
<td>29%</td>
<td>45%</td>
<td>16%</td>
<td></td>
</tr>
<tr>
<td>Dataset 4</td>
<td>50.4 ± 4.6 (24–64)</td>
<td>9%</td>
<td>31%</td>
<td>43%</td>
<td>17%</td>
<td></td>
</tr>
<tr>
<td>Dataset 5</td>
<td>50.5 ± 4.6 (24–64)</td>
<td>10%</td>
<td>30%</td>
<td>43%</td>
<td>18%</td>
<td></td>
</tr>
</tbody>
</table>

that observed for estimates obtained from a complete-case analysis. The estimates of risk associated with an increase of 1 SD in ‘age at menopause’ were attenuated for both methods of single imputation compared with multiple imputation.

Our imputation model assumed that women with ‘age at menopause’ missing (i.e. women with simple hysterectomy) were ‘missing at random’ and that the distribution of ‘age at menopause’ was similar to women with a natural menopause. Pike et al.3 proposed that a simple hysterectomy in some women may damage the blood supply to the ovaries, leading to an earlier cessation of ovarian function and thus an earlier ‘age at menopause’. Literature to support this hypothesis is inconclusive.10–14

We compared the distribution of estradiol levels between women with an intact uterus and women with a simple hysterectomy from a randomly selected subcohort of the MCCS15 and found no evidence to support the hypothesis of Pike et al.3

For the multiple imputations, we had assumed that the distribution of ‘age at menopause’ was Normally distributed. In fact, ‘age at menopause’ was slightly negatively skewed. This deviation from the assumption of Normality would be problematic if the aim of our analyses were to predict, say, the 5th or 95th percentiles of ‘age at menopause’.16

We defined women with simple hysterectomy to be postmenopausal if their age at baseline was greater than 55 years and they did not have estradiol measured. Hence, the postmenopausal women selected for this paper do not represent all possible postmenopausal women in the MCCS and are over-represented in the 55 and above age group. This sampling bias is a problem for all breast cancer cohorts where the analysis is restricted to postmenopausal women. In the paper by Rockhill et al.,5 which also investigated bias in breast cancer analyses, the dataset only included women with simple hysterectomy who were older than the age at which natural menopause had occurred in 90% of the cohort (54 years for current cigarette smokers and 56 years for non-smokers). In the MWS,2 women with a simple hysterectomy were defined as postmenopausal if they were aged 53 years or older.

‘Age at menopause’ was unknown for approximately 17% of the women in this study. Other cohort studies may have a smaller proportion of women with ‘age at menopause’ missing, possibly resulting in smaller differences in the estimates of risk between a complete-case analysis and analyses including all women. Age at menopause was unknown for 11% of all women in the pooled analysis of 50 epidemiological studies.1 For the
Nurses Health Study\textsuperscript{17} and the MWS\textsuperscript{2} the percentage of women with ‘age at menopause’ unknown was not reported. ‘Age at menopause’ was only a weak confounder in our cohort but may be a stronger confounder in other studies, possibly leading to larger differences in the estimates of risk of HRT between the different methods of imputation. In the pooled analysis of 50 epidemiological studies,\textsuperscript{1} failure to adjust for ‘time since menopause’ would have underestimated the relative risk associated with each year of use in current users of HRT (unadjusted percentage increase in relative risk of 0.8\% vs adjusted percentage increase in relative risk of 2.3\%).

Our findings showed that women with ‘age at menopause’ missing did not represent a random sample of the population, and that complete-case analyses resulted in higher estimates of the risk of HRT for breast cancer. We recommend that analyses investigating the association between HRT and breast cancer should present the results in two ways: excluding the women with ‘age at menopause’ missing and including the women with ‘age at menopause’ missing using multiple imputation. For both methods, the estimates of risk for breast cancer, with and without the adjustment of ‘age at menopause’, should be given.

### Acknowledgements

The MCCS was made possible by the contribution of many people, including the original investigators and the diligent team who recruited the participants and who continue working on follow-up. The authors would also like to express their gratitude to the many thousands of Melbourne residents who continue to participate in the study.

### Statements on funding and competing interests

#### Funding

Cohort recruitment was funded by VicHealth and The Cancer Council Victoria. This study was funded by grants from the National Health and Medical Research Council (126402; 209057; 170215; 251533) and VicHealth (1999-0227; 1998-0406) and was further supported by infrastructure provided by The Cancer Council Victoria.

