e-ISSN 2083–4640
ISSN 1507–1367

RESEARCH PAPER

Reports of Practical Oncology and Radiotherapy
2021, Volume 26, Number 6, pages: 839–848
DOI: 10.5603/RPOR.a2021.0088
Received: 04.05.2021
Accepted: 02.07.2021

A data mining based clinical decision support system for survival in lung cancer

Beatriz Pontes[1], Francisco Núñez[2], Cristina Rubio[1], Alberto Moreno[2], Isabel Nepomuceno[1], Jesús Moreno[2], Jon Cacicedo[3], Juan Manuel Praena-Fernandez[4], German Antonio Escobar Rodriguez[2], Carlos Parra[2], Blas David Delgado León[5, 6], Eleonor Rivin del Campo[7], Felipe Couñago[8], Jose Riquelme[1], Jose Luis Lopez Guerra[5, 6]

[1]Department of Computer Language and Systems, Universidad de Sevilla, Seville, Spain
[2]Biomedical Informatics, Biomedical Engineering and Health Economy, Institute of Biomedicine of Seville (IBIS)/Virgen del Rocío University Hospital/CSIC/University of Seville, Seville, Spain
[3]Department of Radiation Oncology, Cruces University Hospital, Barakaldo, Spain
[4]Methodology Unit, University Hospital Virgen del Rocío, Seville, Spain
[5]Department of Radiation Oncology, University Hospital Virgen del Rocío, Seville, Spain
[6]Instituto de Biomedicina de Sevilla (IBIS/HUVR/CSIC/Universidad de Sevilla), Seville, Spain
[7]Department of Radiation Oncology, Tenon University Hospital, Hôpitaux Universitaires Est Parisien, Sorbonne University Medical Faculty, Paris, France
[8]Department of Radiation Oncology, Hospital Universitario Quirónsalud Madrid, Madrid, Spain

ABSTRACT

Background: A clinical decision support system (CDSS) has been designed to predict the outcome (overall survival) by extracting and analyzing information from routine clinical activity as a complement to clinical guidelines in lung cancer patients.

Materials and methods: Prospective multicenter data from 543 consecutive (2013–2017) lung cancer patients with 1167 variables were used for development of the CDSS. Data Mining analyses were based on the XGBoost and Generalized Linear Models algorithms. The predictions from guidelines and the CDSS proposed were compared.

Results: Overall, the highest (> 0.90) areas under the receiver-operating characteristics curve AUCs for predicting survival were obtained for small cell lung cancer patients. The AUCs for predicting survival using basic items included in the guidelines were mostly below 0.70 while those obtained using the CDSS were mostly above 0.70. The vast majority of comparisons between the guideline and CDSS AUCs were statistically significant ($p < 0.05$). For instance, using the guidelines, the AUC for predicting survival was 0.60 while the predictive power of the CDSS enhanced the AUC up to 0.84 ($p = 0.0009$). In terms of histology, there was only a statistically significant difference when comparing the AUCs of small cell lung cancer patients (0.96) and all lung cancer patients with longer (≥ 18 months) follow up (0.80; $p < 0.001$).

Conclusions: The CDSS successfully showed potential for enhancing prediction of survival. The CDSS could assist physicians in formulating evidence-based management advice in patients with lung cancer, guiding an individualized discussion according to prognosis.

Key words: data mining; lung cancer; clinical decision support system; survival; prognosis

Rep Pract Oncol Radiother 2021;26(6):839–848

Address for correspondence: Jose Luis Lopez Guerra, M.D., Ph.D. Department of Radiation Oncology, Virgen del Rocío University Hospital. Manuel Siurot avenue, s/n. 41013, Seville (Spain), tel: (+34) 95 501 2105, fax: (+34) 95 501 2111; e-mail: chanodetriana@yahoo.es

Introduction

Lung cancer is a cancer pathology with the highest mortality in men and the second leading cause of cancer death in women worldwide. It is estimated that 1.8 million new lung cancer cases and approximately 1.5 million lung cancer deaths are reported annually worldwide, which represents one out five of all cancer deaths [1]. Until 2035, the number of lung cancer deaths will increase globally by 86% compared to 2012, with an estimated increase of approximately 1.5 million in 20 years (20% in Europe) [2]. Despite technological and biological advances in recent years, survival in lung cancer is still limited, especially in locally advanced cases [3]. Beyond the limited set of factors related to survival that are already well known, such as the stage or histology, a key challenge is to find the largest number of factors that predict survival in each case individually, and combine them to provide the most accurate information for a specific patient to really customize their therapeutic management [4].

Numerous data sources (e.g. electronic medical records and outcomes data, imaging, laboratory, and pathology data, radiotherapy planning data, etc) provide opportunities for the application of data mining methodologies to improve technical capabilities and the overall quality and safety of cancer care delivery [5]. Variation in sensitivity to radiation depends on multiple factors and recent progress in data mining raises the possibility of customized analysis to characterize individual profiles that predict patient response to radiotherapy [6].

There are a huge amount of medical variables (clinical, physiological, genomic, molecular, therapeutic, etc.) that may affect the survival outcome in lung cancer patients [7, 8]. In order to provide real advances in routine cancer care, it is essential to carry out a strategy to reduce the dimensionality of this set of variables without diminishing its prognosis capacity. In addition, this approach should be able to take into account former outcome values so that they can be incorporated into the generation of new prediction models over time. This way, in a real-life setting, a system for predicting survival outcome aligned with precision medicine and learning healthcare system paradigms may be implemented [9].

