Survival prediction with Bayesian Networks in more than 6000 non-small cell lung cancer patients

A.L.A.J. Dekker[1], A. Hope[2], P. Lambin[1], P. Lindsay[2]

[1]Department of Radiation Oncology (MAASTRO), GROW, University Medical Center Maastricht, the Netherlands
[2]Radiation Medicine Program, Princess Margaret Hospital, Toronto, Canada

Abstract

A model that predicts survival in lung cancer as a function of treatment choices would be valuable for decision support. In this study we built data flow tasks and a data warehouse to collect from clinical databases a large non-small cell lung cancer dataset from MAASTRO (N=1781) and from Princess Margaret Hospital (PMH, N=4591). We learned Bayesian Network (BN) models for survival prediction from the MAASTRO data and evaluated the models in the PMH dataset. The BN model based on stage and radiotherapy dose had a high predictive accuracy (AUC 0.917). The model correctly showed that radical radiotherapy (>60Gy) is beneficial for non-small cell lung cancer patients and that this benefit is disease stage dependent.

Keywords

Outcome prediction, Bayesian networks, Lung cancer, Data warehouse

Introduction

A model that predicts survival in lung cancer as a function of treatment choices would be valuable for decision support. Our current survival prediction models [1, 2] have a number of shortcomings. They often lack accuracy, contain parameters that are difficult to obtain in clinical practice, do not contain treatment choices, are not transparent to the user, cannot handle missing data well and are often based (and thus applicable) on highly selected patients. It is our hypothesis that a Bayesian Network (BN) model learned from unselected data does not have these shortcomings.

A BN is a directed acyclic graph, consisting of nodes and links. A node is a variable that can be observed (e.g. Stage) and/or inferred if unknown (e.g. Two year survival). The links between parent and child nodes are described in Conditional Probability Tables (CPTs) which hold the probability of a child having a certain state (e.g. Two year survival=True) given its parent state (e.g. Stage="IV"). The structure (nodes and links) and parameters (CPTs) of a Bayesian Network can be supplied by a domain expert but can also be learned from data if the data set is large enough.

In this study we aimed to collect a large, unselected non-small cell lung cancer dataset from multiple institutions and aimed to learn and evaluate a BN model from this dataset that can predict survival.

Material and methods

Data flow, source and destination databases

An automated data flow project using MS Visual Studio 2008 was developed to retrieve data from various clinical and research databases in MAASTRO Clinic in Maastricht (MAASTRO), The Netherlands and Princess Margaret Hospital in Toronto, Canada (PMH). During the data flow a common data model was applied and the data was stored in a MS SQL Server 2008 database. The data flow tasks were designed to empty and then completely re-fill the destination database, if updated source databases were available.

For MAASTRO, the source databases were a) CAT data warehouse (XML), in which data from the electronic medical file, PACS and Lantis are stored, b) Research database (SPSS) and c) Dutch government registry for survival information (XML).

For PMH, the source databases were a) Mosaiq (SQL Server), a record and verify and electronic medical file (IMPAC), b) eCancer, an in-house built data-warehouse with medical data which can export to Excel and c) Ontario cancer registry (Excel).

The data flow project was use to extract data on the first treatment of the first lung tumor of patients with a diagnosis of lung cancer. All data extraction was done fully automated. Running the data flow project in December 2009 resulted in a "lung cancer" database containing 2403 patients from MAASTRO and 9972 patients from PMH.

Patient characteristics and subsets

For this study, a total of 6372 patients were selected from the lung cancer database with the following selection criteria 1) not have known small-cell lung

cancer, 2) not have known surgical treatment, 3) have known overall or T or N or M stage. Some characteristics of the MAASTRO and PMH datasets are given in Table 1. For training purposes, a MAASTRO subset was made with patients with known overall stage and known survival (N=963). For validation purposes, a PMH subset was made without patients with known metastatic disease (No M1 or Stage IV, N=1996).

