Developing a prognostic model in the presence of missing data: an ovarian cancer case study

Taane G. Clark*, Douglas G. Altman
Centre for Statistics in Medicine, Institute of Health Sciences, University of Oxford, Old Road, Oxford OX3 7LF, United Kingdom
Received 8 October 2001; received in revised form 15 February 2002; accepted 19 July 2002

Abstract

When developing prognostic models in medicine, covariate data are often missing and the standard response is to exclude those individuals whose data are incomplete from the analyses. This practice leads to a reduction in the statistical power, and may lead to biased results. We wished to develop a prognostic model for overall survival from 1,189 primary cases (842 deaths) of epithelial ovarian cancer. A complete case analysis restricted the sample size to 518 (380 deaths). After applying a multiple imputation (MI) framework we included three real values for each one imputed, and constructed a model composed of more statistically significant prognostic factors and with increased predictive ability. Missing values can be imputed in cases where the reason for the data being missing is known, particularly where it can be explained by available data. This will increase the power of an analysis and may produce models that are more statistically reliable and applicable within clinical practice. © 2003 Elsevier Science Inc. All rights reserved.

Keywords: Ovarian cancer; Prognostic model; Overall survival; Missing data; Multiple imputation

1. Introduction

The identification of factors related to prognosis and their use in predicting the prognosis of individual patients have importance in clinical research and clinical practice. Prognostic factors are used in designing clinical trials, controlling for confounding factors in observational studies of treatment efficacy, counselling patients, formulating strategies for treatment, and for the optimal use of expensive medical tests [1]. Construction of prognostic models or prognostic indices based on multiple factors (e.g., Nottingham Prognostic Index [2] in breast cancer) is often hampered by the small size of available datasets, especially in rare diseases.

Large databases of patients are required to establish reliably the effects of different prognostic factors on long-term outcome. Frequently, however, data on prognostic factors are missing for some patients, and the standard response to this problem is to exclude these individuals from the analysis. This practice not only wastes valuable data, but can also lead to invalid results if the excluded group is a selective (i.e., non-random) sub-sample from the entire sample with respect to prognosis. In addition, the statistical power of such analyses will be reduced and the number of events per potential prognostic variable may decrease resulting in less stable results [3].

Ovarian carcinoma is the commonest cause of death from gynecological malignancy in most of the Western world. About 6,000 women in the United Kingdom develop ovarian cancer each year and about two-thirds of the women will die from the disease. The age-standardized death rate of 15/100,000 has doubled in the past 70 years, and 5-year survival is just 30% [4].

This article discusses an application of a missing data methodology called multiple imputation (MI) to an ovarian cancer dataset, with the purpose of fitting a prognostic model for overall survival. Our focus was solely on missing data for potential prognostic factors, as survival data were available for all patients. We describe the pattern of the missing data, and assess the effects of missing data on the structure and performance of the prognostic model. In addition, we discuss whether the application of MI was necessary.

2. Patients and methods

2.1. Study population

We used data from the 1,189 patients diagnosed at the Western General Hospital (Edinburgh) between 01/01/1984 and 31/12/1999 with primary cases of epithelial ovarian cancer [5]. The date of diagnosis refers to the time at which tissue samples were taken. Follow-up data were available

* Corresponding author. Tel: +44 (0)1865 226874; fax: +44 (0) 1865 226962. E-mail address: taane.clark@cancer.org.uk (T.G. Clark).
up until the end of February 2001. Patients were aged between 15 and 90 years at the time of diagnosis (median 61 years), presented at initial diagnosis with predominantly FIGO stages III and IV (64.6%), and predominantly with a serous papillary histology (51.8%). One thousand one hundred forty (95.9%) patients underwent surgery (their date of diagnosis), 894 (72.3%) patients received chemotherapy, and 860 (75.4%) had both surgery and chemotherapy. Of those patients who received chemotherapy, 654 (73.2%) patients were treated with single agent platinum regimens.

2.2. Outcomes of interest

Overall survival, or equivalently, time to death, was the outcome of interest.