Currently, the health system has been digitized to respond to user needs and improve and streamline workflows. Data Mining, also called Knowledge Discovery from Database, is a complex process which extracts and excavates unknown and valuable knowledge, such as a model or a regular pattern from mass incomplete, fuzzy, noisy, random data [10]. Data mining consists in automated data analysis which allows the observation of patterns representing knowledge [11]. Specifically, clinical data mining uses the data extracted from the health system [12], with the aim of interpreting the available clinical data and aid clinical decision making. Since 2000, there has been a growing interest in the application of data mining techniques to clinical data, with an increase by 10-fold of the number of papers having the term "data mining" in their title and referenced in MEDLINE [12].

Recent studies have shown the capacity of data mining-based models to predict the onset of lung cancer [13, 14]. However, the study of the survival outcome with this approach is sparse. Therefore, a clinical decision support system (CDSS) has been proposed to predict survival and apply the information obtained as a complement of the clinical guidelines in daily practice for lung cancer patients.

Materials and methods

Ethics statement

The project was authorized by the institutional Ethics Committee for clinical research and complies with the declaration of Helsinki and the Institutional Review Board of the participating centers. Written informed consent was obtained from all participants.

Lung cancer dataset

The lung cancer dataset includes clinical information gathered during routine care and agreed upon by a panel of experts in several medical disciplines (pulmonology, radiology, pathology, radiation oncology, surgery, and medical oncology). Prospective multicenter data from 543 consecutive lung cancer patients (Tab. 1) seen in consultation in 2 oncology services from 2013 to 2017 were available to feed the designed tool for creating the prediction models. Forty-two percent of patients were alive at the time of the study and 314 deaths occurred. The data set included 1167 variables, but only a proportion of them available at least in 90% of patients (including age, gender, histology, perfor-

Table 1. Patient characteristics

Characteristics	Number of patients (%) (n = 543)
Sex	
Female	74 (14)
Male	469 (86)
Age [years]	
Median (range)	66 (35-88)
KPS	
100	130 (24)
90	132 (24)
80	138 (25)
≤70	143 (26)
Tumor histology	
ADC	147 (27)
LCC	15 (3)
SCC	218 (40)
NSCLC, NOS	25 (5)
SCLC	138 (25)
Stage	
I–II	54 (10)
IIIA	193 (36)
IIIB	218 (40)
IV	78 (14)
Smoking status	
Current	289 (53)
Former	236 (43)
Never	18 (3)
Surgery	
No	449 (83)
Yes	94 (17)
Radiotherapy	
No	80 (15)
Yes	463 (85)
Radiation dose [Gy] (n = 463)	
< 60	219 (47)
≥ 60	244 (53)
Chemotherapy	
No	113 (19)
Yes	440 (81)
Concomitant chemotherapy (n = 440)	
No	236 (54)
Yes	204 (46)

KPS — Karnofsky performance status; SCC — squamous cell carcinoma; ADC — adenocarcinoma; LCC — large cell carcinoma; NSCLC — non-small cell lung cancer; NOS — not otherwise specified; SCLC — small cell lung cancer

mance status, stage, and treatment approach) and with discriminatory ability according to the data mining algorithms were used (Tab. 2). Different time periods [pre-treatment (variables available before the start of any oncologic treatment] and after the start of any oncologic treatment [only variables related to the treatment such as the therapy and toxicity)] were assessed for prediction.

Additionally, a subset of patients with longer follow-up (≥ 18 months) was also assessed (Tab. 2).

Data mining methods

To implement the technological architecture of the CDSS, we incorporated a series of open source tools that allow us to register information via electronic health records during clinical practice and use this information to conduct various data mining analyses.

In this particular work, we aim to create a model capable of predicting the survival of patients with lung cancer. For this task, we applied two classification methods: eXtreme Gradient Boosting (XGBoost) and Logistic Regression. XGBoost is a scalable tree boosting system which is widely used by data scientists [15], provides state-of-the-art results on many problems and implements the gradient boosting decision tree algorithm (also referred to as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines). Boosting is an ensemble process in which new models are introduced to correct the mistakes created by current models. Models are added sequentially until there can be no further changes. Gradient boosting is an approach [16] that generates new models that estimate previous models´ residuals or errors and then add them together to make the final prediction. It is called gradient boosting because when new models are added, it uses a gradient descent algorithm to mitigate the loss. The implementation used for XGBoost has been the one implemented in the caret package for R [17].

