Medical Decision Making
http://mdm.sagepub.com

Evaluation of Imputation Methods in Ovarian Tumor Diagnostic Models Using Generalized Linear Models and Support Vector Machines
Ioannis Dimou, Ben Van Calster, Sabine Van Huffel, Dirk Timmerman and Michalis Zervakis
Med Decis Making 2010; 30; 123 originally published online Jul 15, 2009;
DOI: 10.1177/0272989X09340579

The online version of this article can be found at:
http://mdm.sagepub.com/cgi/content/abstract/30/1/123

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Society for Medical Decision Making

Additional services and information for *Medical Decision Making* can be found at:

Email Alerts: http://mdm.sagepub.com/cgi/alerts

Subscriptions: http://mdm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations http://mdm.sagepub.com/cgi/content/refs/30/1/123

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

Evaluation of Imputation Methods in Ovarian Tumor Diagnostic Models Using Generalized Linear Models and Support Vector Machines

Ioannis Dimou, MSc, Ben Van Calster, MSc, PhD, Sabine Van Huffel, MSc, PhD, Dirk Timmerman, MD, PhD, Michalis Zervakis, PhD

Neglecting missing values in diagnostic models can result in unreliable and suboptimal performance on new data. In this study, the authors imputed missing values for the CA-125 tumor marker in a large data set of ovarian tumors that was used to develop models for predicting malignancy. Four imputation techniques were applied: regression imputation, expectation-maximization, data augmentation, and hotdeck. Models using the imputed data sets were compared with models without CA-125 to investigate the important clinical issue concerning the necessity of CA-125 information for diagnostic models and with models using only complete cases to investigate differences between imputation and complete case strategies for missing values. The models are based on Bayesian generalized linear models (GLMs) and Bayesian least squares support vector machines. Results indicate that the use of CA-125 resulted in small, clinically nonsignificant increases in the AUC of diagnostic models. Minor differences between imputation methods were observed, and imputing CA-125 resulted in minor differences in the AUC compared with complete case analysis (CCA). However, GLM parameter estimates of predictor variables often differed between CCA and models based on imputation. The authors conclude that CA-125 is not indispensable in diagnostic models for ovarian tumors and that missing value imputation is preferred over CCA. **Key words:** *ovarian tumors; CA-125; imputation; complete case analysis; least squares support vector machines; Bayesian generalized linear model; AUC.* **(*Med Decis Making* 2010;30:123–131)**

Ovarian cancer is one of the most lethal cancers in women.[1] An accurate presurgical diagnosis of the tumor is essential for effective treatment and improved survival.[2] CA-125 is a tumor marker used to diagnose ovarian tumors, with elevated levels in serum, raising concern for malignancy. However, the value of CA-125 in diagnostic models is uncertain,[3] and its use in practice varies widely. Researchers using practice-based data to develop or test diagnostic models often encounter subjects with missing CA-125 values. As shown in a review article,[4] researchers tend to use ad hoc solutions for missing data, such as ignoring incomplete variables, omitting cases with missing values, or using a naive imputation technique (e.g., unconditional mean imputation). Such techniques may create more problems than they solve, distorting estimates, standard errors, and hypothesis tests.[5] Applying a more elaborate imputation technique can provide statistically sound values for the missing cases, thus increasing data quality for developing diagnostic models.

We conducted this study to compare different methods for missing data imputation for the development of clinical decision support systems and to compare models with and without CA-125. Specifically,

Received 11 December 2008 from the Department of Electronics and Computer Engineering, Technical University of Crete, Chania, Greece (ID, MZ); Department of Electrical Engineering (ESAT-SISTA), Katholieke Universiteit Leuven, Leuven, Belgium (BVC, SVH); and Department of Obstetrics and Gynaecology, University Hospitals K.U. Leuven, Leuven, Belgium (DT). Financial support for this study was provided in part by EU: BIOPATTERN (FP6-2002-IST 508803), eTUMOR (FP6-2002-LIFESCIHEALTH 503094); Research Council KUL: GOAAMBioRICS, CoE EF/05/006 Optimization in Engineering (OPTEC); and Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, "Dynamical Systems, Control and Optimization," 2007–2011). The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report. See also the Web appendix at http://mdm/sagepub.com/supplemental. Revision accepted for publication 28 April 2009.