2.3. Prognostic factors

The prognostic factors investigated were age at diagnosis, FIGO stage, grade of tumor (I = well differentiated, II = moderately differentiated, III = poorly differentiated), histology, the presence or absence of ascites, the diameter of the largest residual tumor mass after primary cytoreductive surgery (<2 cm, 2–5 cm, and >5 cm), performance status using the ZUBROD-ECOG-WHO scale (0 = Normal activity, 1 = Symptomatic, but nearly ambulatory, 2 = Some bed time, but needs to be in bed less than 50% of the normal daytime, 3 = Needs to be in bed more than 50% of the normal daytime, and 4 = Unable to get out of bed), and the first CA125, alkaline phosphatase, and albumin laboratory test results that fell between diagnosis and 7 days after the first chemotherapy.

2.4. Auxiliary variables

There were a number of additional variables that are not considered as factors in prognostic studies, but were potentially associated with missing data. These included: surgery (yes, no), clinical trial participation (yes, no), and chemotherapy regimen (none, platinum-based, nonplatinum based). These variables were free of missing data.

2.5. Missing data

We were only interested in missing values of variables that are potentially prognostic. Of the 10 prognostic variables considered, 8 had no complete data with the proportion of missing data varying between 2 and 43%. We did not assume that our missing data were a random sample of the whole dataset, that is, missing completely at random (MCAR) [6]. Instead, we have assumed that the missing data is missing at random (MAR) [6]. This missing data mechanism assumes that the probability that a data value is missing depends on values of variables that were actually measured. In other words, we assumed that the missingness of a variable cannot depend on the values of variables that we did not collect data on. Conceptually, it also excludes a dependence of the occurrence of missing values on the true, but unobserved, value of the variable (called missing not at random—MNAR; [6]).

Most reasonably sophisticated statistical techniques assume the data are MAR. These methods can be broadly classified as either imputation approaches or likelihood-based approaches [7] (where the Expectation-Maximization algorithm is usually applied). The major distinction is that the imputation methods fill in missing data, while likelihood-based approaches do not require the missing data to be estimated explicitly. The latter approach requires correct model specification, software is less readily available for survival data, and it is more difficult to model a mixture of binary, polytomous, and continuous prognostic factors in this setting. Hence, we have applied an imputation procedure known as multiple imputation [6,8].

2.6. Multiple imputation

The simplest form of imputation or “filling in the missing data” is “single imputation,” where missing values are estimated using single “best guesses” (e.g., the mean of other values). However, an analysis using the resulting “complete” dataset, as if the imputed values were real measurements, ignores the uncertainty in the imputation process. In contrast, multiple imputation (MI) is the analysis of multiple “complete” datasets, and incorporates an adjustment of standard errors and other statistics for the imputation uncertainty.

We applied a MI framework to account for missing prognostic factor data, and used Bayesian simulation to generate the missing data. Our approach is similar to that described by Van Buuren et al. [9]. Suppose that the statistic of interest in the analysis—here the log of the hazard ratio (HR)—is represented by a quantity $\theta$. (In reality $\theta$ is a vector of log HRs for different variables, but without loss of generality we have simplified the notation). The steps in the MI procedure we applied were:

1. Investigating the missing data
   a. Quantifying the multivariate patterns of the missing data.
   b. Plotting the proportion of missing data for each potential prognostic factor against diagnosis year to show time trends in measurement practice.
   c. Exploring the relationship between missing data of potential prognostic factors with other prognostic variables, survival information [i.e., (log) survival time and the censoring indicator], and auxiliary variables.
2. Specifying an imputation model.
3. Using the model to generate (via a random sampling procedure) $M$ sets of imputed values for the missing data points, thus creating $M$ completed datasets.
4. For each completed dataset, carrying out a Cox regression (see below), obtaining estimate $\hat{\theta}_i$ of interest and its estimated variance $\text{var}(\hat{\theta}_i)$ for $i = 1 \ldots M$.
5. Combining the results from the different datasets to obtain a prognostic model. The MI estimate of $\theta$ is

$$\hat{\theta} = \frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_i.$$
(i.e., the mean across the models fitted to the imputed datasets) and the MI estimate of variance is

$$\text{var}(\hat{\theta}) = \frac{1}{M} \sum_{j=1}^{M} \text{var}(\hat{\theta}_j) + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1}\right) \sum_{j=1}^{M} (\hat{\theta}_j - \hat{\theta})^2.$$ 

6. Constructing a final “completed data” model (later referred to as Model 2) by removing the covariate with the highest P-value and repeating steps 4 and 5 until all remaining covariates were significant at a 5% level (backward elimination).