Logistic regression is a modeling procedure where a set of independent variables are used to model a categorical dependent variable using a logistic function, which is the cumulative logistic distribution [18]. The predicted variable is the response probability. Therefore, the model can provide a probability of response for every instance, given the estimated parameters for a set of predictor variables. A special case of the Generalized Linear Model can be seen as logistic regression [19] and thus analogous to Linear Regression [20]. However, the Logistic Regression model is based on very dif-

Table 2. Area under the receiver-operating characteristics curve (AUC; mean and 95% confidence interval) for predicting survival using either data mining analyses or basic items included in the guidelines in lung cancer patients

Data	Predictive model for mortality							
	Using data mining				Using guidelines			
	Lung cancer patients							
	All patients (n = 543)		Patients with follow-up > 18 months (n = 451)		All patients (n = 543)		Patients with follow-up > 18 months (n = 451)	
	N*	AUC	N*	AUC	N*	AUC	N*	AUC
Using pre-treatment data	13	0.84 (0.77–0.90)	6	0.74 (0.69–0.79)	2	0.60 (0.56–0.64)	2	0.64 (0.58–0.71)
Using only treatment data	7	0.78 (0.72–0.84)	6	0.81 (0.78–0.84)	3	0.60 (0.56–0.65)	3	0.65 (0.58–0.72)
Using all data	24	0.88 (0.83–0.92)	22	0.80 (0.77–0.83)	5	0.63 (0.58–0.68)	5	0.67 (0.60–0.75)
	Non-small cell lung cancer							
	(n = 405)		(n = 343)		(n = 405)		(n = 343)	
Using pre-treatment data	15	0.79 (0.72–0.85)	11	0.70 (0.64–0.76)	2	0.57 (0.51–0.62)	2	0.58 (0.50–0.67)
Using only treatment data	7	0.77 (0.72–0.82)	8	0.78 (0.73–0.84)	3	0.63 (0.56–0.70)	3	0.66 (0.57–0.75)
Using all data	23	0.81 (0.80–0.83)	20	0.77 (0.71–0.85)	5	0.64 (0.60–0.71)	5	0.66 (0.59–0.74)
	Small cell lung cancer							
	(n = 138)		(n = 108)		(n = 138)		(n = 108)	
Using pre-treatment data	31	0.82 (0.74–0.91)	12	0.73 (0.59–0.87)	2	0.67 (0.52–0.81)	2	0.74 (0.61–0.87)
Using only treatment data	4	0.76 (0.67–0.84)	6	0.90 (0.83–0.97)	3	0.42 (0.34–0.50)	3	0.47 (0.38–0.56)
Using all data	24	0.92 (0.86–0.98)	22	0.96 (0.92–0.99)	5	0.61 (0.54–0.68)	5	0.67 (0.58–0.77)

*Number of variables selected for the analysis; AUC — area under the receiver-operating characteristics curve

ferent assumptions from those of Linear Regression (about the relationship between dependent and independent variables). The implementation used was the Logistic Regression package provided for Python [21].

Statistical methods

All data analyses were carried out using the SPSS statistical software (version 19.0). The primary endpoint was the overall survival. The time of survival was estimated from the date of diagnosis to the date of death or last contact. Area under the receiver-operating characteristics curve (AUC) measured performance. The findings were compared with the AUC obtained using the fundamental elements contained in the guidelines [pretreatment data (stage, histology) and treatment data [surgery, radiation therapy and systemic therapy)] [22, 23]. The one-way analysis of variance (ANOVA) was used to compare the various AUCs obtained from 10 simulations for each rule, obtaining their confidence intervals at 95%. Multiple comparisons were analyzed using the Bonferroni correction and, when the hypothesis of homoscedasticity was not verified, were performed by the Games-Howell correction. Results were considered statistically significant if the p value was 0.05 or less.

Results

The AUCs for each subset are summarized in Table 2. Overall, the highest AUCs (> 0.90) for predicting survival were obtained for SCLC patients. For all SCLC cases, the AUC using data mining was 0.92 and 0.96 for patients with a minimum follow-up of 18 months. In contrast, the lowest AUCs (< 0.50) were observed for SCLC patients when us-