	MAASTRO (N=1781)		PMH (N=4591)	
	#	%*	#	%*
Stage				
I	219	14%	296	8%
II	79	5%	125	3%
IIIa	245	16%	479	12%
IIIb	452	29%	588	15%
IV	548	36%	2416	62%
Unknown	*238*	*13%*	*687*	*8%*
Two year survival				
True	206	19%	425	13%
False	866	81%	2849	87%
Unknown	*709*	*40%*	*1317*	*29%*
Prescribed EQD2 to the thorax				
<60 Gy	1194	73%	4042	88%
>=60 Gy	440	27%	549	12%
Unknown	*147*	*8%*	*0*	*0%*
Chemotherapy given				
True	738	54%	2313	65%
False	621	46%	1219	35%
Unknown	*422*	*24%*	*1059*	*23%*

Table 1: Patient characteristics. EQD2: Equivalent dose in fractions of 2 Gy; *: Percentage is relative to the number of known cases, except percentage unknown is relative to the total number of cases

Bayesian Network learning and model evaluation

Bayesian Network learning and inference was done in Hugin Researcher 7.1 (Hugin Expert A/S, Aalborg, Denmark). Different model structures were supplied by a domain expert, while the CPTs were learned from data using the expectation-maximization algorithm. The MAASTRO set was solely used for training the BN model, while the PMH set was only used to validate the BN model. During training of the model, maximum likelihood (ML) and the Bayesian information criterion (BIC) were used to evaluate different models, the latter having a penalty for model complexity. The performance metric for classifying two-year survival was the Area-Under-the-Curve (AUC) of the receiver operation curve in the PMH set.

BN models

In the current study, two BN models were built and evaluated. The first is a very simple model "Stage" consisting of two nodes: Stage and Survival. This model was learned on the described MAASTRO subset with only complete staging and survival information. This model is shown in Figure 1.

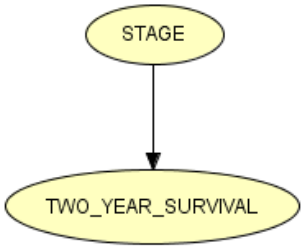

Figure 1: BN model "Stage" with stage as the prognostic factor for survival

A second more extensive model "Stage and RT" was learned that included stages cT, cN, cM and the prescribed radiotherapy dose to the thoracic region (Figure 2). The latter was learned using the full MAASTRO set.

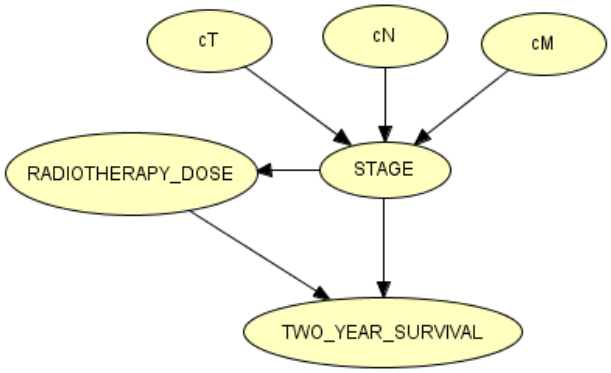

Figure 2: BN model "Stage and RT" with stage as a prognostic factor and radiotherapy as a predictive factor.

Results and discussion

Data

Table 1 shows that we were successful in extracting a large lung cancer dataset (>6000 patients) from electronic data sources in two hospitals. Probably the most striking difference between the two sets is that the PMH dataset contains more stage IV patients (62% vs. 36% at MAASTRO). The most likely cause is that PMH is an integrated cancer center in which multi-disciplinary clinics take place either in-house or in the hospital network, which means data is collected for all cancer patients. On the other hand, MAASTRO is a regional radiotherapy institute in which the radiation oncologist participates in multi-disciplinary clinics (and enters data) in the referring hospital. Only when a

patient is actually referred for radiotherapy will the patient data be entered into the MAASTRO environment. The difference in stage IV patients is the most likely cause for the difference in survival, prescribed dose and chemotherapy use.

BN model "Stage"

The CPT of the survival node is shown in Table 2. Due to its simple structure and the way it as learned (using a dataset without any missing data), the CPT is equal to the probabilities in the dataset itself. The AUC of this model in the full Toronto set is 0.915 while in the non-metastatic patients it is 0.816.

Stage	I	II	IIIa	IIIb	IV
$P_{\text{two year survival}}$	37%	21%	25%	25%	4%
Experience	145	46	56	172	280

Table 2: CPT of the BN model "Stage" for prediction of survival as a function of stage. $P_{\text{two year survival}}$: Probaility of two year survival. Experience: the number of observations the BN used in determining the CPT.

The BN model confirms that patients with more advanced stages have a lower probability of survival and the AUC shows that stage is quite a powerful predictor of outcome.

BN model "Stage and RT"

The CPT of survival node is shown in Table 3. The AUC of this model in the full Toronto set is 0.917 while in the non-metastatic patients it is 0.820.

	EQD2<60 Gy	EQD2>=60 Gy
Stage I		
$P_{\text{two year survival}}$	17%	42%
Experience	71	207
Stage II		
$P_{\text{two year survival}}$	13%	33%
Experience	50	65
Stage IIIa		
$P_{\text{two year survival}}$	17%	29%
Experience	106	184
Stage IIIb		
$P_{\text{two year survival}}$	20%	27%
Experience	226	279
Stage IV		
$P_{\text{two year survival}}$	4%	11%
Experience	679	7

Table 3: CPT of the BN model "Stage and RT" for prediction of survival as a function of stage and radiotherapy dose. EQD2: Equivalent dose in fractions of 2 Gy. $P_{\text{two year survival}}$: Probaility of two year survival. Experience: the number of observations the BN used in determining the CPT.

The BN model confirms that patients receiving radical (=>60Gy) have a better chance of survival and that the survival gain is stage dependent. The first result (radical radiotherapy increases survival) should be interpreted with caution as there may be a number of reasons why a stage I-IIIb patient receives a low dose. Next to non-survival related factors such as the position of the tumor and changes in practice (e.g. IMRT or IGRT introduction making higher dose possible), prescribed dose could simply be a surrogate for survival-related factors such as the general condition of the patient, the size of the tumor etc. We will investigate these relations in future work. The second result (radical radiotherapy works better in lower stage patients) should also be interpreted as it is likely that the lower stage patients simply received a higher dose.

Conclusion

We have combined clinical databases from two hospitals to populate a lung cancer research database containing more than 6000 non-operated, non-small cell lung cancer patients in which staging information is present. We have built Bayesian Network models that show that stage is a very strong predictor of outcome. The addition of radiotherapy information to the model does not deteriorate its performance and predicts that radiotherapy is beneficial to non-small cell lung cancer patients.

References

[1] C. Dehing-Oberije, D. De Ruysscher, H. van der Weide, M. Hochstenbag, G. Bootsma, W. Geraedts, C. Pitz, J. Simons, J. Teule, A. Rahmy, P. Thimister, H. Steck, and P. Lambin, "Tumor volume combined with number of positive lymph node stations is a more important prognostic factor than TNM stage for survival of non-small-cell lung cancer patients treated with (chemo)radiotherapy," *Int J Radiat Oncol Biol Phys,* vol. 70, pp. 1039-44, Mar 15 2008.

[2] C. Dehing-Oberije, S. Yu, D. De Ruysscher, S. Meersschout, K. Van Beek, Y. Lievens, J. Van Meerbeeck, W. De Neve, B. Rao, H. van der Weide, and P. Lambin, "Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy," *Int J Radiat Oncol Biol Phys,* vol. 74, pp. 355-62, Jun 1 2009.