Prognostic Model for Terminally Ill Cancer Patients on Laboratory Data

Kenji Hira, MD[1,2,3], Noriaki Aoki, MD[3,4], Akitoshi Hayashi, MD[5], Janez Demsar[6], Blaz Zupan, PhD[6,4], Kim Dunn, MD[2,3], William Jack Schull, PhD[2], Tsuguya Fukui, MD[1]

[1]Department of General Medicine and Clinical Epidemiology, Kyoto University Graduate School of Medicine, Kyoto, Japan, [2]The Schull Institute, Houston, Texas, [3]School of Health Information Sciences, University of Texas-Houston Health Science Center, Houston, Texas, [4]Information Research and Planning/ Department of Medicine, Baylor College of Medicine, Houston, Texas [5]Japan Baptist Hospital, Kyoto, Japan, [6]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Background

An estimate of prognosis of patients with terminal cancer is important for everybody involved in the case of these patients. An accurate estimation allows the appropriate planning of their medical care and might improve the patient's quality of dying. However, a prognostic model for such patients has not yet been established. In fact, several studies showed physicians' predictions of survival were often erroneous and optimistic [1,2].

The purpose of this study was to develop prognostication models for terminal cancer patients using selected data mining techniques.

Methods

A total of 311 patients with terminally ill cancer, who admitted to the hospice ward at the Japan Baptist Hospital from 1995 to 2000, were included in this analysis. Twenty-three variables including age, gender, blood count and serum laboratory measurements at the initial hospital admission were retrospectively collected. For model validation purposes, the data was split to training set (237 patients) and test set (74 patients).

Three statistical and data mining techniques, i.e., logistic regression analysis (LR), naive Bayesian (NB) and decision tree induction (TI) were used to develop prognostic models on the training data. All models were designed to predict 30-day survival after the first admission to the hospice ward. Classification accuracy (CA), sensitivity, specificity and area under the receiver operating characteristic curve (AUC) were calculated to evaluate the accuracy of each model.

Results

The numbers of 30-day survival were 65.4% (155 patients) in the training data and 58.1% (43) in the test data, respectively (P=0.255). LR identified 7 variables; lymphocyte, platelet, total protein, total bilirubin, AST, urea nitrogen and Chloride. NB identified 12 variables; lymphocyte, urea nitrogen, calcium, sodium, total bilirubin, total protein, eosinocyte, gender, AST, chloride, white blood cell and platelet. TI identified 4 variables; total bilirubin, lymphocyte, chloride and eosinocyte. CA, sensitivity, specificity and AUC of each model in the test data are listed in Table 1.

Table 1 Test Characteristics of Each Prognostic Model in the test data

	CA	Sensitivity	Specificity	AUC
LR	73.0	58.1	83.7	74.8
NB	77.0	67.7	83.7	80.6
TI	60.8	45.2	72.1	65.1

Conclusion

Three different data mining techniques independently identified total bilirubin, fraction of lymphocyte and electrolytes as important predictive factors for terminally ill cancer patients. Data mining techniques are thus useful tools in uncovering important relationships among clinical factors and patient outcomes within an existing database.

References

1. Parkes CM. Accuracy of predictions of survival in later stages of cancer. BMJ. 1972; 2: 29-31
2. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. BMJ. 2000; 320: 469-72.