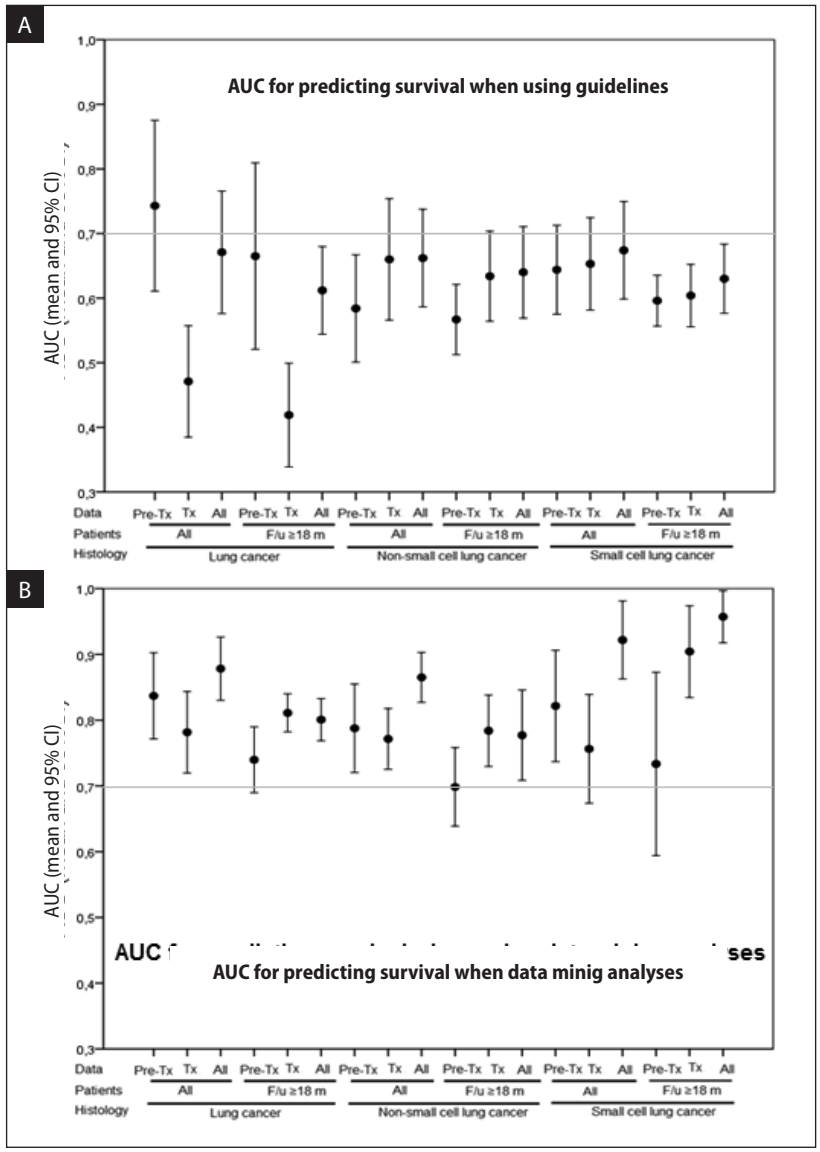

Figure 1. Area under the receiver-operating characteristics curve for predicting survival using either (**A**) basic items included in the guidelines or (**B**) data mining analyses in lung cancer patients. Tx — treatment; AUC — area under the receiver-operating characteristics curve; CI — confidence interval; F/u — follow-up

ing only the treatment variables recommended in the guidelines. The AUC was 0.47 and 0.42 for all patients and for those with a minimum follow-up of 18 months, respectively.

The AUCs for predicting survival using data mining analyses were mostly above 0.70 while those obtained using basic items included in the guidelines were mostly below 0.70 (Fig. 1). The vast majority of the comparisons between the AUCs obtained by data mining versus using the basic variables recommended in the guidelines were statistically significant regardless of the time period of the data used (pre-treatment data, treatment data, all data) or the follow up (Fig. 2–4). For instance, using the guidelines, the AUC for predicting survival in all lung cancer patients in the pretreatment setting was 0.60 while the predictive power of the CDSS enhanced the AUC up to 0.84 ($p = 0.0009$; Fig. 2).

In contrast, there were no significant differences in the AUCs when comparing all patients with only those with longer follow up (≥ 18 months). Additionally, there were no significant differences when comparing AUCs obtained with only pre-treatment data, only treatment data, or using all data. In terms of histology, there was only a statistically significant

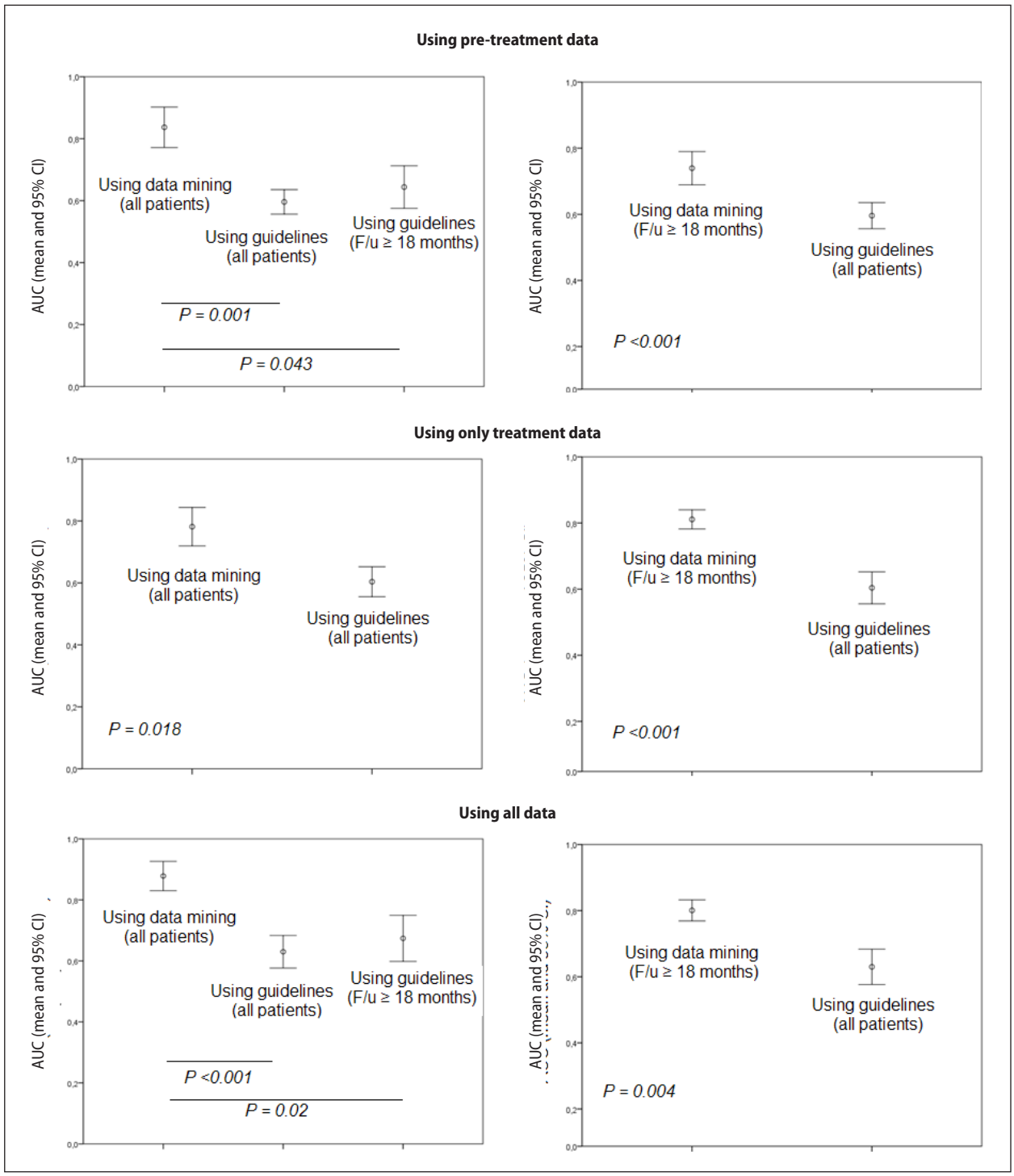

Figure 2. Comparison of the area under the receiver-operating characteristics curve [AUCs; mean and 95% confidence interval (CI)] for predicting survival in all lung cancer patients when using data mining analyses *vs.* the guidelines. The first column starts the comparison with all patients when using data mining while the second column starts the comparison with patients with longer follow up. F/u — follow-up

difference when comparing the AUC of all lung cancer patients with a minimum follow-up of 18 months (0.80) and the AUC of SCLC patients (0.96) using data mining ($p < 0.001$).

Discussion

Recent publications [13, 14] have shown that models based on data mining improve the diagnos-

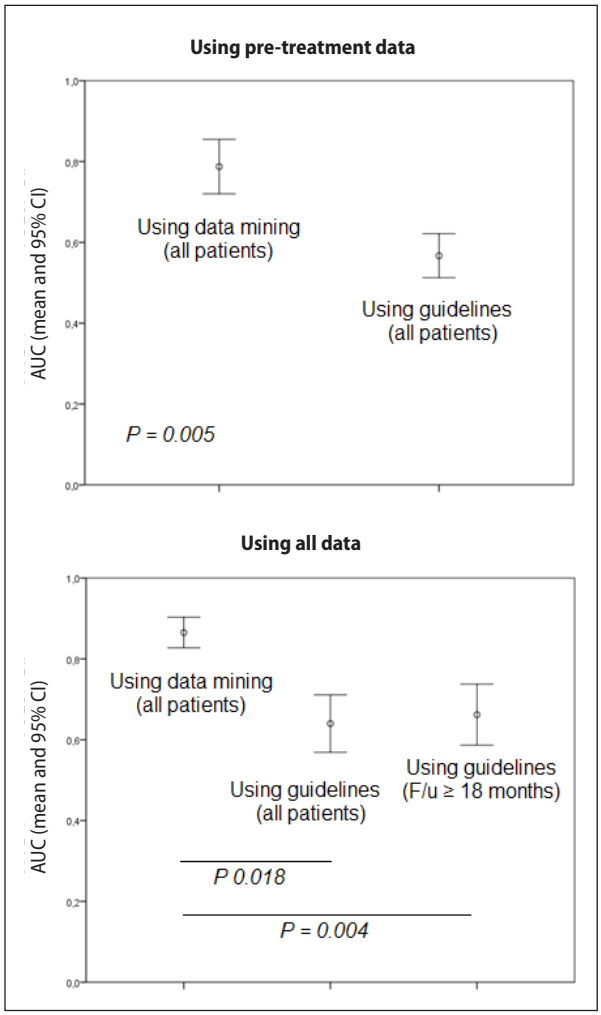

Figure 3. Comparison of the area under the receiver-operating characteristics curve [AUCs; mean and 95% confidence interval (CI)] for predicting survival in non-small cell lung cancer patients when using data mining analyses vs. the guidelines. F/u — follow-up

tic effect and profound significance for diagnostic of early stage lung cancer. A Chinese group [14] showed that the AUC level of each discriminative model was improved by about 10% based on multiple tumor markers and data mining compared with the diagnostic model based on different tumour markers, which indicates that the sensitivity and specificity of diagnosis can be substantially improved through combining different tumor markers compared to an individual tumor marker. However, the information about data mining based decision support systems for survival in lung cancer is sparse. At present, the influence of artificial intelligence on radiation oncology has been relatively minimal and may rightly seem more distant to many, given the specialty's largely interpersonal and complex interventional existence [24]. Our pertinent findings can be summarized as follows: first, we found that the highest AUCs for predicting survival were obtained when using data mining. Specifically, the highest AUCs were obtained in SCLC patients. Second, the vast majority of the comparisons between the AUCs obtained by data mining versus using the basic variables recommended in the guidelines were statistically significant regardless of the time period of the data used or the follow up. Third, there were no significant differences in the AUCs when comparing all patients with only those with longer follow up. Additionally, there were no significant differences when comparing AUCs obtained with only pre-treatment data, only treatment data, or using all data. Finally, there were no statistically significant differences in the majority of the AUCs comparisons according to histology.

