



# A comparison between clinical decisions made about lung cancer patients and those inherent in the corresponding Scottish Intercollegiate Guidelines Network (SIGN) guideline

Health Informatics Journal

16(4) 260–273

© The Author(s) 2010

Reprints and permission: sagepub.

co.uk/journalsPermissions.nav

DOI: 10.1177/1460458210380520

<http://jhi.sagepub.com>

**Derek Sleeman, Laura Moss and Elias Gyftodimos**

University of Aberdeen, UK

**Marianne Nicolson and Graham Devereux**

Aberdeen Royal Infirmary, UK

## Abstract

Treatment and survival for patients with lung cancer vary between and within countries. We have undertaken a multifaceted study of a clinical dataset of 635 patients, to see if clinician treatment decisions were being made consistently and in accordance with the appropriate Scottish Intercollegiate Guidelines Network (SIGN) document. Subsequently, we created a dataset of 117 patients who should have undergone surgery according to the SIGN guideline. As analyses of this dataset did not provide clear distinctions between the main treatment groups, a clinician reviewed the case notes and dataset, checking for inconsistencies. The revised dataset was processed by a decision tree algorithm which suggests clinically plausible decisions. Further, statistical analyses compared the 54 patients offered surgery with the 52 who were not. These analyses suggest that there are significant differences: the most discriminating feature is *significant co-morbidity* ( $p < 0.001$ ). The article concludes with suggestions for how future guidelines might be enhanced.

## Keywords

co-morbidity, data curating, decision trees, guidelines, lung cancer

## Introduction

Lung cancer is the leading cause of cancer death in men, and in women is surpassed only by breast cancer. The survival rate for people diagnosed with lung cancer is poor, e.g. in Scotland 50 per cent

---

### Corresponding author:

Derek Sleeman, PhD, FBCS FRSE, Research Professor, Department of Computing Science, University of Aberdeen,

Aberdeen AB24 3FX, UK

Email: [d.sleeman@abdn.ac.uk](mailto:d.sleeman@abdn.ac.uk)

of lung cancer patients die within four months of diagnosis.<sup>1</sup> In recent years, whilst there have been improvements in the one-year lung cancer survival rate in the UK, five-year survival rate remains poor at 7 per cent.<sup>2</sup> When compared internationally, patients diagnosed with lung cancer in the UK are less likely to undergo surgical resection of the tumour and survive five years after diagnosis. Surgical resection rates in Britain (11%) compare unfavourably with those in Europe (17%) and North America (21%).<sup>3</sup> There is a threefold variation in surgical resection rates between English health authorities, and within the UK there is a fourfold variation in five-year survival rates.<sup>4,5</sup>

Several possible factors contribute to regional and international variation in rates of surgical resection and survival. First, there may be differences in patient profiles, particularly coexisting co-morbidities and willingness to seek medical advice early in their disease. Second, there may be differences in the investigation, assessment and management of patients presenting with potential lung cancer. The development of clinical databases established primarily for audit and clinical purposes provides an opportunity to analyse the decision-making processes of clinicians investigating and assessing patients with lung cancer. We have analysed such a clinical dataset with a number of machine learning algorithms<sup>6</sup> and statistical approaches to check for consistency of decision making and for whether the clinical decisions made are consistent with those inherent in the Scottish Intercollegiate Guidelines Network (SIGN) guideline on the management of lung cancer.<sup>7</sup> Our initial aim was to predict the treatment proposed by the multidisciplinary clinical team (MDT) for each patient, but given the noisy nature of the data this was downgraded to predicting the simpler decision of whether patients were offered potentially curative surgical resection. Ultimately we are interested in applying machine learning techniques to datasets from several referral centres to ascertain whether clinicians working in different hospitals differ in their decision-making processes, and to determine how this impacts on patient treatment and survival.

The article is organized as follows. The next section provides a literature review. This is followed by an outline of the various analyses applied to this lung cancer dataset. Further sections discuss the results and implications of the study, and planned further work.

## Literature review

Many national and international clinical groups produce guidelines in order to capture best clinical practice; these guidelines are then made available to clinicians. Guideline production usually involves setting up an expert review panel which attempts to capture and evaluate relevant evidence and where possible make evidence-based recommendations. The majority of the guidelines produced are in natural language with some flowcharts and other diagrammatic materials included.<sup>7</sup>

Guidelines are largely complex natural language documents, and it has been suggested that further computer support should be provided in creating, checking, and applying them clinically. Thus several systems and projects<sup>8</sup> have been developed to realize computer-assisted guideline management. Computer-interpretable languages and representation formalisms (e.g. Asbru,<sup>9</sup> GLIF,<sup>10</sup> PROforma<sup>11</sup>) have been developed and tools allow the computer-based execution of a guideline (e.g. Tallis<sup>12</sup>). Remaining challenges for the subfield include the treatment of patients with co-morbidities, and guidelines which are incomplete or inconsistent.

Guidelines often capture the treatment of one illness; however, the treatment of patients with several medical conditions may involve multiple guidelines, requiring computer support to allow concurrent merging of the multiple guidelines. Abidi et al., for example, suggest an ontology-based approach to enable the merging of guidelines at both the knowledge and execution levels.<sup>13</sup> A further challenge is that guidelines can be incomplete or inconsistent. Guidelines are also often static; however, 'living guidelines' are flexible guidelines that are updated on a more continuous

basis. Subsequently, this has led to the development of automatic guideline verification techniques which can be divided into those which verify the logical structure of the guideline and others which verify the clinical accuracy of the guideline. One particular approach to verifying the logical structure of a guideline is the use of a model checker; Bottrighi et al. describe the coupling of GLARE, a system for acquiring, representing and executing guidelines, with SPIN, a widely used verification tool.<sup>14</sup> Another approach to verification of the logical structure of a guideline involves the application of decision-table techniques to verify the completeness and consistency of a guideline;<sup>15</sup> ‘a guideline is said to be complete when an action is defined for every possible value-combination of parameters used within the guideline. It is considered consistent, if each of its rules, consisting of a certain condition and an assigned action, is unique.’<sup>16</sup> Bottrighi et al. describe a set of tools available for determining conformance to clinical guidelines, identifying discrepancies between actual clinical actions and those specified in the guideline; further, explanations are provided which attempt to justify the discrepancy.<sup>17</sup> Groot et al. have proposed a methodology for critiquing (the clinical accuracy of) guidelines, where the actual actions are based on real-world patient data.<sup>18</sup> In this instance, they employed model checking to investigate whether a patient’s actual treatment is consistent with the guideline. Patient data have also been used to verify clinicians’ actions against clinical guidelines in the Asgaard project which provided a methodology for the ‘retrospective review and critiquing of guideline-based medical care given to patients’.<sup>19</sup>

The literature reports many applications of statistics and machine learning (ML) approaches to medical datasets. For a recent review of some ML approaches see Berka et al.<sup>20</sup> Sleeman et al. describe a fairly standard application of several ML techniques to dialysis datasets which are likely to have clinical benefits.<sup>21</sup> Additionally, McQuatt et al.<sup>22</sup> describe the analysis of ML-derived decision trees by head injury physicians which resulted in the detection of anomalous situations. Studies applying machine learning techniques to the clinical processes of lung cancer treatment are few and have taken place in countries with relatively high surgical intervention rates (compared with the UK) and have tended to focus on the management of inoperable non-small cell lung cancer.<sup>23–25</sup> This study also focuses on non-small-cell lung cancer.

## Analyses of lung cancer dataset

All patients with possible lung cancer within NHS Grampian are assessed at Aberdeen Royal Infirmary by the Department of Respiratory Medicine in collaboration with the Departments of Radiology, Pathology, Thoracic Surgery and Oncology. In 2004 an electronic patient record was developed by AxSys Technology Ltd (Glasgow, UK) to facilitate patient management and audit. Data on all patients referred to the lung cancer service are routinely entered into the database by several lung cancer specialist nurses. Details recorded include demographic information, symptoms, smoking history, performance status,<sup>26</sup> co-morbidities, investigations performed, body mass index, ventilatory function, tumour stage,<sup>27</sup> decisions made at MDT meetings and treatment plan(s). The primary treatment options reported by this group and recorded in the database were: surgery with curative intent, radical radiotherapy with curative intent, palliative chemotherapy, palliative radiotherapy, palliative chemo-radiotherapy or best supportive care. For this study, data from 635 patients diagnosed with lung cancer in 2005–6 were extracted from the database. The principal study objective was to determine whether, at the principal clinical review of these patients (i.e. the MDT meeting), treatment decisions were being made consistently. That is, our initial modelling goal was to predict for each patient the treatment decision made by the MDT; the possible clinical decisions are given above. Our analysis has involved a number of stages; these are summarized below.

## *Initial data analysis and cleaning*

In the first round of pre-processing, we performed some data cleaning, removed clearly irrelevant attributes, derived some new attributes, and reformatted the full data as a single database table. The resulting dataset now contains 618 cases, each described by 90 attributes.

## *Application of machine learning algorithms*

In an attempt to model the treatment decision made by the multidisciplinary group for each patient we applied a number of machine learning algorithms to the dataset, including: Bayesian networks, naïve Bayesian classifiers, decision rule learning, neural networks and instance-based learning.<sup>6</sup> Given the relatively small amount of data in some of the classes, the 10-fold cross-validation technique was used.<sup>6</sup> The aims of these analyses were twofold: first, to build models which generalize the training data and can be used to make predictions (treatment decision) on new (unseen) cases; second, to detect interesting descriptive patterns in the derived models (rules or correlation between attributes) which can provide clinical insights. However, we observed that these approaches did not produce meaningful results, as the extracted generalized rules were not significantly better than guessing. Furthermore, the expected associations between clinical stage, tumour histology (small-cell/non-small cell) and treatments were not present in the models constructed. Further, the absence of expected associations by subsequent statistical analysis led us to suspect that there were significant inaccuracies in the dataset. Nevertheless, we decided to proceed with the next planned stage which was to compare the decisions made by the clinicians with those inherent in the corresponding SIGN guideline.<sup>7</sup>

## *Manual analysis of dataset against the corresponding SIGN guideline*

First we reviewed in detail the SIGN guideline for lung cancer<sup>7</sup> and then extracted rules from this document which covered decisions about the types of treatment to be offered lung cancer patients. The rules took the form:

IF condition C is true THEN apply action A

Figure 1 shows examples of rules which were extracted from the SIGN guideline. In total 15 rules were extracted which covered all the treatment options outlined above.

The rules extracted from the SIGN guideline were checked with clinical colleagues before they were applied to the dataset. For each case, the actual decision taken by the clinical team (typically: seven respiratory physicians, two radiologists, two oncologists, one or two thoracic surgeons, two pathologists, one cytologist, and two lung cancer nurse specialists) was extracted from the dataset; and the result of running the case (manually) against the extracted rule set was also recorded. It was not possible to evaluate all the rules extracted from the SIGN guideline against the information held in the dataset, as some of the information needed to evaluate particular rules is simply not available in the dataset. For example, rule 4 in Figure 1 is only to be applied if the patient's cancer is deemed to be 'inoperable'; similarly, rule 5 involves an assessment of the geometry of the patient's tumour. As it happens the uncertainties all occur in the rules which suggest non-surgical treatments. Thus it has been necessary to *estimate* the degree of match. Two of the authors independently carried out the matching exercise manually, and have concluded that based on tumour stage alone the agreement between the extracted SIGN rules and the decisions made by the Aberdeen MDT is ~ 64 per cent (i.e. ~ 400/618 instances). Further it should be noted that cases of reasonable doubt have been recorded as matches, which means that the above figure is an over-estimate of agreement.

- |  |
|--|
| <ol style="list-style-type: none"> <li>1) IF small cell carcinoma AND limited disease<br/>THEN no SURGERY</li> <li>2) IF non-small cell AND stage I/II<br/>THEN consider SURGERY</li> <li>3) IF non-small cell AND stage IIIA<br/>THEN consider CHEMOTHERAPY<br/>IF positive response<br/>THEN consider SURGERY</li> <li>4) IF non-small cell AND stage I/II AND inoperable<br/>THEN RADICAL RADIOTHERAPY</li> <li>5) IF non-small cell AND stage IIIA/IIIB AND tumour can be encompassed in a radiotherapy field AND<br/>(patient) performance status 0-1 AND weight loss &lt; 10%<br/>THEN RADICAL RADIOTHERAPY</li> </ol> |
|--|

**Figure 1.** Examples of rules extracted from the SIGN guidelines

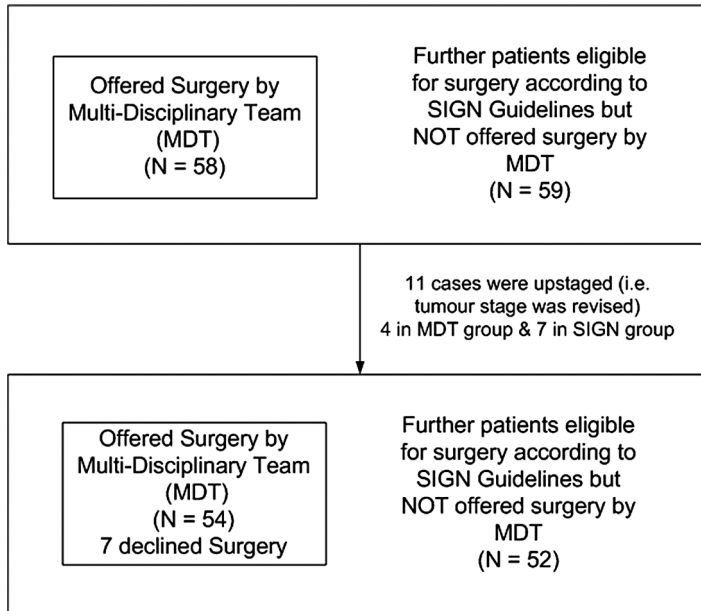
The high level of disagreement between the treatments recommended by the SIGN guideline and those actually made raised the possibility that the local MDT was seriously discordant from the SIGN guideline and/or there were inaccuracies in the database. To explore these issues further it was decided to focus the analysis on a subset of the database.

### *Analyses of subset of patients recommended surgery by the SIGN guideline*

As a result of the initial unsuccessful machine learning exercise and the discrepancies outlined above, we decided to focus the subsequent analyses on a subset of the cases of particular relevance to long-term lung cancer survival, namely potentially resectable cases of non-small-cell lung cancer.

There were three reasons for this decision. First, the most important and sensitive clinical decision is whether to offer surgery to a lung cancer patient. Second, it is possible to obtain unambiguous rules for the surgery/no-surgery decisions from the SIGN guideline. Third, given the poor prediction rates earlier of the machine learning algorithms on this task, it was reasonable to refocus the study on a less demanding goal (namely the binary decision to offer surgery: yes or no). Therefore, we identified from the database all cases (58) where surgery had been recommended by the MDT (9.1% of the total), and the 117 cases (18.4%) recommended for surgery by the SIGN clinical guideline, i.e. cases where the tumour histology and staging data (stages I, II and T3N1 IIIa) suggest surgery should be offered. The 58 cases recommended by the MDT were a subset of those recommended by the SIGN guideline (see Figure 2).

This data subset was then subjected to the machine learning and statistical analyses (outlined earlier), where the aim was now to predict for each patient the binary goal (operate or non-operate); once again the results of both the ML and the statistical analyses were not significant. Particularly worrying was the absence of association(s) between the decision to operate and tumour stage, patient performance status or ventilatory function. Consequently, it was decided that an experienced clinician (GD) would carry out an extensive review of each case in the subset (117 cases), using the case notes and the dataset. This review revealed a number of interesting issues related to attributes or features (the terms are used interchangeably in this article), as follows.



**Figure 2.** The number of cases recommended for surgery by the multidisciplinary clinical team and by the SIGN guideline

*Inconsistent use of attributes.* First, it was noted that the value for the tumour stage attribute was often updated as a result of further clinical investigations, and although the revised staging information would be recorded in the case notes (and rendered some patients inoperable), it would often not be recorded in the dataset. Consequently the clinicians would make decisions on accurate values but these were not available to the analysis programs.

When lung cancer patients are evaluated, a series of decisions is made and the outcomes of the several phases are often ‘intertwined’. For example, a patient is offered surgical resection but declines the operation for personal reasons; the decision recorded in the database is other than surgery, when in fact the *clinical* decision was to proceed to surgery. Similarly, a patient is offered and accepts an operation but at operation the tumour is unresectable; in some cases we have seen this also reported in the dataset as an option other than surgery. In this medical domain, we note there are a number of distinct decisions including:

- Was the patient offered surgery?
- Did the patient decide to proceed with surgery?
- What was the outcome of the surgery? (Options include: successful operation, operation aborted, patient dies during surgery, etc.)

*Implicit features used in decision making.* An earlier brainstorming session between the investigators, after the initial unsatisfactory analyses, had hypothesized that several ‘hidden’ or implicit attributes were being used by the multidisciplinary clinical team when making decisions about their patients. These included:

- post-surgery estimated ventilatory function (a parameter dependent on pre-operative lung function and the extent of the proposed surgical resection)
- likelihood of surviving operation (a composite factor influenced by a number of features contained in the current dataset).

Thus as part of the detailed review the clinician added values for the ‘post-surgery estimated ventilation function’ feature, but not for the ‘likelihood of surviving operation’ feature, which he felt could only be made following a patient consultation. Subsequent analyses of the enhanced datasets (see later) show whether these were important additions.

### *Resulting datasets*

As noted earlier (Figure 2), the SIGN guideline suggested that 117 patients should be offered surgery, whereas the MDT only recommended this treatment for a subset of 58 of those patients. Moreover, the detailed analysis of the patient records and database by the clinician noted that the tumour stage data for 11 of the patients had been updated by subsequent investigations, thus rendering surgical intervention inappropriate; this left a total of 106 patients according to the SIGN guideline who should have been offered surgery. Four of the excluded patients were in the subset which were initially offered surgery by the MDT – making that a subset of 54 patients. Of that subset, seven decided not to undergo surgery but for the purposes of analysis were included in the ‘offered surgery’ group. Of the 52 patients in the non-offered subset (106 – 54), 42 were not offered surgery by the MDT, and the remaining 10 were referred to the surgeons for final decisions. On these occasions each of the 10 referred patients was not offered surgery; so the final number of ‘no operation’ patients recorded in the dataset was 52 (as summarized in Figure 2). We were particularly interested to determine why those 52 patients, who should have been offered surgery according to tumour stage and the SIGN guideline, were not offered this treatment by the MDT.

Additionally, as a result of the above analyses, five additional attributes were added to each of the cases. The resulting datasets, both having 80 attributes, are: the full dataset of over 600 cases (with the new attributes marked ‘missing’) and the revised case subset of 106 records.

### *Analyses of the resulting datasets*

The analysis of the ‘full’ dataset using machine learning algorithms (essentially as described earlier) has not produced models which give highly accurate predictions; this is not surprising as values are missing for the majority of the added features. However, with the revised datasets for the 106 patients some interesting patterns emerged. This dataset has been subjected to various analyses; some of the more interesting ones are summarized below.

- The initial results with the revised dataset were disappointing in that the best overall true positive (TP) rate obtained with the C4.5 algorithm<sup>28</sup> on the test dataset was 51.9 per cent (using 10-fold cross-validation) (predicting surgery, sensitivity 0.56, specificity 0.53; predicting non-surgery, sensitivity 0.48, specificity 0.51).
- As a result of this experience the clinician selected 22 features (from the set of 80) which he thought were clinically relevant. In this case an overall TP rate of 49.1 per cent was achieved (using 10-fold cross-validation) (predicting surgery, sensitivity 0.52, specificity 0.50; predicting non-surgery, sensitivity 0.46, specificity 0.48).

- The clinician then suggested that only 17 of those features should be used; this resulted in an overall TP rate of 57.5 per cent (using 10-fold cross-validation) (predicting surgery, sensitivity 0.56, specificity 0.59; predicting non-surgery, sensitivity 0.60, specificity 0.56).

It was then decided to use the attribute selection mechanism of the WEKA package,<sup>28</sup> where initially all 80 features were provided to the process. The results obtained are as follows:

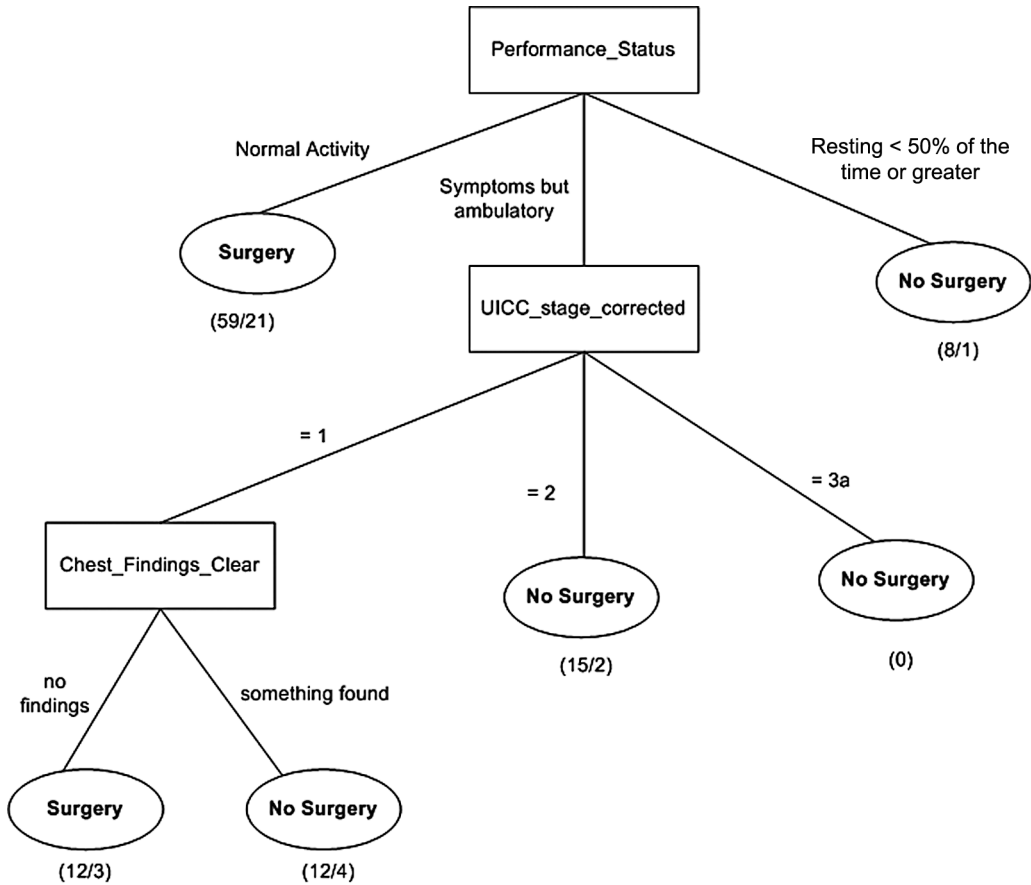
- Best overall TP rate of 67 per cent is achieved by the Bayesian network approach (using 10-fold cross-validation) (predicting surgery, sensitivity 0.65, specificity 0.69; predicting non-surgery, sensitivity 0.69, specificity 0.66).
- C4.5 creates a classifier which is particularly successful at predicting surgery cases, resulting in a TP rate of 0.74; overall TP rate achieved is 62 per cent (using 10-fold cross-validation) (predicting surgery, sensitivity 0.74, specificity 0.61). (See Figure 3 for the tree produced on the training set with this same algorithm and dataset.)
- Similarly, C4.5 creates a classifier which is particularly successful at predicting non-surgery cases, resulting in a TP rate of 0.75 (using 10-fold cross-validation); overall TP rate achieved is 63 per cent (using 10-fold cross-validation) (predicting non-surgery, sensitivity 0.75, specificity 0.60). (See Figure 4 for the tree produced on the training set with this same algorithm and dataset.)
- Using the attributes described in the SIGN guideline produced a classifier which resulted in an overall TP rate of 56 per cent on the test set (using 10-fold cross-validation) (predicting surgery, sensitivity 0.67, specificity 0.56; predicting non-surgery, sensitivity 0.46, specificity 0.57).

In relation to the above, note that variations in misclassification costs can effectively create different classifiers; this facility has been exploited here to create one classifier which is particularly effective at detecting surgery cases, and another for non-surgery cases.

We thought it would be informative to investigate how other guidelines compare with SIGN, so we did a comparable analysis using the National Institute for Health and Clinical Excellence (NICE) guideline (<http://www.nice.org.uk/guidelines>) which is used in non-Scottish UK centres. Using the attributes from NICE's comparable guideline produced a classifier which results in an overall TP rate of 63 per cent (predicting surgery, sensitivity 0.46, specificity 0.71; predicting non-surgery, sensitivity 0.81, specificity 0.59). Further, it achieved a TP rate of 0.80 for non-surgical cases (using 10-fold cross-validation) (see Figure 5 for the tree produced on the training set with this same algorithm and dataset).

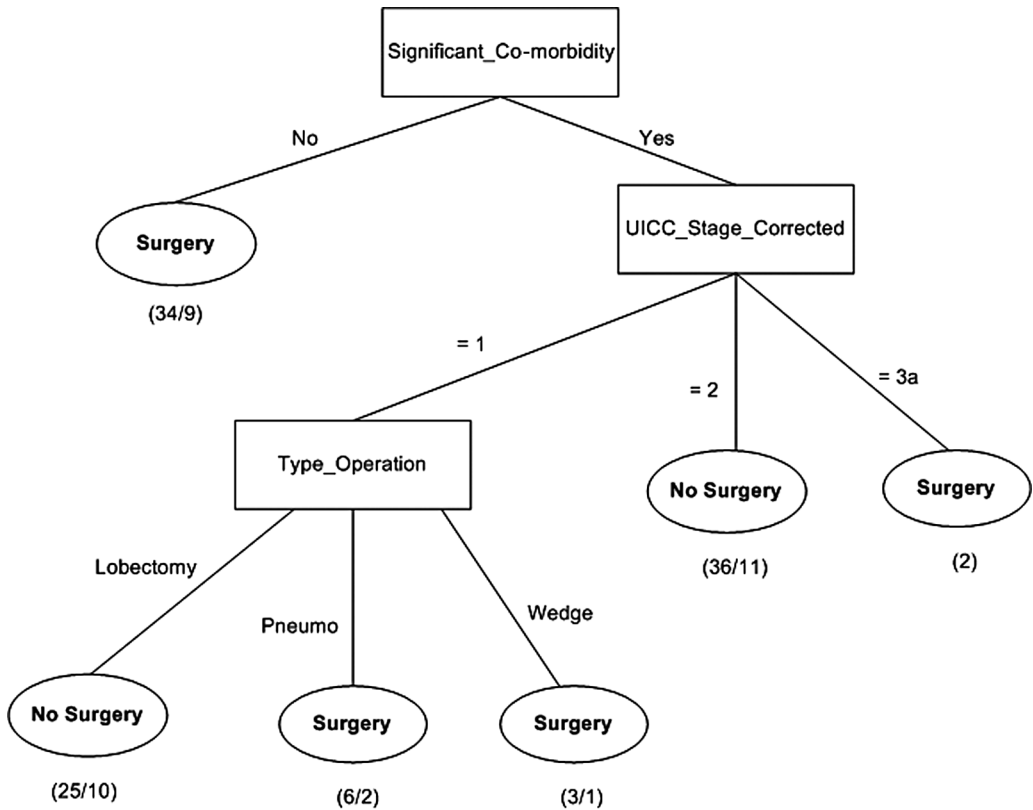
## Discussion

We note that the decision tree in Figure 3 (predicting surgery) contains just three features, as does Figure 4 (predicting no surgery). On the other hand, the tree based on the NICE guideline incorporates a single feature (Significant\_Co-morbidity). In fact just five features appear in all three of these decisions trees: Significant\_Co-morbidity (twice) and UICC\_stage\_corrected (twice); with Chest\_Findings\_Clear, Performance\_Status, and Type\_Operation all occurring just once. More specifically, the 'surgery' classifier has Performance\_Status as its 'top' node; similarly the 'no surgery' classifier has Significant\_Co-morbidity as its most significant feature. One of the clinicians has commented that each of the features in these classifiers 'are intuitively relevant to their decisions'. Further, we note that the newly introduced feature (revised *tumour stage*, i.e. UICC\_stage\_corrected) influences the decisions significantly. Table 1 explains the various features in these decision trees.



**Figure 3 .** A Classifier for “surgery” (which does not cover extreme values for the Performance\_Status descriptor, ie resting for 50% or more of the time).

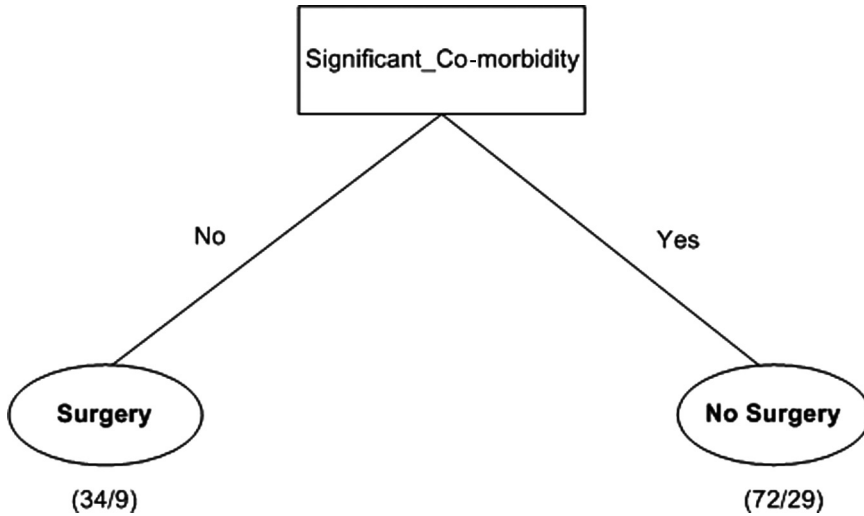
Although some (clinically) intuitive rules are being proposed, we are still unable to obtain very high accuracy predictions with the present datasets. This is not completely unexpected, as in the full dataset, as noted earlier, there is a significant amount of missing data. We believe there are a number of reasons why the accuracy is not higher with the subset of 106 records. First, the revised dataset contains a small number of cases compared to the complexity of the dataset. Second, for this cohort of 106 patients, the initial decision was: 54 should be offered surgery, 42 were not to be offered surgery, and the remaining 10 were referred to surgeons for the final decisions. Eventually, as noted earlier, all the referred patients on this occasion were *not* offered surgery. However, the reasons for the final decisions for these 10 patients are not accessible either to the MDT meeting or to the machine learning algorithms. The decision of the surgeon not to operate after reviewing the patient is likely to be influenced not only by objective parameters but also by subjective parameters (e.g. surgeon’s experience and training). As each of these 10 patients could have been assigned to either the surgery or the no surgery categories, the maximum accuracies which could be expected are for the surgery group 84.4 per cent (i.e.  $54/(54+10)$ ) and for the no surgery group 80.8 per cent (i.e.  $42/(42+10)$ ).



**Figure 4.** A classifier for “no-surgery”

We reported earlier that the overall TP rate achieved by the surgery classifier was 74 per cent and those by the no surgery classifiers were 75 and 80 per cent. The last point above indicates that these actual values are close approximations to the maximum values which can be expected in situations where a significant number of patients are effectively ‘unclassified’.

As shown in Figure 2, 54 patients were offered surgery by the MDT and 52 were not. As noted earlier, we wished to investigate whether there are significant differences between these two patient groups, so in addition to the machine learning analyses we subsequently ran a range of statistical tests on the two groups. The results of these tests, reported in Table 2, suggest that there are important differences between the patients who were offered surgery by the MDT and those who were not. (It is clear that the objectives of the statistical and the machine learning (ML) analyses are similar. Rightly or wrongly, what we have reported here is the actual order in which the investigations were undertaken (i.e. ML approaches initially, and statistical approaches as confirmation). In retrospect, it is also clear that the overall results would have been similar if we had used the approaches in the alternative order.) The most discriminating feature is significant co-morbidity ( $p < 0.001$ ); age, ventilatory function (mean FEV1), diagnosed chronic obstructive pulmonary disease (COPD), and ability to perform tasks of daily living (performance status) are further possible discriminants (but at lower confidence levels). In practice it might be important to establish whether there are just a small number of diseases contributing to the co-morbidity, and for these to be articulated.



**Figure 5.** A classifier based on attributes described in the NICE guidelines

## Conclusions and further work

For the dataset to be useful for audit and epidemiological purposes, features in the dataset have to be recorded consistently. In this study we noted a number of inconsistencies, some of which were totally unexpected, e.g. that the recorded decision of ‘no surgery’ covered at least three scenarios. A likely factor contributing to the inconsistencies is that when care of a patient is transferred within the MDT, the new clinician concentrates on his/her part of the database and fails to record results of investigations from other specialties. Consequently, whilst the original decision data were accurate at the time of entry, subsequent decisions based on further investigations and patient consultations by other members of the care pathway were often not recorded.

This study suggests that when clinicians make decisions they (often) use information which is not explicitly recorded in the patient notes, including their (implicit) clinical experience. For

**Table 1.** Description of features used in decision trees

UICC_stage_corrected	The clinical tumour stage, after case review
Significant_Co-morbidity	The presence, on initial clinical assessment, of any of the following conditions: COPD, cerebrovascular disease, heart disease, peripheral vascular disease, diabetes, renal disease, alcohol excess, other cancers, dementia
Performance_Status	An assessment of how the lung cancer is affecting the ability of the patient to complete daily tasks <sup>26</sup>
Type_Operation	Three potentially curative surgical interventions are conducted locally: pneumonectomy, lobectomy and wedge resection
Chest_Findings_Clear	The absence of any abnormal clinical signs on examination (expansion, percussion, breath sounds)

**Table 2.** