http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (6): 1125-1133, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.1125.1133

Efficient Missing Data Technique for Prediction of Nasopharyngeal Carcinoma Recurrence

[1]Panrasee Ritthipravat, [1]Orrawan Kumdee and [2]Thongchai Bhongmakapat
[1]Department of Biomedical Engineering, Faculty of Engineering, Mahidol University,
999 Puttamonthon 4, Salaya, Nakornpathom, Thailand
[2]Department of Otolaryngology, Faculty of Medicine, Ramathibodi Hospital, Bangkok, Thailand

Abstract: This study aims to investigate efficient missing data techniques for prediction of nasopharyngeal carcinoma (NPC) recurrence. Initially, clinical data of patients with NPC who received treatment at Ramathibodi hospital, Thailand, were collected. In total, 495 records were employed for the cancer recurrence prediction. Due to the fact that these data contain different missing values, appropriate missing data techniques (MDTs) must be examined. In this study, complete-case analysis, mean imputation, k-nearest neighbor imputation and Expectation Maximization (EM) imputation are mainly focused. The completed data are then used for developing three different predictive models, i.e., single-point model, multiple-point model and sequential neural network. The experimental results showed that EM imputation was superior to the other missing data techniques in which it provided highest predictive performance in all models. The average area under the receiver operating characteristic curve (AUC) of 0.72 could be achieved. The Hosmer and Lemeshow goodness of fit test was used for evaluating goodness of fit of each model. The results confirmed that EM imputation was the best missing data technique. The sequential neural network outperformed the other models. It provided the highest predictive performances in terms of the average AUC (0.73) and the Chi-square statistic (4.30). In addition, survival curves generated from these predictive models were compared with that of the Kaplan-Meier survival curve. The curves based on EM imputation were closest to the Kaplan-Meier model. From the log-rank test, however, these curves were significantly different (p-value < 0.05).

Key words: Missing data techniques, sequential neural network, prediction of nasopharyngeal carcinoma recurrence

INTRODUCTION

Nasopharyngeal carcinoma (NPC) is one of common head and neck cancers generally found in Southern part of China, Hong Kong, Taiwan and Thailand (Devi *et al*., 2004). In Europe and America, the annual incidence is low, only 1 in every 100,000 white children. NPC is difficult to early detect because it occurs in the nasopharynx which is not simply accessible. When it is discovered, NPC has already spread to regional lymph nodes. Initially observed symptoms include existences of a painless neck mass, serous otitis and nasal obstruction with bloody drainage etc. Various treatments can be applied such as chemotherapy, radiotherapy, or combination of these treatments. Though NPC can be cured, it may redevelop after the treatment. Patients must be followed up regularly in order to detect the recurring cancer. Frequently checking is necessary for high-risk patient group. However, this may take time and incur unnecessary expenses for low-risk patient group. Prediction of NPC recurrence for each patient is thus necessary in which the patient can realize his/her health condition and can make decision about the treatment. In addition, hospital resources, time and money can be effectively planned and managed.

To predict the presence or absence of cancer recurrence, related clinical data and recurring time are collected in the retrospective manner. Because these data may be incompletely recorded, efficient Missing Data Techniques (MDTs) must be investigated. From previous research, there are two main approaches for treating missing data (Nelwamondo *et al*., 2007; Schafer and Graham, 2002; Magnani, 2004; D'Agostino, 2007; Acock, 2005; Little and Rubin, 2002). They are data deletion and data imputation. For the deletion technique, all incomplete records are excluded from the analysis. This is different from the imputation technique in that either statistical methods or Artificial Intelligence (AI) based

Corresponding Author: Panrasee Ritthipravat, Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, 999 Puttamonthon 4, Salaya, Nakornpathom, Thailand Tel: +66 28892138 ext. 6351

techniques are used to assign values for the missing data. Both approaches have been used in various research areas, such as data classification (Acuna and Rodriguez, 2004; Penny and Chesney, 2006; Ennett *et al*., 2001), prediction (Jerez *et al*., 2006), epidemiologic studies (Barzi and Woodward, 2004) etc. Acuna and Rodriguez (2004) investigated four missing data techniques, i.e., complete case analysis, mean imputation, median imputation and K-Nearest Neighbor (KNN) imputation in a classification problem. These techniques were not considerably different when the number of missing records was small. With the increased number of missing data, KNN imputation was superior to the others. Jerez *et al*. (2006) compared three MDTs in prediction of breast cancer relapse. The techniques included complete case analysis, mean imputation and hot-deck imputation. Both imputation techniques were superior to the deletion. Barzi and Woodward (2004) studied many techniques for dealing with missing values of total cholesterol in 28 cohort studies. The techniques included complete case analysis, several imputation methods and multiple imputations. Multiple imputations were superior to the other techniques, even though, 60% of data were missing. However, they were relatively complicated and required specific software for implementation.

As seen above, several methods have been employed to treat the missing data. Proper techniques are task dependent. In the previous study of NPC recurrence prediction (Kumdee *et al*., 2008), four different missing data techniques were applied, i.e., complete-case analysis, mean imputation, KNN imputation and EM imputation. The results showed that EM imputation was superior to the other techniques when tested with the single-point model. In this study, different predictive models commonly used in medical prognosis (Ohno-Mochado, 2001; Lin *et al*., 2008; Ohno-Machado and Musen, 1997; Park and Edington, 2001; De Laurentiis *et al*., 1999) are investigated. The models are multiple-point model and sequential neural network. The most efficient missing data technique is reinvestigated. In addition, the best predictive model is examined.

MATERIALS AND METHODS

Patients' data: In this study, clinical data of patients with nasopharyngeal carcinoma who were treated at Ramathibodi Hospital, Thailand, during the period 1982-2007 were collected and recorded. This research has been conducted under the approval of the Ethics Committee of Faculty of Medicine, Ramathibodi hospital, Mahidol University. The study period was set at 5 years for each patient. In total, 495 records were taken into consideration. Within the 5-year study period, 125 patients had redeveloping cancer while 159 patients were lost to follow up or withdrew from the study. The others (211 patients) had no cancer recurrence. The collected data composed of 13 prognostic factors selected by an expert medical doctor. These factors included (1) Age, (2) Sex, (3) Duration of the firstly observed symptoms before checkup at the Otolaryngology department (4-6) TNM stages (T: Tumor stage, N: Node involvement and M: Metastasis), (7) Cancer stage (Stage I-IV), (8) Cell type determined from the biopsy, (9) IgG quantity, (10) IgA quantity, (11) Chemotherapy (yes or no), (12) Dose of radiation and (13) Presence of neck fibrosis (yes or no). In addition, either recurring time or censoring time was collected for each patient. Table 1 presents a list of all prognostic factors including range, mean, median, standard deviation and the number of missing records. Table 2 shows the numbers of patients who withdrew or were lost to follow up and those who had and did not have the cancer relapse within a specific time-point.

As seen in Table 1 and 2, both missing records and censored data are encountered in the analysis. Previous research showed that these types of missing data were treated differently. Missing data techniques commonly focused on either data deletion or imputation methods (Nelwamondo *et al*., 2007; Schafer and Graham, 2002; Magnani, 2004; D'Agostino, 2007; Acock, 2005; Little and Rubin, 2002). For censoring data, modifications of a predictive model were mostly exploited (De Laurentiis *et al*., 1999; Baesens *et al*., 2004; Ravdin and Clark, 1992). In this study, the most efficient missing data technique for NPC recurrence prediction is merely investigated.

Missing data techniques: Missing data problem is usually found in a retrospective study. The missing records, in general, can arise from various reasons. For example, a patient may not answer a certain question because he/she may not know the required information (e.g., he/she may not be able to correctly specify the time when the symptoms were firstly observed). In some cases, missing data occur from inconsistent measures, e.g., only some patients were asked about the presence of the neck fibrosis. Measurement error and recorder error are also the other causes of missing records. Normally, missing data are categorized into 3 types according to the missing mechanisms (Nelwamondo *et al*., 2007; Schafer and Graham, 2002; Magnani, 2004;

Table 1: All prognostic factors from 495 patients that used in the analysis and their range, mean, median, standard deviation and the number of missing records

Factor	Range/coding	Mean	Median	SD*	No. of missing records
Age (years)	10-84	45.67	46	12.64	1
Duration time (days)	0-3600	159.90	90	257.64	57
IgG quantity (1:10-1:2560)	10-2560	159.42	160	191.44	329
IgA quantity (1:10-1:640)	10-640	58.83	20	79.09	264
Dose of radiation (cGy)	0-8000	6846.17	7000	868.45	41
Sex	0: male	0.37	0	0.48	0
	1: female				
T stage* (T1-T4)	1000: T1	2.72	3	1.06	3
	0100: T2				
	0010: T3				
	0001: T4				
N stage* (N0-N3)	1000: N0	1.58	2	1.00	3
	0100: N1				
	0010: N2				
	0001: N3				
M stage* (M0,M1,Mx)	100: M0 = no metastasis	0.18	0	0.53	3
	010: M1 = metastasis				
	001: Mx = unidentified				
Cancer stage (I-IV)	1000: stage 1	3.54	4	0.81	5
	0100: Stage 2				
	0010: stage 3				
	0001: stage 4				
Cell types	1000: type 1	2.35	2	0.60	18
	0100: type 2				
	0010: type 3				
	0001: type 4				
Chemotherapy	0: absence	0.83	1	0.38	37
	1: presence				
Neck fibrosis	0: absence	0.88	1	0.32	130
	1: presence				