Commonly used clinical models for survival prediction after radiation therapy for lung cancer can be limited by the lack of individual risk scores and disproportionate prognostic groups [25]. In this setting, different approaches have been developed to overcome that limitation. For instance, with nomograms it is possible to assess individualized probabilities for endpoints, and relevant prognostic factors can be evaluated. A multicenter cohort study led by the MAASTRO clinic [25] of lung cancer patients treated with stereotactic radiosurgery for brain metastases showed that two nomograms predicted early death (AUC ≥ 0.70) and long-term survival (AUC ≥ 0.67) more accurately than commonly used prognostic scores (range AUC = 0.51-0.68). Similar results were observed in our series when using data mining analyses. The AUCs for predicting survival were mostly above 0.70 while those obtained using basic items included in the guidelines were mostly below 0.70.

In lung cancer, the applicability of information discovery in database methods, based on data mining techniques, has been tested previously [26, 27]. Rivo et al. [27] reported a data mining project developed on a data warehouse containing records for 501 patients operated for lung cancer with curative intention. Data mining objectives were stated so as to discover risk factors for surgical mortality. The model which was finally selected had an AUC of 0.82 (0.74–0.89) ($p < 0.05$). There are substantial differences between the profile of patients and the endpoint in that study and ours. All patients were

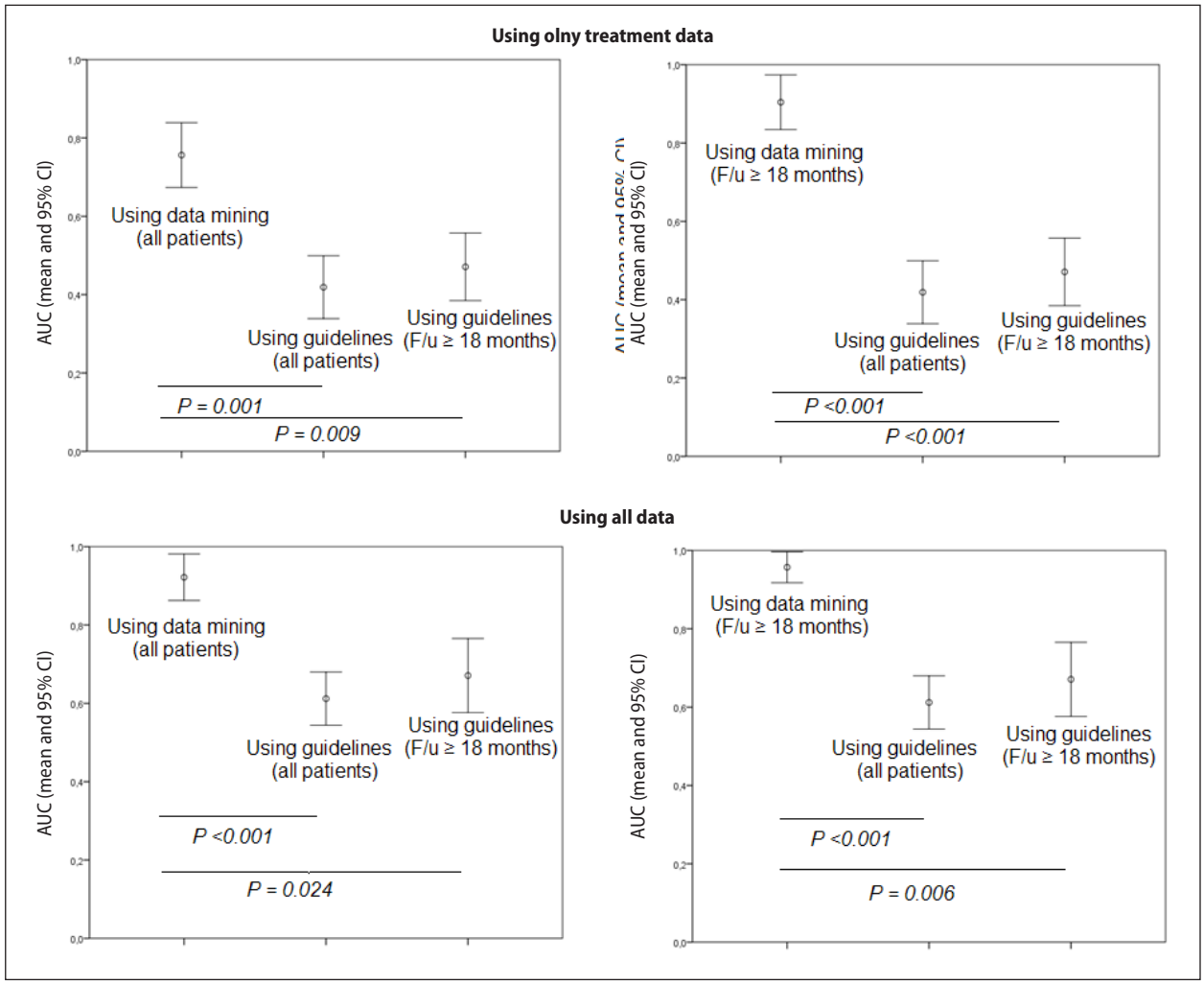

Figure 4. Comparison of the area under the receiver-operating characteristics curve [AUCs; mean and 95% confidence interval (CI)] for predicting survival in small cell lung cancer patients when using data mining analyses *vs.* the guidelines. The first column starts the comparison with all patients when using data mining while the second column starts the comparison with patients with longer follow up. F/u — follow-up

operated while this proportion is only 18% in our series. The endpoint, mortality, occurred in 30 (7%) patients [post-operative mortality (within 30 days)] in the surgical report while there were 314 (68%) deaths in our cohort at the time of the study. Although both approaches seem adequate, they probably can be only used in the setting of the patient profiles that were included to develop these CDSSs.