Address correspondence to Ioannis Dimou, MSc, Department of Electronics and Computer Engineering, Technical University of Crete, Akrotiri Campus, 73100 Chania, Greece; e-mail: jdimou@gmail.com.

DOI: 10.1177/0272989X09340579

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

we imputed missing values of the CA-125 tumor marker for ovarian tumors and evaluated the imputations by developing diagnostic models predicting malignancy of the tumor.

Few studies have applied and compared imputation techniques in this context.[6,7]

We used data from the International Ovarian Tumor Analysis (IOTA) group multicenter study, which obtained extensive standardized diagnostic information for a large number of patients.[8,9] Reflecting the differences concerning the use of CA-125 in clinical practice, some IOTA investigators by default did or did not measure CA-125 levels, whereas others omitted measuring CA-125 levels only when the tumor looked clearly benign on ultrasound examination.

We applied 4 different imputation techniques for the missing CA-125 values: regression imputation,[10] expectation-maximization (EM),[11] data augmentation (DA),[12] and hotdeck imputation.[13] We developed models to predict tumor malignancy using imputed data sets, a data set without CA-125, and a complete case data set. This methodology allowed us to evaluate 1) the effect of imputation and 2) to investigate the necessity of CA-125 information for predicting malignancy.

A statistical limitation in real-life imputation studies, as compared to simulation studies, is that the missing values and the missing data mechanism are unknown such that the validity of the imputations cannot be checked. In the present study, the effectiveness (rather than the validity) of the imputations was primarily evaluated through the diagnostic performance of the developed models. In addition, imputation and complete case analysis (CCA) were compared by investigating estimated model parameters.

METHODS

Patients and Sources

The IOTA data set consists of 1066 nonpregnant women with at least 1 persistent ovarian tumor, collected at 9 clinical centers throughout Europe. When a tumor was present on both ovaries, data from the most complex mass were included in the study. The main outcome of the study was the diagnosis of the tumor as benign or malignant. The variables collected in the IOTA data set involved demographic and cancer history information, pain during examination, around 40 morphologic and blood flow measurements to describe the tumor, and the serum tumor marker CA-125. Apart from CA-125, the measurement of which was merely encouraged, investigators had to provide complete data. An extensive description of the data set and the data collection protocol can be found in Timmerman and others.[9] Except for CA-125, missing values in the IOTA data set were structural missings, in the sense that nothing could be measured. For example, if no papillary structures were present, none of the variables describing the papillary structures were applicable. After discussion with the clinical experts, these structural missings were imputed with zero.

Most imputation techniques assume that missing data are missing at random (MAR) or missing completely at random (MCAR). MAR implies that missingness depends on observed information but not on missing information,[10] whereas MCAR implies that the missing values are a random subsample of the data set. MCAR is usually unrealistic. Under MAR, a missing CA-125 value may depend on values of other measurements but not on the (unavailable) value of CA-125. Note that MAR allows that missingness of CA-125 is related to CA-125 levels, yet only indirectly through associations between both CA-125 levels and CA-125 missingness with other variables.[14] Data are missing not at random (MNAR) when missingness depends on the missing values. Imputation methods assuming MNAR are considerably more complex, and no uniform methods are available. Based on the available information, CA-125 missingness in our data set is likely to largely follow the MAR assumption.

The missingness mechanism of CA-125 was investigated by checking the association of CA-125 missingness (a binary indicator with a value of 1 if CA-125 was missing and 0 otherwise) with other variables using likelihood ratio chi-squared tests. The statistically significant missingness predictors of CA-125 are shown in Table 1. These results indicate that there was a strong relationship between CA-125 missingness and other covariates, such that it is unlikely that the missing values are MCAR. When MCAR does not hold, CCA very often leads to biased results. Because CA-125 was less often measured for tumors that looked clearly benign on ultrasound, CCA is likely to be biased in our application.