The first term after the equals sign relates to the variance within the imputed datasets, whereas the second term captures the uncertainty due to the variability in the estimates $\hat{\theta}_j$ from the imputed datasets. The term $1 + 1/M$ is a bias correction factor.

Simulation studies have shown that the required number of repeated imputations ($M$) can be as low as three for data with 20% of missing entries [9]. We had three predictors with approximately twice this percentage, and decided that 10 repeated imputations (i.e., $M = 10$) would be a conservative choice. Unless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than 10 imputations [6]. In the following sections we describe some aspects of steps 1 to 3, where steps 2 and 3 are the Bayesian simulation. These sections are quite technical, especially the sections describing the imputation model and the sampling process.

### 2.7. Investigating the missing data

As we assumed that the missing data were MAR, we looked for evidence of MAR data by assessing associations between missing data and observed variables within our dataset. The relationships between missing potential prog-

### Table 1

<table>
<thead>
<tr>
<th>Potential prognostic factors</th>
<th>$n = 1189$</th>
<th>(N) (%)</th>
<th>Median (range)</th>
<th>(P)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age at diagnosis (years)</strong></td>
<td>1189</td>
<td>(100.0)</td>
<td>61 (15–90)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>FIGO stage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>281</td>
<td>(23.6)</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>II</td>
<td>119</td>
<td>(10.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>590</td>
<td>(49.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>178</td>
<td>(15.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>21</td>
<td>(1.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>I</td>
<td>131</td>
<td>(11.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>278</td>
<td>(23.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>641</td>
<td>(53.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>139</td>
<td>(11.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Histology</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Serous papillary</td>
<td>616</td>
<td>(51.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endometrioid</td>
<td>240</td>
<td>(20.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mucinous</td>
<td>131</td>
<td>(11.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mesonephroid (clear cell)</td>
<td>101</td>
<td>(8.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed mesodermal</td>
<td>42</td>
<td>(3.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>38</td>
<td>(3.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undifferentiated</td>
<td>21</td>
<td>(1.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ascites</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Presence</td>
<td>707</td>
<td>(59.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absence</td>
<td>417</td>
<td>(35.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>65</td>
<td>(5.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Performance status (Zubrod)</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>0</td>
<td>328</td>
<td>(27.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>228</td>
<td>(19.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>84</td>
<td>(7.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3+4</td>
<td>38</td>
<td>(3.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>511</td>
<td>(43.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Residual disease</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>$\leq 2 \text{ cm}$</td>
<td>641</td>
<td>(53.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–5 cm</td>
<td>165</td>
<td>(13.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&gt; 5 \text{ cm}$</td>
<td>302</td>
<td>(25.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>81</td>
<td>(6.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA125</td>
<td>749</td>
<td>(63.0)</td>
<td>208 (6–22878)</td>
<td>—</td>
</tr>
<tr>
<td>Log(CA125)</td>
<td>749</td>
<td>(63.0)</td>
<td>5.3 (1.8–10.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Alkaline phosphatase (ALP)</td>
<td>793</td>
<td>(66.7)</td>
<td>94 (26–1810)</td>
<td>—</td>
</tr>
<tr>
<td>Log(ALP)</td>
<td>793</td>
<td>(66.7)</td>
<td>4.5 (3.3–7.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Albumin</td>
<td>797</td>
<td>(67.0)</td>
<td>39 (20–50)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
nostic factors and survival time were explored using Kaplan-Meier curves stratified by a missing value indicator for each of the potential prognostic factors of interest. The log-rank method was used to test for differences between the distribution of survival times for the missing data and nonmissing data for a particular potential prognostic factor. For each prognostic variable the relationship between missingness (yes vs. no) and other variables was assessed using univariate logistic regressions. A more correct approach would be to produce adjusted survival curves and multivariate logistic regressions, but as the adjustment requires prognostic factors, some of which are missing data, this was not possible.