#### Competing interests

None identified.

### References


3. Pike MC, Ross RK, Spicer DV. Problems involved in including women with simple hysterectomy in epidemiological studies

### Table 3

<table>
<thead>
<tr>
<th>Predictor variable</th>
<th>Hazard ratio$^a$</th>
<th>95% CI</th>
<th>Relative percentage change$^b$</th>
</tr>
</thead>
</table>
| **Model 1 (n = 11 035; 328 cases)**
(without adjustment for ‘age at menopause’)
HRT
| Never user | 1.00 | | |
| Past user | 1.27 | 0.90–1.80 | |
| Current user | 1.90 | 1.44–2.51 | |
| **Model 2 (n = 11 035; 328 cases)**
(complete-case analysis)
HRT
| Never user | 1.00 | | |
| Past user | 1.28 | 0.90–1.81 | 9\% |
| Current user | 1.93 | 1.46–2.55 | 14\% |
| Age at menopause$^c$ | 1.20 | 1.06–1.35 | –2\% |
| **Model 3 (n = 13 357; 396 cases)**
(without adjustment for ‘age at menopause’)
HRT
| Never user | 1.00 | | |
| Past user | 1.13 | 0.83–1.56 | |
| Current user | 1.65 | 1.29–2.12 | |
| **Model 4 (n = 13 357; 396 cases)**
(multiple imputation of ‘age at menopause’)
HRT
| Never user | 1.00 | | |
| Past user | 1.17 | 0.85–1.61 | |
| Current user | 1.69 | 1.31–2.17 | |
| Age at menopause$^c$ | 1.21 | 1.08–1.36 | |
| **Model 5 (n = 13 357; 396 cases)**
(single imputation of 55 years for ‘age at menopause’)
HRT
| Never user | 1.00 | | |
| Past user | 1.12 | 0.82–1.55 | –4\% |
| Current user | 1.64 | 1.28–2.11 | –3\% |
| Age at menopause$^c$ | 1.13 | 1.01–1.26 | –7\% |
| **Model 6 (n = 13 073; 393 cases)**
(single imputation by equating age at hysterectomy to ‘age at menopause’)
HRT
| Never user | 1.00 | | |
| Past user | 1.18 | 0.86–1.62 | 1\% |
| Current user | 1.71 | 1.33–2.19 | 1\% |
| Age at menopause$^c$ | 1.13 | 1.03–1.24 | –7\% |

$^a$All models include adjustment for education level, age at menarche, age at first live birth and parity (combined variable), smoking status, alcohol intake, body mass index and ethnic group.

$^b$Relative percentage change from the multiple imputation estimate.

$^c$Hazard ratio for ‘age at menopause’ represents the risk associated with an increase of 1 standard deviation (i.e. 4.8 years).

HRT, hormone replacement therapy.


7 Barzi F, Woodward M. Imputations of missing values in *(categorical and continuous) were assumed to be Normally distributed. The posterior predictive distribution of the data was assumed to be the predictive distribution of the missing data given the observed data. The posterior predictive distribution of the data was assumed to be ‘missing at random’. ‘Missing at random’ means that the probability of the data being missing may depend on any of the observed variables but is not dependent on any unobserved variable. Alternatively, for the imputation model, women with a bilateral oophorectomy were excluded. These women have an earlier ‘age at menopause’ compared with women who have had a natural menopause, resulting in a lower risk of breast cancer.19 We assumed that the distribution of ‘age at menopause’ for women with a simple hysterectomy was the same as for women with a natural menopause.

Step 2: The imputation model was used to generate five sets of imputed values for the missing data points thus creating five completed datasets. Because Rubin10 has shown that there is little benefit from using more than five to ten imputations and for our proportion of missing data (<20%), we considered that five imputations were adequate. Data were imputed from the posterior predictive distribution of the five different datasets were combined using the multiple imputation components of variation.