Our study had several limitations, amongst which are those inherent to the data mining process. Data mining enables the user to discover data trends and relationships, cannot guarantee perfect outcomes, cannot clarify why an outcome happens, and cannot fix data issues. We included only the data coming from two Institutions, collected prospectively, to minimize this limitation. Secondly, as mentioned above, due to the specific lung cancer patient profile data set (mostly unresectable stage III disease treated with radiation therapy) used for training of the algorithms and development of the CDSS, it may not be adequate for other data sets with differences in type of treatment or disease stage. Finally, the AUC is the most common metric used to measure the capacity of predictive and prognostic models to differentiate between individuals who develop the endpoint and those who do not. However, the AUC is frequently criticized [28, 29], and its interpretation has been a challenge since its introduction in medicine. Most of this AUC criticism can be traced back to the ROC curve, indicating that the appreciation of the AUC could be altered by a more intuitive interpretation

of the ROC. Therefore, we consider that the CDSS proposed should be evaluated with additional metrics and validated in an external cohort. This is part of an ongoing study of our research group.

Data mining can assist decision making, which is a mainstay in radiation oncology, but it can also improve knowledge and quality management. The data mining based decision support system for survival in lung cancer presented could be easily incorporated into the process system of the thoracic unit. Once implemented, this tool could facilitate a factual approach to decision making. Outputs from a data mining program can easily be re-directed to a benchmarking program aimed at continual improvement. Data mining is always hypothetical because its conclusions automatically become the hypothesis for the next data mining cycle. One model is never developed to be perfect and permanent but dynamic and, most importantly, useful.

Conflict of interests

The authors report no conflict of interest.

Funding

This research was funded by the S32 project (Carlos III Institute of Health of Spain, PI16/02104), OncoAID (Regional Ministry of Health of Andalusia, PIN-0476-2017) and supported by the ITC-Bio research suite (Ministry of Economy and Competitiveness of Spain, FPAP13-1E-2429).

References

1. Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. Adv Exp Med Biol. 2016; 893: 1–19, doi: 10.1007/978-3-319-24223-1_1, indexed in Pubmed: 26667336.
2. Didkowska J, Wojciechowska U, Mańczuk M, et al. Lung cancer epidemiology: contemporary and future challenges worldwide. Ann Transl Med. 2016; 4(8): 150, doi: 10.21037/atm.2016.03.11, indexed in Pubmed: 27195268.
3. Bradley JD, Paulus R, Komaki R, et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. Lancet Oncol. 2015; 16(2): 187–199, doi: 10.1016/S1470-2045(14)71207-0, indexed in Pubmed: 25601342.
4. Rajan JR, Chelvan AC. A Data Mining Approach to Diagnose Cancer for Therapeutic Decision Making. Altern Ther Health Med. 2019; 25(S1): 2–7, indexed in Pubmed: 30626737.
5. de Jong EEC, van Elmpt W, Rizzo S, et al. Applicability of a prognostic CT-based radiomic signature model trained on stage I-III non-small cell lung cancer in stage IV non-small cell lung cancer. Lung Cancer. 2018; 124: 6–11, doi: 10.1016/j.lungcan.2018.07.023, indexed in Pubmed: 30268481.
6. Roelofs E, Persoon L, Nijsten S, et al. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol. 2013; 108(1): 174–179, doi: 10.1016/j.radonc.2012.09.019, indexed in Pubmed: 23394741.
7. He RQ, Cen WL, Cen JM, et al. Clinical Significance of miR-210 and its Prospective Signaling Pathways in Non-Small Cell Lung Cancer: Evidence from Gene Expression Omnibus and the Cancer Genome Atlas Data Mining with 2763 Samples and Validation via Real-Time Quantitative PCR. Cell Physiol Biochem. 2018; 46(3): 925–952, doi: 10.1159/000488823, indexed in Pubmed: 29669324.
8. Xu T, Wei Q, Lopez Guerra JL, et al. HSPB1 gene polymorphisms predict risk of mortality for US patients after radio(chemo)therapy for non-small cell lung cancer. Int J Radiat Oncol Biol Phys. 2012; 84(2): e229–e235, doi: 10.1016/j.ijrobp.2012.03.032, indexed in Pubmed: 22608953.
9. Hsu ER, Klemm JD, Kerlavage AR, et al. Cancer Moonshot Data and Technology Team: Enabling a National Learning Healthcare System for Cancer to Unleash the Power of Data. Clin Pharmacol Ther. 2017; 101(5): 613–615, doi: 10.1002/cpt.636, indexed in Pubmed: 28139831.
10. Jiang P, Liu XS. Big data mining yields novel insights on cancer. Nat Genet. 2015; 47(2): 103–104, doi: 10.1038/ng.3205, indexed in Pubmed: 25627899.
11. Huang Z, Juarez JM, Li X. Data Mining for Biomedicine and Healthcare. J Healthc Eng. 2017; 2017: 7107629, doi: 10.1155/2017/7107629, indexed in Pubmed: 29065638.
12. Iavindrasana J, Cohen G, Depeursinge A, et al. Clinical data mining: a review. Yearb Med Inform. 2009: 121–133, indexed in Pubmed: 19855885.
13. Shahhoseini R, Ghazvini A, Esmaeilpour M, et al. Presentation of a model-based data mining to predict lung cancer. J Res Health Sci. 2015; 15(3): 189–195, indexed in Pubmed: 26411666.
14. Wang Z, Feng F, Zhou X, et al. Development of diagnostic model of lung cancer based on multiple tumor markers and data mining. Oncotarget. 2017; 8(55): 94793–94804, doi: 10.18632/oncotarget.21935, indexed in Pubmed: 29212267.
15. Torlay L, Perrone-Bertolotti M, Thomas E, et al. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform. 2017; 4(3): 159–169, doi: 10.1007/s40708-017-0065-7, indexed in Pubmed: 28434153.
16. Luo Y, Ye W, Zhao X, et al. Classification of Data from Electronic Nose Using Gradient Tree Boosting Algorithm. Sensors (Basel). 2017; 17(10), doi: 10.3390/s17102376, indexed in Pubmed: 29057792.
17. Jalal H, Pechlivanoglou P, Krijkamp E, et al. An Overview of R in Health Decision Sciences. Med Decis Making. 2017; 37(7): 735–746, doi: 10.1177/0272989X16686559, indexed in Pubmed: 28061043.
18. Adeli E, Li X, Kwon D, et al. Logistic Regression Confined by Cardinality-Constrained Sample and Feature Selec-