Imputation Methods

The 4 imputation techniques that were applied assume MAR. Some other techniques were deemed unsuitable for this specific study due to experimental limitations or the underlying assumptions. For

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

Table 1 Statistically Significant CA-125 Missingness Predictors

Variable	*P* Value
Negatively associated with missingness	
Color score of intratumoral flow (1–4)	0.039
Papillations (yes-no)	0.038
Time-averaged maximal velocity, cm/s	0.016
Number of papillations	0.014
Number of locules (0, 1, 2, 3, 4, 5–10, >10)	0.009
Maximal diameter of papillation, mm	0.006
Fluid in pouch of Douglas, mm	0.006
Hysterectomy (yes-no)	0.005
Septum thickness, mm	0.003
Volume of largest solid component, mL	0.002
Ratio solid component to lesion	0.001
Volume of the ovary, mL	0.001
Volume of the lesion, mL	0.001
Years postmenopause	<0.001
Ascites (yes-no)	<0.001
Postmenopausal (yes-no)	<0.001
Age, years	<0.001
Maximal diameter of largest solid component, mm	<0.001
Maximal diameter of the lesion, mm	<0.001
Maximal diameter of the ovary, mm	<0.001
Positively associated with missingness	
Resistance index	0.045
Pulsatility index	0.015
Acoustic shadows (yes-no)	0.001
Categorical variables	
Origin of tumor (ovarian or other)	0.005
Locularity (6 categories)	<0.001

example, multiple imputation (MI)[5,14,15] is a statistical method that accounts for uncertainty in the imputations by generating multiple imputed data sets such that standard errors of estimated quantities are not underestimated. MI was not used in our study because it was impractical for this study's experimental setup and because our primary focus was to investigate how well we predict malignancy in new data.

To describe the imputation methods, some mathematical notation is required. Suppose we have N data points $(\mathbf{x}_n, y_n)$, $n = 1, \ldots, N$, where $\mathbf{x}_n \in \mathrm{P}^q$ consists of all q measured variables, and $y_n \in \{0, 1\}$ indicates whether the tumor is benign or malignant. Then, $\mathbf{x}_n$ can be split up in the observed values $\mathbf{a}_n$ and the missing values $\mathbf{m}_n$.

Regression Imputation

Regression imputation (RI) aims at imputing missing values using a regression model based on the complete cases. The regression model predicts CA-125 using the other variables (except the tumor outcome because that is the outcome of the diagnostic models developed later). A regression model was built using stepwise variable selection. CA-125 was bilog transformed to approach a normal distribution. Regression imputation is a form of conditional mean imputation.

Intuitively, one may be tempted to exclude a significant missingness predictor from a regression imputation model because the regression parameter for the available cases could be different from the parameter for the incomplete cases. However, the use of missingness predictors is encouraged because it renders the MAR assumption more plausible (v. MNAR).[14]

EM Imputation

Each $\mathbf{x}_n$ is generated by the underlying distribution $p(\mathbf{x}|\theta) = p(\mathbf{a},\mathbf{m}|\theta)$, where θ are the distribution's parameters. If there are no missing values, θ can be estimated easily. When missing values exist, the EM algorithm is a deterministic method to circumvent the problem of estimating θ without the missing values $\mathbf{m}$ by integrating $\log[p(\mathbf{a},\mathbf{m}|\theta)]$ over $\mathbf{m}$, yielding the expected log-likelihood $E(\ell)$.[11,16] In practice, one iteratively alternates between an E-step in which the missing values $\mathbf{m}$ are estimated using initial parameter values θ_0 (first iteration) or estimated parameter values θ_{t-1} (iteration t) and an M-step in which $E(\ell)$ is maximized with respect to θ. The M-step yields θ_t, the updated estimate of the model parameters to be used in the next E-step.