*SD: Standard deviation, *T stage: Size of primary tumor and whether it has invaded nearby tissue, *N stage: Regional lymph node involvement and *M stage: Metastasis

Table 2: The numbers of patients who withdrew or were lost to follow up and those who had and did not have the cancer relapse within a specific time-point

Time (years)	Withdrew/lost to follow up	Recurrence	No recurrence
0-1	51	65	379
0-2	99	96	300
0-3	122	112	261
0-4	140	118	237
0-5	159	125	211

D'Agostino, 2007; Acock, 2005; Little and Rubin, 2002). These types include Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing at Random (NMAR). For MCAR data, the missing information is either unrelated to its value or the value of the other variables. This represents the fact that the probability of being missing data is identical for every record. In this case, excluding the incomplete records provides unbiased results. For MAR data, the missing information is only related to the observable values of the variables. In this case, the data being missing can be directly estimated from the measurable data. The last type of the missingness is not missing at random (NMAR). It describes the case that the missing information depends on its values. It may arise when the collected data are measured from a sensor that cannot detect values at a particular range. NMAR is the most problematic type because the missing data cannot be easily modeled or guessed from the existing data.

Appropriate missing data technique can be selected when the missing mechanism is correctly identified, for instance, excluding the incomplete records can be performed without biased analysis for MCAR data. Imputation can be used when the missing data are MAR. However, it is not easy to test that the type of missing data is MCAR, MAR, or NMAR. In fact, the proper missing data technique is application dependent. As presented previously, there are 2 main approaches for dealing with missing data. The first approach is to disregard the missing records from the analysis. The second approach is to approximate value of missing data from non-missing attributes. In this study, four techniques which are in both approaches are considered. They are presented as follows.

Complete case analysis (CCA): As the name implies, only complete records are taken into consideration. All missing cases are removed from the analysis. This is the simplest technique and available in many statistical software packages. When the number of missing data is huge, however, deletion can drastically reduce the size of data.

The remaining data may be insufficient for the study. In case that the collected data do not have complete records, CCA cannot be applied (Lakshminarayan *et al*., 1999). CCA can be appropriately used when the missing data is MCAR (Little and Rubin, 2002). For MAR and NMAR data, the use of CCA may deviate statistical values of the data, such as mean, standard deviation etc. Therefore, care must be taken when the complete case analysis is exploited.

Mean imputation (MI): Rather than removing the incomplete record, mean of a missing variable is calculated from its non-missing values and used in place of the missing values of that variable. This technique is frequently employed because it is simple and straightforward. However, distribution of the imputed data is usually different from the actual one. The standard deviation of the variable after the imputation is normally underestimated though the missing data are MCAR. The shortcoming occurs from adding the constant mean value. Dispersion of the missing values is thus not taken into account.

KNN imputation: This technique estimates value of a missing data from its k neighboring records. The neighborhood can be determined from a distance function, such as Euclidean distance, Mahalanobis distance etc. The k neighbors are selected from the records that are most similar to the missing data of interest. In this study, Euclidean distance is employed. The neighboring records are weighted with the values that are inversely proportional to the distances. Sum of the weighted records is used in place of the missing data. The KNN imputation does not require modeling of the missing data. On the contrary, it uses information from the similar observations. For the large database, however, searching the k-nearest neighbors may relatively take time.

EM imputation: EM algorithm is a likelihood-based modeling procedure. It tries to estimate the distribution model of the variables. In the analysis, multivariate normal model is assumed (Acuna and Rodriguez, 2004). In general, this technique composes of two-step iteration, i.e., expectation (E-step) and maximization (M-step). The E-step computes an expectation of the log likelihood function determined based on the current estimation of the distribution of the existing data. The M-step tries to find parameters that maximize the expected log likelihood found from the E-step. These two steps run repeatedly until the expected log likelihood function is maximized.