2.8. Imputation model

Step 2 is an explicit attempt to model the MAR process. Imputation models were specified for each potential prognostic factor with missing data irrespective of the quantity of missing data. The purpose of the models was to provide a set of plausible values for the missing potential prognostic factor data, and involved two modeling choices: the form of the model (linear, logistic, etc.) and the set of predictors that enter the model. The form is considered to be less important, and is largely immaterial unless the uncertainty associated with the missing entries is small [9].

For binary variables (e.g., the presence or absence of ascites) we used a logistic model, for categorical variables with three or more ordered levels (e.g., performance status) we applied a polytomous (>2 levels) logistic model, and for continuous variables (e.g., log CA125) we used normal linear regression truncated where appropriate to the credible range of values.

Composition of the sets of predictors in these models was based on guidelines in Van Buuren et al. (1999) [9]. In particular, all variables that appear in a model constructed using complete cases should be included in the imputation models. Failure to do so may bias the complete data analysis. It has also been observed that including as many predictors in the imputation model as possible tends to make the MAR assumptions more plausible. However, we recognized that there may be no substantial increase in explained variance after selecting a certain number of predictors, and identified computational problems associated with multicollinearity and inclusion of many predictors with missing data. We proposed to confine the imputation model for each prognostic factor to include all other prognostic variables from Table 1, survival variables, and auxiliary variables.

For example, the imputation model for ascites would have a logistic form with age, FIGO stage, grade, histology, residual disease, performance status, log CA125, log alkaline phosphatase, albumin, log survival time, death, surgery, clinical trial participation, and chemotherapy regimen as predictors.

2.9. Creating 10 complete datasets

To generate replacement data for the missing values in each of the 10 imputations, we did not assume a particular form for the multivariate distribution of our data. This contrasts with the approach of Schafer (1997), which assumes the distribution to be a multivariate Normal for continuous data, a log-linear model for categorical data, or a general location model for a mixture of continuous and categorical. We assumed only that a multivariate distribution existed, and that Gibbs sampling [10] of the conditional distributions (based on the models above) could generate samples from it. Specifically, this was an iterative process in which the missing data for a prognostic variable were estimated using its imputation model, and then these completed data were used in the estimation of the missing values for another variable. Each iteration ended when all variables had been updated. The application of Gibbs sampling ensured that the imputation process was not deterministic, there was random variation between the completed datasets, and is the main reason the approach is considered (misleadingly) to be a form of Bayesian simulation. The Gibbs sampling algorithm was run for 80 iterations updating for each of the 10 imputations. The time taken for each variable was a function of the prevalence of missing values. It has been found that in the presence of large amounts of missing data, convergence can be obtained in as few as 10 iterations [9]. By plotting the standard deviations and means of the imputations by iteration, an assessment of convergence was made. These plots allowed us to assess whether between imputation-variability had stabilized and whether estimates were free of trend.

The frequency distributions of prognostic variables in the 10 datasets with imputed missing values were compared with the frequency distribution in the original dataset using histograms and basic summary statistics (e.g., medians, ranges, and proportions). This comparison assessed the extent to which the imputed datasets were consistent among themselves and with the original data.

2.10. Statistical models

Univariate analyses on continuous prognostic factors were performed using Cox regression [10], and categorical (including categorized continuous) prognostic factors were analyzed using the Kaplan-Meier and the log-rank methods. The linearity of continuous variables was assessed using fractional polynomials [11]. Univariate analyses were performed on complete cases in the original dataset.

The fundamental method of multivariate analysis was Cox regression [12]. The proportional hazard assumption for each predictor was tested using an approximate score statistic of linear correlation between the rank order of failure times in the sample and Schoenfeld partial residuals [13].

Like the final “completed data” model, the final “complete case” model (later referred to as Model 1) was obtained using backward elimination, but was based on the original data and not the completed data.

2.11. Measures of model performance

We evaluated the predictive performance of models by considering measures of discrimination and calibration.
Discrimination refers to the ability to distinguish between high risk and low risk patients, and was quantified using the c-index [13] (a generalization of the area under the Receiver Operating Characteristic (ROC) curve [13]) and the Nagelkerke $R^2 (R_N^2)$ [13]. The c-index is the probability of concordance between predicted and observed survival based on pairs of individuals, with $c = 0.5$ for random predictions and $c = 1$ for a perfectly discriminating model. Similarly, $R_N^2 = 0$ indicates no predictive ability and $R_N^2 = 1$ indicates perfect predictions. The term predictive ability is synonymous with discrimination and describes the ability of a set of prognostic factors to explain variation in outcome. In theory, high predictive ability implies accurate prediction for individual patients.

Calibration refers to whether the predicted probabilities agree with observed probabilities, or equivalently, indicates the ability of the model to make unbiased estimates of outcome. Calibration was quantified using an estimate of slope shrinkage and was based on partitioning the patients into 10 risk groups [13]. Shrinkage is the flattening of the plot of predicted (x-axis) and observed (y-axis) probabilities away from the 45 degree line, caused by overfitting. The estimate of slope was based on the mean slope from 200 bootstrap samples of the data. An adjusted slope close to one indicates little shrinkage and good calibration.

2.12. Statistical software

All statistical analyses were carried out with S-Plus 2000 using the Hmisc, Design, and MICE libraries. MICE is available from www.multiple-imputation.com.

3. Results

3.1. Long-term outcome

Follow-up information was available on all 1,189 patients, but some were lost to follow-up before the end of study period. 842 (70.8%) patients had died at the time of censoring the data. Median follow-up in the 347 (29.2%) patients who were not known to have died was 1,665 days (range 29–5,852 days). Five-year survival in the cohort was
and missing strata within each of the factors (log) CA125, 3.3. Evidence of MAR data

in the ovarian cancer literature. in prominence of performance status as a prognostic factor

literature until the mid-1980s, and there has been decrease by the fact that CA125 was not commonly reported in the

maximum of 71% in 1995. These trends can be explained increasing trend in time with a minimum of 18% in 1986 and

early in time from 85 to 21% between 1984 and 1999, and contrast, the proportion of missing CA125 data decreased lin-

albumin, grade, and residual disease were constant. In con-

show that the proportions for ascites, alkaline phosphatase,

seven hundred thirty-nine (85.0%) of the missing cells re-

four or more missing values, but only 4 (0.4%) patients with

senting less than two missing values per individual in the

2,045 (17.2%) were missing in 831 (69.9%) patients, repre-

were missing for 43.0, 37.0, 33.3, and 33.0% of patients, respectively. Of the 1,1890 prognostic factor data values,

were missing in 831 (69.9%) patients, representing less than two missing values per individual in the entire cohort. Two hundred thirty-six (19.8%) patients had four or more missing values, but only 4 (0.4%) patients with seven or more missing prognostic factors. One thousand seven hundred thirty-nine (85.0%) of the missing cells resulted from missing information on albumin, alkaline phosphatase, CA125, and performance status. The number of patients contributing to a complete case analysis using all the prognostic factors in Table 1 would be 358 (245 deaths).

Plots of the proportion of missing data by diagnosis year show that the proportions for ascites, alkaline phosphatase, albumin, grade, FIGO stage, and performance status (Fig. 1), showed no visual or statistical evidence of significant differences (Table 2). However, there was a difference between the survival distributions of patients with and without missing data for ascites (P < .002) (Fig. 1), albumin (P < .003), alkaline phosphatase (P = .020) (Fig. 1), and residual disease (P = .020) (Fig. 1). Those patients missing albumin and alkaline phosphatase results had a better prognosis, suggesting that eliminating the patients with missing values would lead to an underestimate of the true survival of the cohort. The opposite effect was seen for ascites and residual disease. Those patients missing ascites data may be more likely to have ascites present at diagnosis. However, as the survival curves are not adjusted for other variables, these effects may not be real and, of course, in theory one might miss an effect if masked by some other association.

The univariate logistic models indicated that the missingness of potential prognostic factors were (potentially) associated with other potential prognostic factors and auxiliary variables. Histology and clinical trial participation were associated with the missingness of all but one prognostic variable (Table 2). These results must be interpreted with the same caution because they were from univariate analyses.

3.4. Imputed data

We completed 10 data sets by imputing 2,045 values in each. As a consequence, 6,265 additional real data values were incorporated into each dataset. The summary statistics for the datasets with imputed data are shown in Table 3. The narrow ranges of imputation values for each potential prognostic variable coincides with the visual impression that the distributions for each of the potential prognostic factors in the 10 imputed datasets were similar. The prevalences (%) of categorical prognostic factors in the original data (ignoring missing data) were consistent with those from the 10 imputations. The median and range of albumin, log CA125, and alkaline phosphatase in the original data were consistent with the median of the median of the 10 imputation distributions, and the extreme values of these distributions, respectively.
3.5. Univariate survival analysis

Univariate analyses suggested that age, FIGO stage, the presence or absence of ascites, performance status, histology, residual disease, albumin, grade, and (log) CA125 and alkaline phosphatase were significant (P < .05) prognostic factors of overall survival (Table 1). Investigations of linearity of CA125, alkaline phosphatase, age, and albumin in the univariate models suggested that CA125 and alkaline phosphatase required a natural logarithmic transformation.

3.6. Cox models

Table 4 shows the results from fitting the Cox models. The first column represents an analysis of complete cases in the original data (Model 1). As four factors, each with missing values, were found not to be prognostic, the analysable dataset was 518 (380 deaths), that is, larger than the 358 patients available when fitting the full model. The second column represents a pooled analysis using 10 complete datasets with imputed missing values (Model 2). Age, FIGO stage, performance status, residual disease, and (log) alkaline phosphatase were significant (P < .05) prognostic factors in both models. Specifically, there was a greater risk of mortality associated with older age, higher FIGO stage, the presence of ascites, worse performance status, and greater
Standard approach is to exclude those individuals for whom data are missing on one or more prognostic variables. Although missing data are very common, they are infrequently discussed. The standard approach is to exclude those individuals for whom data are missing from the analysis. In addition to being a waste of data that are often costly to collect, this practice could lead to invalid results if the excluded group is a selective subsample (nonrandom sample) from the entire sample with respect to outcome.

We developed a prognostic model for overall survival in 1189 ovarian cancer patients. An analysis based on 518 patients with complete prognostic factor data found that age, FIGO stage, histology, performance status, residual disease, and alkaline phosphatase were significant (P < .05) prognostic factors for overall survival (Model 1). We applied Bayesian simulation to construct 10 complete datasets for all 1,189 patients. Within a multiple imputation (MI) framework we analyzed these individually and pooled the results, adjusting standard errors and other statistics for the uncertainty consequent upon imputing the missing data. By imputing missing data we added three actual data values for each imputed value, and doubled our sample size. With the increased statistical power, grade and ascites were additionally found to be significant (Model 2). The hazard ratios of the prognostic factors common to both models were similar, except that FIGO stage had reduced effect sizes and a weaker linear trend in the augmented data.

Additionally, we found that the models applied to completed data (i.e., the 10 datasets with imputed missing values) had better calibration (i.e., greater ability to produce unbiased estimates of outcome), and superior discrimination (i.e., improved ability to provide accurate predictions for individual patients) than models based on complete cases only. This is a result of increased sample size and possibly less bias. There was little difference between the discrimination measures of Model 1 and Model 2 when applied to the completed data. An explanation for the small difference is the focus of this discussion. In particular, given the evidence of only a small difference in predictive ability between Models 1 and 2, was the application of MI procedure necessary? The answer to these points lies in a discussion of the assumptions and limitations of the statistical methodology.

The major assumption associated with the complete case construction of Model 1 is that the reason for missingness is not related to outcome. If we assume the prognostic model is correct, bias will not arise if the reason for missingness of a variable is associated with other variables in the model or the true values of the missing prognostic variable itself. However, the deletion of noncomplete cases in Model 1 (671 of 1,189 patients included) with (≥1) missing prognostic factor data will result in less precision. There is evidence, however, that the missingness may be related to outcome. In particular, univariate survival analyses indicated that those patients missing albumin, alkaline phosphatase, residual disease, and ascites had either better or worse survival than those with recorded values. However, only adjusted survival analyses could determine whether the associations were real, but these were not possible because of the multivariate nature of the missingness.

Model 2 assumes that the missing mechanism is missing at random (MAR), that is, the missingness is not re-

3.7. Measures of model performance

Table 5 shows the measures of predictive accuracy for Model 1 and Model 2 applied to the original data and completed data (i.e., original data augmented with imputed missing values). Models applied to the completed data had slope shrinkage values close to 1, indicating less overfitting and less need for recalibration of prognostic factor effects. In addition, the models applied to completed data had substantially better measures of discrimination.

Model 1 was marginally inferior to Model 2 in overall discrimination when applied to the completed data. However, high values of the c-index (~0.78) and Nagelkerke $R^2$ (~0.50) for both models applied to the completed data indicate that their respective set of prognostic factors were explaining the variation in outcome reasonably well.

If we compare the discrimination between Model 1 and Model 2 when applied to the datasets used in their construction, there is clear difference. Figure 2 (left) shows this difference. Four risk groups were created for each of Model 1 and Model 2 by partitioning linear predictions from the respective models into quartiles. The plot shows there is greater separation between risk groups using Model 2, particularly at lower survival times. The differences are a result of increased sample size and two extra prognostic factors. Figure 2 (right) shows the effect of increased sample size only, with the same risk groups from Model 2 as before, applied to all patients and only those with complete data for Model 2. This plot shows that a complete case analysis using Model 2 may overestimate the survival in both risk group 2 after 3 years and risk group 4 in the first year. In both plots, the confidence intervals around the survival curves are narrower in the augmented data.

4. Discussion

Most prognostic studies are performed retrospectively. The construction of prognostic models is problematic when data are missing on one or more prognostic variables. Although missing data are very common, they are infrequently discussed. The standard approach is to exclude those individuals for whom data are
lated to the true (unobserved) values of the missing data. A complete case analysis does not make this assumption. However, MAR also assumes that the missingness of a variable depends on the values of variables that we collected data on. This implies that the imputation approach may produce valid inferences in the presence of response related missingness. Our imputation model for each variable consisted of all other prognostic factors, auxiliary variables (not incorporated into Model 1) and survival information. The survival time and censoring indicator were incorporated into the imputation model because there was evidence that missingness was related to survival. The somewhat circular approach allows us to include all patients. As a general rule, using all available information yields multiple imputations that have minimal bias and maximal certainty [9]. This principle implies that the number of predictors should be as large as possible. In addition, and most importantly, it has been observed that including many predictors in the imputation model makes the MAR assumptions more plausible, and reduces the need to make special adjustments for missingness associated with the unobserved true values of missing data [9,14]. Ascites was present in Model 2 and not Model 1, and this may not be entirely a result of increasing sample size, but may also result from reducing bias in the estimation of its effect. However, ascites is a relatively weak prognostic factor and its inclusion has little effect on the predictive ability.

An appealing feature of the application of MI is that it yields actual imputed values, and potentially allows scope for investigating the properties of the resulting prognostic models. Table 3 shows that the imputed data are consistent with the original data, but this need not be true in other applications. The MI process has incorporated the uncertainty of imputing those variables with greater proportions of missing data. Given that using our method allows us to include three extra real data values for each imputed one, the confidence limits have narrowed in most cases (Table 4), less in variables with high levels of missing data (e.g., performance status), and more in those with low levels of missingness (e.g., FIGO stage).
Most data are multivariate in nature, so a small proportion of missing data for several variables can lead to a severely depleted complete case analysis. MI seems appropriate in this setting if the original dataset is not too small. The similarity of the predictive abilities of Model 1 and Model 2 may be explained by the fact that there are no major deviations away from their respective modeling assumptions. This does not imply that the complete case analysis was appropriate in the first place. Using imputed data we are incorporating patients that are removed merely because one or more of their prognostic factors are missing and, as a result, increasing power and adding precision to an analysis (see Fig. 2). Of course, we cannot know what the effect of imputation will be without performing the analysis. In other cases we might find a greater effect for imputation, and this may suggest a complete case analysis is not appropriate. From this point of view our approach may be viewed as a sensitivity analysis, and ultimately we need to use judgement about the plausibility of assumptions in a particular situation to assess which is the primary analysis.

Acknowledgments

We wish to thank Moira Stewart, Hani Gabra, and John Smyth for making the data available, and Christina Davies and Victoria Cornelius for reading earlier drafts of the manuscript. Taane Clark holds a National Health Service (UK) Research Training Fellowship. Douglas Altman is supported by Cancer Research UK.

References