The algorithm's convergence relies on the data set and the stopping criteria. In practice, EM's convergence rate depends on the ratio of missing information and the size of the data set. A variation of the EM algorithm, called regularized EM, was used in this study.[16] For the imputation using EM, multivariate normality was assumed. We relied on the robustness of the algorithm against violations of this assumption. A bilog transformation was introduced for the highly skewed CA-125 levels to approach a normal distribution.

Data Augmentation

DA[12] is a stochastic rather than deterministic method to circumvent the problem of estimating θ without the missing values $\mathbf{m}$. In DA, missing values are randomly imputed given a missing data distribution based on assumed values for θ. Given $\mathbf{a}$ and the imputed values for $\mathbf{m}$, a posterior distribution for θ is constructed from which updated estimates for θ

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

are drawn. This procedure is implemented using the Gibbs sampling framework. In this 2-step iterative process, the distributions for the missing values and the model parameters stabilize, leading to convergence in distribution, whereas EM converges in its estimates. For the convergence rate of DA, the arguments given for EM apply here too. The difference is that, for DA, the values still change after convergence, but their distribution remains fixed.

Both the EM and DA algorithms are based on a Bayesian estimation framework. As a consequence, they rely on the choice of a prior distribution for the model parameters reflecting prior knowledge of the data. In the existing literature, often an uninformative prior is selected to reflect lack of a priori knowledge.[4,12] We experienced little variation when using a different prior. The DA implementation also assumed multivariate normality, and CA-125 measurements were bilog transformed to approach normality. As for EM, we relied on the algorithm's robustness against violations of this assumption.

As in the case of EM imputation described above, the CA-125 variable was bilog transformed to account for its highly skewed distribution, whereas other predictors used in the imputation model had to be log transformed to alleviate skewness. Also, to avoid numerical issues, we scaled the variables for both EM and DA methods to the $[-1, +1]$ interval.

Hotdeck Imputation

Hotdeck refers to the imputation of missing values with observed values from other cases (the donors). Many approaches for selecting a donor exist, and in this article, we used a nearest neighbor method.[17] In this method, the donor for an incomplete case is the most similar complete case according to a similarity metric. A commonly used similarity metric is the sum of absolute differences for all variables (except the outcome). This method depends on the existence of similar cases in the data set, which is more likely when N is large. To improve the performance, we rescaled the numerical variables to the $[-1, +1]$ interval.

Of all N cases, there are C complete cases $(\mathbf{x}_c, y_c)$, $c = 1, \ldots, C$. The missing CA-125 level of patient $i (i = 1, \ldots, N - C)$ is imputed using sample $\mathbf{x}_c$, where c is chosen using

$$\arg\min_c [d(\mathbf{a}_c, \mathbf{a}_i)]. \quad (1)$$

The distance d, as explained above, is

$$d(\mathbf{a}_n, \mathbf{a}_i) = |\mathbf{a}_n - \mathbf{a}_i|. \quad (2)$$

Evaluation Using Diagnostic Models

Two types of diagnostic models were implemented: least squares support vector machines (LS-SVMs)[18] and generalized linear models (GLMs).[19] Due to space limitations, LS-SVMs are described in the appendix (available online at http://mdm/sagepub.com/supplemental). Support vector machines are flexible models that can automatically account for nonlinearity; therefore, we used LS-SVMs to complement the GLM models.

Developing Diagnostic Models Using Generalized Linear Models

A linear model for binary classification typically links the logit of the probability of "success" (malignancy) to a linear combination of input variables:

$$\text{logit}[P(Y = 1|\mathbf{x})] = \mathbf{w}^T\mathbf{x} + b. \quad (3)$$

This model has the form of a logistic regression model, a member of the family of GLMs.[19] Note, however, that we used this algorithm in a simple way without consideration of interactions or transformations of variables (similar to Van der Heijden and others[6]). A Bayesian approach based on the evidence procedure was used for the GLM models, which naturally protects for overfitting by keeping model parameters small.[20]

Experimental Setup

The IOTA data set is split up into a training set of 754 patients and a test set of 312 patients, using stratification for outcome and clinical center. All 4 imputation methods were applied to the training and test set data simultaneously.