These four missing data techniques are applied to treat missing values from the collected data. The completed data are used in NPC recurrence prediction. In this study, three predictive models presented in the next section are focused.

PREDICTION OF NPC RECURRENCE

Three predictive models based on artificial neural network are mainly investigated in this study. They are single-point model, multiple-point model and sequential neural network. These models are commonly used in cancer prognosis. To predict the presence or absence of cancer recurrence, type I censoring observations can be used directly. For type II censoring data, they are handled differently in each model.

Single-point model: This model provides the prediction of cancer recurrence within a specific period (Ohno-Mochado, 2001) as shown in Fig. 1. In the Fig. 1, prognostic factors are used as inputs to the model. Output, ranging from 0 to 1, represents the recurrence probability within the first year of treatment (year 1). Recurrence status is used as the training target by setting to 0 for no redeveloping cancer and 1 for cancer relapse. As an example, the target vector for a patient with cancer relapse at 3.5 years is set to [0 0 0 1 1]. For the relapse-free patient (type I censoring data), the target vector is [0 0 0 0 0]. Type 2 censoring data are used until their censoring time. The number of training data is thus gradually reduced over time. This may cause unreliable predictions at distant time (Baesens *et al*., 2004). Multilayer perceptron with back-propagation training is used to provide mapping between the inputs and the output. For cancer management, in general, it normally considers the cancer relapse within 5 years. Therefore, 5 single-point models are developed in this study.

Though this model is relatively simple, several models must be generated. In case that there are multiple time-points of interest, it is very time-consuming and not scalable well. In addition, inconsistent predictions may

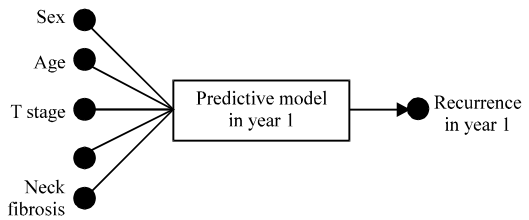

Fig. 1: The single-point model provides the prediction of cancer recurrence within a specified period

occur because the interdependencies among the predictions at different time-points are not taken into account. Survival curve gained by aggregating the single-point models is more likely to be non-monotonic. Ohno-Machado and Musen (1997) had shown that the single-point model could not differentiate patterns of disease progression and might be unreliable for long-term prediction.

Multiple-point model: Rather than separately considering several predictive models, multiple-output neural network is used as shown in Fig. 2. Similar to the single point model, recurrence status is used as the training target by setting to 0 as long as the redeveloping cancer is not found and 1 for another case. Training targets for patients who have or do not have the cancer recurrence within the study period have can be set directly. However, for type II censoring observation, his/her recurrence status after the censoring time cannot correctly specify. Probability that the patient relapses before a time-point of interest, F(t), is then used in place of the unknown target. This probability is determined from 1-S(t) where S(t) is the probability of relapse-free gained from the Kaplan-Meier model (Bewick *et al*., 2004a). For example, the training target of a patient who is lost to follow-up after 3 years of treatment is [0 0 0 0.23 0.29] where the last two terms represent the probabilities that the patient relapses before 4 and 5 years, respectively. In doing so, all type 2 censoring data can be used in the prediction. However, the probabilities of cancer relapse may be underestimated because the Kaplan-Meier model provides the relapse-free probabilities of the whole patients, not for individuals. In addition, monotonic survival curve cannot be ensured.

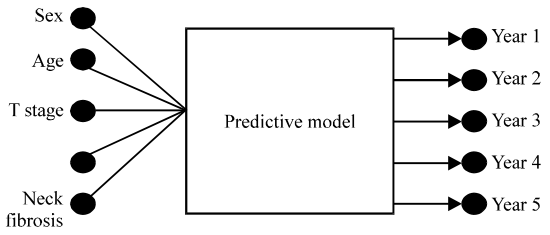

Fig. 2: The multiple-point model provides the recurrence prediction at several time-points

Sequential neural network: From the previous predictive models, interdependencies among multiple predictions are not taken into consideration. Ohno-Mochado (2001) introduced a sequential neural network that composes of multiple neural networks connected sequentially. The output of a network is added as an additional input of the subsequent network as presented in Fig. 3. By doing this, interdependencies of predictions at several time-points can be seamlessly integrated. Survival curve generated by this technique tends to gradually decrease over time. Training targets are set similar to the single point model. Therefore, type II censoring observations can be used until their censoring time.