tion. IEEE Trans Pattern Anal Mach Intell. 2020; 42(7): 1713–1728, doi: 10.1109/TPAMI.2019.2901688, indexed in Pubmed: 30835210.
19. Zhao H, Hodges JS, Carlin BP. Diagnostics for generalized linear hierarchical models in network meta-analysis. Res Synth Methods. 2017; 8(3): 333–342, doi: 10.1002/jrsm.1246, indexed in Pubmed: 28683516.
20. Jan SL, Shieh G. Sample size calculations for model validation in linear regression analysis. BMC Med Res Methodol. 2019; 19(1): 54, doi: 10.1186/s12874-019-0697-9, indexed in Pubmed: 30866825.
21. Hsieh MH, Sun LM, Lin CL, et al. Development of a prediction model for pancreatic cancer in patients with type 2 diabetes using logistic regression and artificial neural network models. Cancer Manag Res. 2018; 10: 6317–6324, doi: 10.2147/CMAR.S180791, indexed in Pubmed: 30568493.
22. Kalemkerian GP, Loo BW, Akerley W, et al. NCCN Guidelines Insights: Small Cell Lung Cancer, Version 2.2018. J Natl Compr Canc Netw. 2018; 16(10): 1171–1182, doi: 10.6004/jnccn.2018.0079, indexed in Pubmed: 30323087.
23. Ettinger DS, Aisner DL, Wood DE, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 5.2018. J Natl Compr Canc Netw. 2018; 16(7): 807–821, doi: 10.6004/jnccn.2018.0062, indexed in Pubmed: 30006423.
24. Thompson RF, Valdes G, Fuller CD, et al. Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? Radiother Oncol. 2018; 129(3): 421–426, doi: 10.1016/j.radonc.2018.05.030, indexed in Pubmed: 29907338.
25. Zindler JD, Jochems A, Lagerwaard FJ, et al. Individualized early death and long-term survival prediction after stereotactic radiosurgery for brain metastases of non-small cell lung cancer: Two externally validated nomograms. Radiother Oncol. 2017; 123(2): 189–194, doi: 10.1016/j.radonc.2017.02.006, indexed in Pubmed: 28237400.
26. Edgerton ME, Fisher DH, Tang L, et al. Data mining for gene networks relevant to poor prognosis in lung cancer via backward-chaining rule induction. Cancer Inform. 2007; 3: 93–114, indexed in Pubmed: 19455237.
27. Rivo E, de la Fuente J, Rivo Á, et al. Cross-industry standard process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. Clin Transl Oncol. 2012; 14(1): 73–79, doi: 10.1007/s12094-012-0764-8, indexed in Pubmed: 22262722.
28. Yaffe MJ. Emergence of "Big Data" and Its Potential and Current Limitations in Medical Imaging. Semin Nucl Med. 2019; 49(2): 94–104, doi: 10.1053/j.semnuclmed.2018.11.010, indexed in Pubmed: 30819400.
29. Gleason KT, Dennison Himmelfarb CR. Big Data: Contributions, Limitations, and Implications for Cardiovascular Nurses. J Cardiovasc Nurs. 2017; 32(1): 4–6, doi: 10.1097/JCN.0000000000000384, indexed in Pubmed: 27918361.