Diagnostic models were developed for 6 different scenarios: once for each imputation method, once for a CCA, and once for a complete variable analysis (CVA) in which CA-125 is ignored. Binary variables were coded as -1 v. $+1$. In the training set, ordinal and continuous variables were rescaled into the $[-1, +1]$ interval. In the test set, these variables were transformed accordingly. Variable selection proceeded in 2 steps. First, 20 variables were selected based on previous variable selection processes done on the same training set using logistic regression and LS-SVMs.[9,21] These 20 variables (including CA-125) can be seen as the most important variables for predicting tumor malignancy. Second, separately for each scenario, the 20 variables (19 for CVA) were ranked according to their importance. This was done using automatic relevance determination,[20] a Bayesian variable selection method. For the imputation and CCA scenarios, CA-125 was by definition ranked as the most important variable.

After the variable selection process, we developed diagnostic models using the $3, 4, \ldots, 20$ (or 19 for CVA) most important variables, resulting in 18 (17 for CVA) models. These models were applied to the test set. Test set performance was evaluated using the area under the receiver operating characteristic curve (AUC).

In a next step, we constructed 100 new train-test set splits (repeated data splitting) with stratification for outcome. For each of the 6 scenarios, the 18 (17 for CVA) models were trained on each training set and tested on the accompanying test set. The 100 test set AUCs were summarized by their median.

Finally, to compare model parameters, we fitted a GLM model with all 20 variables on all data: 1066 cases for imputed data sets and complete cases for CCA. For each imputation method, we computed the percent differences of the parameter estimates with the CCA parameter estimates.

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

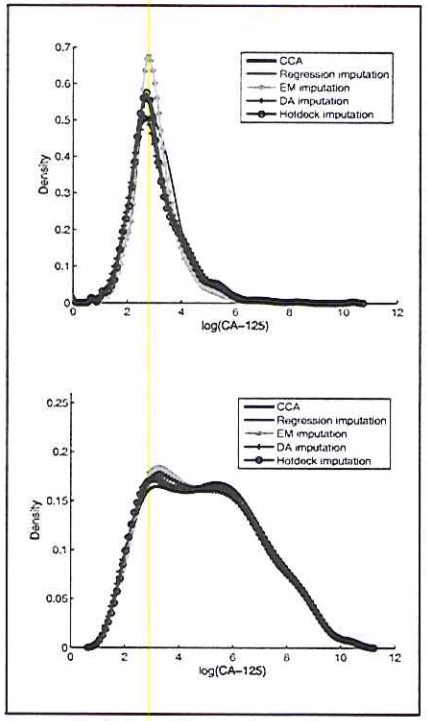


Figure 1 Distribution of CA-125 in the complete and imputed data sets.

RESULTS

Preliminary Results

Of the 1066 tumors, 800 were benign (75%) and 266 were malignant (25%). CA-125 was unavailable in 233 patients with a benign tumor (29%) and in 24 of the patients with malignant tumor (9%). This reflects the fact that some clinical investigators did not measure CA-125 when the tumor looked clearly benign. Figure 1 presents the estimated distributions of CA-125 for complete cases as well as for all cases after imputation of missing values.

Using CCA ($n = 809$), CA-125 alone has an AUC of 0.826. The AUC slightly dropped when missing values were imputed ($n = 1066$): 0.807 for regression imputation, 0.821 for EM, 0.800 for DA, and 0.805 for hotdeck.

Diagnostic Models

The AUC results for the diagnostic models based on GLMs are given in Figures 2 and 3. The results when using LS-SVMs were highly similar; therefore, we have put the LS-SVM figures in the online appendix (Figures 4–5; see http://mdm/sagepub.com/supplemental). Figures 2 and 4 show the performance for the imputation scenarios in the original test set and after repeated data splitting. Figures 3 and 5 compare the results after regression imputation with the results based on CCA and CVA. As expected, the original test set results exhibit less performance stability than the results after repeated data splitting.

Differences between imputation methods. Focusing mainly on the repeated data-splitting results, there are minor, not meaningful, differences in AUC between the various imputation scenarios. De facto, regression imputation very often resulted in the best median test set AUC. For this reason, the results for the regression-imputed data set were shown in Figures 3 and 5.

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

Figure 2 Performance of generalized linear models (GLMs) in the original test set (left) and after repeated data splitting (right).

Figure 3 Performance of generalized linear models (GLMs) using regression imputation, complete case analysis (CCA), and complete variable analysis (CVA) in the original test set (left) and after repeated data splitting (right).

Differences between imputation and CCA. The CCA results were comparable with the results for the regression imputation scenario. After repeated data splitting, no meaningful differences in median test set AUC were observed. Yet, imputing missing values has the added benefits of making available the whole data set and alleviating bias when estimating particular parameters. Such bias is suggested by the parameter estimates obtained by fitting GLM models on all data. Table 2 presents the parameter

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

Table 2 Generalized Linear Model (GLM) Parameter Estimates (β) for CCA Analysis and the 4 Missing Value Imputation Methods

Variable*	CCA β (SE)	Regression β (SE)	EM β (SE)	DA β (SE)	Hotdeck β (SE)	% Difference between Imputation Methods and CCA
Intercept	0.834	0.694	1.03	0.827	0.402	
Unilocular-solid tumor[b]	−0.049 (0.143)	−0.107 (0.138)	−0.081 (0.132)	−0.102 (0.132)	−0.110 (0.139)	+66 to +125
Age, years	0.527 (0.306)	0.809 (0.303)	0.795 (0.301)	0.823 (0.296)	0.812 (0.294)	+51 to +56
Pain[b]	−0.266 (0.140)	−0.355 (0.132)	−0.355 (0.132)	−0.342 (0.130)	−0.334 (0.130)	+26 to +34
Maximal diameter of lesion, mm	0.915 (0.427)	1.17 (0.403)	1.17 (0.402)	1.14 (0.388)	1.12 (0.382)	+22 to +28
Multilocular-solid tumor[b]	−0.247 (0.136)	−0.328 (0.130)	−0.301 (0.124)	−0.299 (0.123)	−0.317 (0.130)	+21 to +33
Maximal diameter solid part, mm	1.06 (0.436)	1.26 (0.474)	1.25 (0.468)	1.29 (0.460)	1.27 (0.446)	+18 to +22
Acoustic shadows[b]	−0.668 (0.346)	−0.871 (0.308)	−0.768 (0.306)	−0.789 (0.301)	−0.808 (0.286)	+15 to +30
Unilocular tumor[b]	−0.791 (0.287)	−0.960 (0.287)	−0.908 (0.283)	−0.900 (0.271)	−0.929 (0.263)	+14 to +21
Irregular cyst walls[b]	0.519 (0.162)	0.538 (0.148)	0.545 (0.148)	0.561 (0.145)	0.572 (0.144)	+4 to +10
Entirely solid tumor[b]	0.415 (0.163)	0.422 (0.155)	0.450 (0.147)	0.466 (0.145)	0.415 (0.154)	+0.2 to +12
Hormonal therapy[b]	−0.303 (0.144)	−0.297 (0.133)	−0.293 (0.133)	−0.305 (0.132)	−0.319 (0.132)	−3 to +5
Ascites[b]	0.683 (0.182)	0.647 (0.168)	0.639 (0.169)	0.686 (0.165)	0.734 (0.163)	−7 to +7
Postmenopausal bleeding[b]	0.219 (0.228)	0.212 (0.195)	0.218 (0.195)	0.221 (0.192)	0.177 (0.192)	−19 to +1
History of ovarian cancer[b]	0.628 (0.416)	0.694 (0.352)	0.670 (0.350)	0.631 (0.343)	0.520 (0.355)	−17 to +11
Log(CA-125), U/mL	1.66 (0.431)	1.94 (0.414)	1.87 (0.397)	1.44 (0.331)	1.37 (0.329)	−26 to +5
Papillary blood flow[b]	0.396 (0.187)	0.344 (0.159)	0.350 (0.159)	0.369 (0.157)	0.355 (0.157)	−7 to −13
Color score	0.791 (0.165)	0.673 (0.149)	0.671 (0.148)	0.684 (0.146)	0.681 (0.145)	−14 to −15
Number of papillary projections	0.430 (0.231)	0.374 (0.207)	0.378 (0.206)	0.367 (0.204)	0.360 (0.201)	−12 to −16
Lesion unequal to ovary[b]	−0.248 (0.194)	−0.201 (0.162)	−0.201 (0.163)	−0.231 (0.162)	−0.235 (0.159)	−5 to −19
Multilocular tumor[b]	−0.245 (0.174)	−0.164 (0.161)	−0.128 (0.159)	−0.129 (0.155)	−0.150 (0.158)	−33 to −48