PERFORMANCE EVALUATION

Predictive performances of all models are evaluated by three different measures, i.e., area under the receiver operating characteristic curve (AUC), Hosmer-Lemeshow goodness of fit test and the logrank test:

- Area under the receiver operating characteristic (AUC) curve (Bewick *et al*., 2004b) is used to measure accuracy of the predictive models. The perfect predictors provide AUC of 1 whereas the random predictor has AUC of 0.5
- The goodness of fit test (Lin *et al*., 2008; Lemeshow and Hosmer, 2000) is a way to assess the model fit. It sorts the predicted recurrence probabilities and separates them into 10 groups. The Pearson chi-square statistic is then computed based

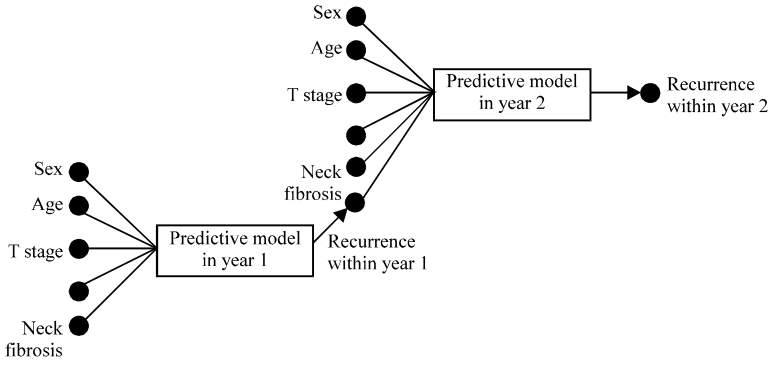

Fig. 3: Sequential neural network connects multiple recurrence predictive models sequentially

on the difference of expected and observed recurrences in each group. In this study, the statistic value is compared to the Chi-square distribution with 8 degrees of freedom. The statistic value smaller than 15.51 indicates that the model fits well

- The logrank test (Bland and Altman, 2004) is used to compare whether the survival curve generated from each predictive model is significantly different from the survival curve from the Kaplan-Meier model. The survival curve of the model is created from the predicted outputs. Initially, the best cut-point at a year is determined from the area under the ROC curve. Recurring time of a patient is determined from the first output that provides the value above its best cut-point. For example, the cut-point of a model is assumed to be [0.3 0.4 0.5 0.2 0.1]. The model predicts the cancer recurrence probability of a patient as [0.1 0.3 0.7 0.1 0.5]. The recurring time of this patient is then after 2 years. After the recurring time of all patients is determined, the survival curve of the model is then generated using the Kaplan-Meier estimator. The actual and the predicted survival curves are then compared with the logrank test. The p-value larger than 0.05 indicates that there is no difference between these curves

EXPERIMENTS

Completed data gained from four missing data techniques were employed to predict nasopharyngeal carcinoma recurrence. The techniques were Complete Case Analysis (CCA), imputations by mean (MEAN), K-Nearest Neighbor (KNN) and Expectation Maximization (EM). For CCA, only 80 complete records were used in the prediction. For the imputation approaches, all 495 records could be used. The number of neighboring records was set to 5 for KNN imputation. Three predictive models based on multilayer perceptron with back-propagation training, i.e., single-point model, multiple-point model and sequential neural network, were used in the comparison.

In each model development, the completed data were separated into 2 groups, 80% for model generation and 20% for model validation. For the model generation, ten-fold cross validation was employed. The model was trained on 90% of the separated data and tested on the others 10% at every 1000 epochs of training. The training was stopped when achieving the smallest testing set error. In the study, learning rate was varied from 0.05, 0.1-0.5. Momentum was tuned from 0.6 - 0.9. Hidden nodes were adjusted from 1-5 nodes. The best parameter setting was selected. The best model was then validated with the rest 20% of the completed data. Predictive performances from the validation set were used in comparison.

RESULTS

The results of area under the receiver operating characteristic curve (AUC) averaged over 5 years are summarized in Table 3. In every predictive model, all imputation techniques were superior to the deletion. EM imputation was the most efficient MDT in which it provided highest average AUC whichever predictive model is used. The average AUC of EM imputation (0.72) was higher than KNN (0.69) and MEAN (0.60). CCA provided the lowest average AUC (0.53). These results corresponded to the Hosmer and Lemeshow goodness of fit test presented in Table 4. The average Chi-square statistic of EM (5.57) was smaller than that of KNN (6.55) and MEAN (6.87) while CCA was the worst missing data technique (7.42). Normally, the smaller value of the Chi-square statistic presents the better fit of the model. The sequential neural network provided the highest predictive performances in both average AUC (0.6425) and the Chi-square statistic (5.77). The second best predictive model was multiple-point model with the average AUC of 0.625 and the average Chi-square statistic of 6.05.