Step 4: For each of the five completed datasets (including women who had had a bilateral oophorectomy), a Cox proportional hazard regression model was performed with attained age as the time axis and inclusion of all predictor variables), from which the estimate of \( \hat{\theta} \) (log of the hazard ratio) of interest and its estimated variance var(\( \hat{\theta} \)) was obtained, for \( i = 1, 2, 3, 4, 5. \)

Step 5: The results from the Cox proportional hazard regression modelling of the five different datasets were combined using the rules given by Rubin20 to produce a multiple imputation estimate of \( \theta \) (log of the hazard ratio) and var(\( \hat{\theta} \)). The multiple imputation estimate of \( \hat{\theta} \) is simply the average of the \( \hat{\theta} \)'s, whereas the multiple imputation estimate of var(\( \hat{\theta} \)) allows for the between- and within-imputation components of variation.

\[ \hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i \]

\[ \text{var}(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \text{var}(\hat{\theta}_i) + \left(1 + \frac{1}{m-1} \right) \sum_{i=1}^{m} \left(\hat{\theta}_i - \hat{\theta}\right)^2, \]

where \( m \) is the number of sets of imputed values.

APPENDIX

The multiple imputation procedure consisted of the five steps outlined below:  

**Step 1:** An imputation model was devised using the recommendations of Van Buuren et al.18 The model included all of the predictor variables and the outcome variable, diagnosis of breast cancer (yes/no). Follow-up time was not included since the majority of women (92%) were followed up to 1 June 2002 with no diagnosis of breast cancer. The data on ‘age at menopause’ were assumed to be ‘missing at random’. ‘Missing at random’ means that the probability of the data being missing may depend on any of the observed variables but is not dependent on any unobserved variable.16

For the imputation model, women with a bilateral oophorectomy were excluded. These women have an earlier ‘age at menopause’ compared with women who have had a natural menopause, resulting in a lower risk of breast cancer.19 We assumed that the distribution of ‘age at menopause’ for women with a simple hysterectomy was the same as for women with a natural menopause.

**Step 2:** The imputation model was used to generate five sets of imputed values for the missing data points thus creating five completed datasets. Because Rubin10 has shown that there is little benefit from using more than five to ten imputations and for our proportion of missing data (<20%), we considered that five imputations were adequate. Data were imputed from the posterior predictive distribution of the missing data given the observed data. The posterior predictive distribution of the data was assumed to be multivariate Normal, so that all of the predictor variables (categorical and continuous) were assumed to be Normally distributed. This assumption is reasonable when no values are missing for the categorical predictor variables, as is the case for our data.21 The Expectation–Maximization (EM) algorithm (a likelihood-based approach to handling missing data) was applied first to obtain maximum likelihood estimates that were used as starting values for the data augmentation step. Data augmentation first performs a random imputation of the missing data by assuming the parameters of the posterior predictive distribution equal the maximum likelihood estimates from the EM algorithm. New parameters are then drawn from the posterior distribution of the observed and imputed data. This procedure is repeated many times creating a Markov Chain that eventually converges. Convergence of the means and standard deviations of ‘age at menopause’ to their posterior distribution was assessed by examination of time-series and autocorrelation function plots.21

**Step 3:** The imputed values generated in Step 2 using multiple imputation were imputed from the following conditional distributions:

(a) ‘Age at menopause’ must be less than or equal to ‘age at first hormone replacement therapy (HRT) use plus 12 months’ for women whose age at first HRT was known and not before age at hysterectomy (\( n = 771 \)).

(b) ‘Age at menopause’ must be greater than or equal to the ‘age at hysterectomy’ for women who had a simple hysterectomy and the age at hysterectomy was known (\( n = 2038 \)).

The distribution of ‘age at menopause’ for the imputed values was compared with the observed distribution (excluding women who had had a bilateral oophorectomy).

**Step 4:** For each of the five completed datasets (including women who had had a bilateral oophorectomy), a Cox proportional hazard regression model was performed with attained age as the time axis and inclusion of all predictor variables), from which the estimate of \( \hat{\theta} \) (log of the hazard ratio) of interest and its estimated variance var(\( \hat{\theta} \)) was obtained, for \( i = 1, 2, 3, 4, 5. \)

**Step 5:** The results from the Cox proportional hazard regression modelling of the five different datasets were combined using the rules given by Rubin20 to produce a multiple imputation estimate of \( \theta \) (log of the hazard ratio) and var(\( \hat{\theta} \)). The multiple imputation estimate of \( \hat{\theta} \) is simply the average of the \( \hat{\theta} \)'s, whereas the multiple imputation estimate of var(\( \hat{\theta} \)) allows for the between- and within-imputation components of variation.