Standard errors (SE) were estimated using 1000 bootstrap samples. The percentage differences between the imputation-based estimates and the CCA estimate are presented as a range.
CCA, complete case analysis; DA, data augmentation; EM, expectation-maximization.
a. Ordinal and continuous variables were scaled to the [−1, +1] interval.
b. Binary variables: −1 = no, +1 = yes.

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

estimates as well as the percent differences of the CCA estimates with the estimates obtained after missing value imputation. For 5 variables, all imputation-based estimates suggested that CCA had underestimated the value by at least 20%. For 1 variable, all imputation-based estimates suggested overestimation of at least 20% when CCA was used. Furthermore, the variable rankings also differed between CCA and the imputation scenarios, suggesting that different variables would be selected after missing value imputation. Variable rankings for the 4 imputation scenarios correlated between 0.97 and 0.998, whereas the variable ranking based on CCA correlated only between 0.73 and 0.79 with the rankings after imputation.

Necessity of CA-125. An important clinical question is whether CA-125 is indispensable when constructing diagnostic models for ovarian tumors. Focusing on the repeated data-splitting figures, missing value imputation resulted in slightly higher AUCs compared with CVA. Difference in median test set AUC between the regression imputation scenario and CVA is never larger than 0.01, which is clinically nonsignificant.

DISCUSSION

Missing values are frequently encountered in medical research and can result in biased inferences and in suboptimal performance of diagnostic models. In this study, we investigated the use of CA-125 for predicting ovarian tumor malignancy in a data set with 24% missing values for this tumor marker. Next to analyzing complete cases, we also imputed the missing CA-125 values using 4 single imputation techniques. Diagnostic models were developed using LS-SVMs and GLMs and were evaluated using the test set AUC. We found that missing value imputation and CCA resulted in comparable AUCs with minor AUC differences between the 4 imputation scenarios. However, CCA clearly differed from imputation regarding model parameter estimates and variable selection. In addition, the use of imputation resulted in minor, clinically nonsignificant improvements over CVA. We will now elaborate on these findings and their implications.

First, all 4 imputation methods led to diagnostic models with very similar AUCs. Yet, regression imputation often resulted in slightly higher AUCs de facto. This does not mean that this imputation method produced the most accurate imputations, however. Imputation accuracy could not be verified because the true CA-125 values for missing observations were not known, and hence it is hard to conclude which imputation method was preferred for our application. Moreover, the AUC performance is not the optimal index to investigate imputation accuracy. Methods such as EM and DA are superior in theory.[22] For example, EM and DA imputations take into account uncertainty in the imputations by averaging over the distribution for the missing values.

Second, it was also observed that CCA and missing value imputation resulted in similar AUCs. Notwithstanding this result, the obtained GLM model parameters suggested that imputation was preferred over CCA. For 6 of 20 variables, all 4 imputation methods suggested that CCA overestimated (or underestimated) the model parameter by at least 20%. Also, CCA resulted in different variable selection results compared with the imputation scenarios. This suggests that missing value imputation has alleviated bias when estimating particular parameters. Note in this respect that this is conditional on the use of a reasonable imputation model—that is, a rich model including variables related to the imputed variable, variables related to the missingness of the imputed variable, and variables of interest in the analysis following imputation.[12,23]

Third, because the advantage in AUC of models including CA-125 over models excluding CA-125 (CVA) was very small, we conclude that it is not necessary to use this marker in diagnostic models for ovarian tumor malignancy. This is an important clinical message that is consistent with earlier findings based on CCA using logistic regression modeling.[3] Our results strengthen these findings because we imputed missing values rather than omitting them and because we also used the flexible and regularized LS-SVM classifier to build diagnostic models.

Recently, it has been suggested to use the outcome when imputing missing values.[24] Even though this may seem a "self-fulfilling prophecy," it makes the imputation model more reasonable because otherwise, the imputation model assumes that the association between the outcome and the variable that is being imputed is zero. Therefore, we added the outcome to the regression imputation model to obtain new imputations. Redoing the analyses using GLMs did not alter the conclusions: the median test set AUCs hardly increased and were at most 0.01 higher than those based on CVA. This is most likely due to the fact that the imputation model used several other variables that were predictive of malignancy.

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010

REFERENCES

1. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. CA Cancer J Clin. 2008;58:71–96.
2. Vergote I, De Brabanter J, Fyles A, Bertelsen K, Einhorn N, Sevelda P. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. Lancet. 2001;357:176–82.
3. Timmerman D, Van Calster B, Jurkovic D, Valentin L, Testa AC, Bernard J-P. Inclusion of CA-125 does not improve mathematical models developed to distinguish between benign and malignant adnexal tumors. J Clin Oncol. 2007;25:4194–200.
4. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004;91:4–8.
5. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley; 1987.
6. Van der Heijden GJMG, Donders ART, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol. 2006;59:1102–9.
7. Perez A, Dennis RJ, Gil JFA, Rondon MA, Lopez A. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Stat Med. 2002;21:3885–96.
8. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. Ultrasound Obstet Gynecol. 2000;16:500–5.
9. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML. A new logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumour Analysis (IOTA) Group. J Clin Oncol. 2005;23:8794–801.
10. Little RJA, Rubin DB. Statistical Analysis With Missing Data. 2nd ed. New York: John Wiley; 2002.
11. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc. 1977;39:1–38.
12. Schafer JL. Analysis of Incomplete Multivariate Data. London: Chapman & Hall; 1997.
13. Ford BL. An overview of hot-deck procedures. In: Madow WG, Olkin I, Rubin DB, eds. Incomplete Data in Sample Surveys. New York: Academic Press; 1983. p 185–207.
14. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. Multivariate Behav Res. 1998;33:545–71.
15. Bernaards CA, Farmer MM, Qi K, Dulai GS, Ganz PA, Kahn KL. Comparison of two multiple imputation procedures in a cancer screening survey. J Data Sci. 2003;1:293–312.
16. Schneider T. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J Climate. 2001;14:853–71.
17. Chen J, Shao J. Nearest neighbor imputation for survey data. J Off Stat. 2000;16:113–31.
18. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least Squares Support Vector Machines. Singapore: World Scientific; 2002.
19. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. New York: John Wiley; 2000.
20. MacKay DJC. Probable networks and plausible predictions: a review of practical Bayesian methods for supervised neural networks. Network. 1995;6:469–505.
21. Van Calster B, Timmerman D, Lu C, Suykens JAK, Valentin L, Holsbeke CV. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. Ultrasound Obstet Gynecol. 2007;29:496–504.
22. Schafer J, Graham J. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147–77.
23. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc. 1996;91:473–89.
24. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59:1092–101.

Downloaded from http://mdm.sagepub.com at Katholieke Universiteit Leuven on February 1, 2010