Survival curves generated from these models were compared with that of the Kaplan-Meier estimator as shown in Fig. 4-6. The logrank test was applied and the results were summarized in Table 5. All survival curves based on EM imputation were closest to the Kaplan-Meier model, even though the p-values from the logrank test showed that these curves were significantly different ($p<0.05$).

Table 3: Comparison of average area under the receiver operating characteristic curve (AUC) from different missing data techniques

Missing data techniques	Predictive models			
	Single	Multiple	Sequential	Avg.
CCA	0.520	0.550	0.5100	0.53
MEAN	0.610	0.580	0.6100	0.60
KNN	0.690	0.660	0.7200	0.69
EM	0.720	0.710	0.7300	0.72
Avg.	0.635	0.625	0.6425	

CCA: Complete case analysis, MEAN: Mean imputation, KNN: K-nearest neighbor imputation and EM: Expectation maximization

Table 4: Comparison of average chi-square statistics for investigating the goodness of model fit

Missing data techniques	Predictive models			
	Single	Multiple	Sequential	Avg.
CCA	7.88	6.56	7.83	7.42
MEAN	8.69	7.73	4.20	6.87
KNN	5.93	6.97	6.76	6.55
EM	9.49	2.92	4.30	5.57
Avg.	8.00	6.05	5.77	

Chi-square statistic>15.51 indicates the poor fit, *CCA: Complete case analysis, MEAN: Mean imputation, KNN: K-nearest neighbor imputation, EM: Expectation maximization

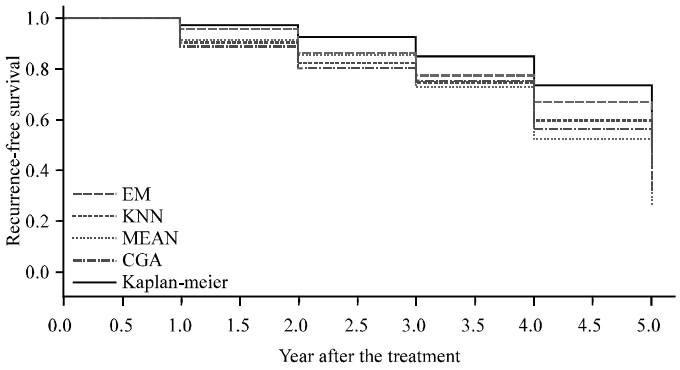

Fig. 4: Recurrence-free survival curves generated from single-point model

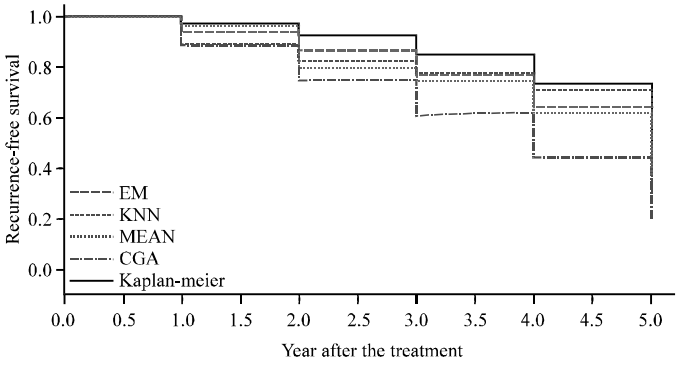

Fig. 5: Recurrence-free survival curves generated from multiple-point Model

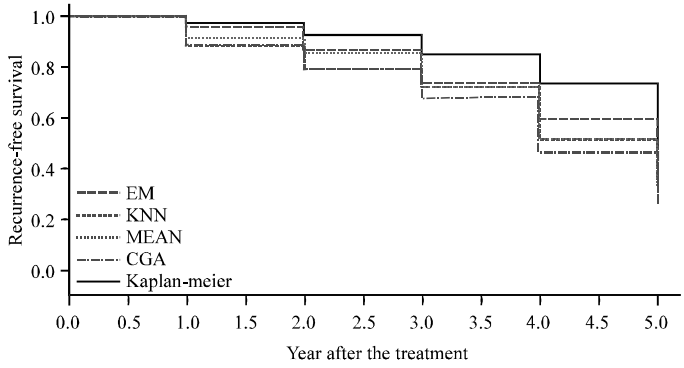

Fig. 6: Recurrence-free survival curves generated from sequential neural network

Table 5: Results of the logrank test for survival curve comparison

Missing data techniques	Predictive models			
	Single	Multiple	Sequential	Avg.
CCA*	0.0029	<0.0001	0.0002	0.0010
MEAN*	<0.0001	<0.0001	<0.0001	<0.0001
KNN*	<0.0001	0.0008	<0.0001	0.0003
EM*	0.0357	0.0142	0.0003	0.0167
Avg.	0.00965	0.00375	0.000125	

*CCA: Complete case analysis, *MEAN: Mean imputation, *KNN: K-nearest neighbor imputation and *EM: Expectation maximization

DISCUSSION

In the study, the most efficient missing data technique in the prediction of nasopharyngeal carcinoma recurrence is mainly investigated. The results showed that the Complete Case Analysis (CCA) provided the lowest performances in every predictive model. This is not surprising because the existing data after the deletion were only one-fifth of the total records. Poor

performances may arise from either the number of the existing data is not sufficient to generate the accurate models or the excluded data contain important information about the NPC recurrence. This is different from the imputation techniques in which the sample size can be maintained. Missing values are approximated from the available collected data using different approaches.

EM technique provided the best estimated missing values because it tries to fit models to the incomplete data. It is a global estimation that takes both the mean and data dispersion of each variable into account. This result conforms to the previous studies (Azen *et al*., 1989; Musil *et al*., 2002 ; Karadogan *et al*., 2011) in which EM imputation was outperform to various missing data techniques.

For KNN imputation, it is the second best technique. KNN is a local imputation method. It only takes information from the k-nearest complete records into account. Therefore, information from the other records is not used in estimating the missing values. In case that there exist many missing attributes in an incomplete record, the estimated missing values may be inaccurate.

Imputation by mean value was poorer than the other imputation techniques. This indicates that the completed data tend to be biased and cannot be used to generate accurate predictive models. In addition, the mean value is sensitive to the presence of outliers.

For predictive models, sequential neural network is superior to the others because interdependencies among predictions are taken into consideration. Output of the prediction from a given interval can be used to improve the prediction of the subsequent interval. However, multiple models must be generated. This is impractical particularly when various time-point predictions are required. This model also suffers from the decreasing number of patients over time. This is because type II censoring data can be used until their latest follow-up time. In the study, 159 censoring records were disregarded from the study at 5 year.

For the multiple-point model, probability that the patient relapses before a time-point of interest, F(t), is used in place of unknown target for type II censoring data. Accurate prediction could be achieved, even though this probability is derived from the Kaplan-Meier model which provides the prediction for a group of patients, not for individuals. All 495 records could be used in the study. From AUC and Chi-square statistic comparisons, the predictive performances of the multiple-point model were slightly lower than that of the sequential neural network. However, the model can simultaneously provide predictions at different time-points within a single model.

For the single-point model, its shortcomings are similar to the sequential neural network in which multiple models must be generated. The number of cases reduces over time. Though the logrank test showed that the single-point model provided the survival curve closest to the actual one, this model was slightly inferior to the sequential neural network. For survival curve comparison, the curve may not truly represent the predictive performance of its model. This is because the survival curve is created from the number of relapse-free patients. The models may provide correct prediction about this information. However, they may incorrectly identify which patients do not have cancer recurrence at a time.

From these predictive models, relapse-free survival function of a patient cannot be guaranteed to be monotonically decreasing curve. For overall patients, recurrence-free probability was approximately 70% at 5 years.

CONCLUSION

This study has investigated the efficient missing data technique for NPC recurrence prediction problem. Three predictive models, i.e., single-point model, multiple-point model and sequential neural network, are used in the investigation. The results showed that EM imputation was superior to the other missing data techniques particularly when the sequential neural network was employed. This is because models of missing data are appropriately fitted. Average AUC of 0.72 could be achieved.

ACKNOWLEDGMENT

This research was supported by Mahidol University.

REFERENCES

Acock, A.C., 2005. Working with missing values. J. Marriage Family, 67: 1012-1028.

Acuna, E. and C. Rodriguez, 2004. The Treatment of Missing Values and Its Effect in the Classifier Accuracy. In: Classification, Clustering and Data Mining Applications, Banks, D.L., F.R. McMorris, P. Arabie and W. Gaul, (Eds.). Springer, New York, pp: 639-648.

Azen, S.P., M. Van Guilder and M.A. Hill, 1989. Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomly incomplete data. Stat. Med., 8: 217-228.

Baesens, B., T.V. Gestel, M. Stepanova and D.V.D. Poel, 2004. Neural network survival analysis for personal loan data. Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium.

Barzi, F. and M. Woodward, 2004. Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. Am. J. Epidemiol., 160: 34-45.

Bewick, V., L. Cheek and J. Ball, 2004a. Statistics review 12: Survival analysis. Critical Care, 8: 389-394.

Bewick, V., L. Cheek and J. Ball, 2004b. Statistics review 13: Receiver operating characteristic curves. Critical Care, 8: 508-512.

Bland, J.M. and D.G. Altman, 2004. The logrank test. Br. Med. J., Vol. 328. 10.1136/bmj.328.7447.1073

D'Agostino Jr., R.B., 2007. Overview of missing data techniques. Methods Mol. Biol., 404: 339-352.

De Laurentiis, M., S. De Placido, A.R. Bianco, G.M. Clark and P.M. Ravdin, 1999. A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. Clin. Cancer Res., 5: 4133-4139.

Devi, B., P. Pisani, T.S. Tang and D.M. Parkin, 2004. High incidence of nasopharyngeal carcinoma in native people of Sarawak, Borneo Island. Cancer Epidemiol. Biomarkers Prev., 13: 482-486.

Ennett, C.M., M. Frize and C.R. Walker, 2001. Influence of missing values on artificial neural network performance. Stud. Health Technol. Inf., 84: 449-453.

Jerez, J.M., I. Molina, J.L. Subirats and L. Franco, 2006. Missing data imputation in breast cancer prognosis. Proceedings of the 24th IASTED International Conference on Biomedical Engineering, February 15-17, 2006, Innsbruck, Austria, pp: 323-328.

Karadogan, S.G., L. Marchegiani, L.K. Hansen and J. Larsen, 2011. How efficient is estimation with missing data? Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 22-27, 2011, Prague, Czech Republic, pp: 2260-2263.

Kumdee, O., P. Ritthipravat, T. Bhongmakapat and W. Cheewaruangroj, 2008. Dealing with missing values for effective prediction of NPC recurrence. Proceedings of the SICE International Conference on Instrumentation, Control and Information Technology, August 20-22, 2008, Tokyo, Japan, pp: 1290-1294.

Lakshminarayan, K., S.A. Harp and T. Samad, 1999. Imputation of missing data in industrial databases. Applied Intell., 11: 259-275.

Lemeshow, D.W. and S. Hosmer, 2000. Applied Logistic Regression. 2nd Edn., Wiley, New York, ISBN 0-471-35632-8.

Lin, R.S., S.D. Horn, J.F. Hurdle and R.A.S. Goldfarb, 2008. Single and multiple time-point prediction models in kidney transplant outcomes. J. Biomed. Inform., 41: 944-952.

Little, R.J.A. and D.B. Rubin, 2002. Statistical Analysis with Missing Data. 2nd Edn., Wiley, New York, ISBN-13: 9780471183860, Pages: 381.

Magnani, M., 2004. Techniques for dealing with missing data in knowledge discovery tasks. Department of Computer Science, University of Bologna, Italy, pp: 1-10.

Musil, C.M., C.B. Warner, P.K. Yobas and S.L. Jones, 2002. A comparison of imputation techniques for handling missing data. West. J. Nurs. Res., 24: 815-829.

Nelwamondo, F.V., S. Mohamed and T. Marwala, 2007. Missing data: A comparison of neural network and expectation maximization techniques. Curr. Sci., 93: 1514-1521.

Ohno-Machado, L. and M.A. Musen, 1997. Sequential versus standard neural networks for pattern recognition: An example using the domain of coronary heart disease. Comput. Biol. Med., 27: 267-281.

Ohno-Mochado, L., 2001. Modeling medical prognosis: Survival analysis techniques. J. Biomed. Inform., 34: 428-439.

Park, J. and D.W. Edington, 2001. A sequential neural network model for diabetes prediction. Artif. Intell. Med., 23: 277-293.

Penny, K.I. and T. Chesney, 2006. Imputation methods to deal with missing values when data mining trauma injury data. Proceedings of the 28th International Conference on Information Technology Interfaces, June 19-22, 2006, Cavtat, Dubrovnik, pp: 213-218.

Ravdin, P.M. and G.M. Clark, 1992. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. Breast Cancer Res. Treatment, 22: 285-293.

Schafer, J.L. and J.W. Graham, 2002. Missing data: Our view of the state of the art. Psychol. Methods, 7: 147-177.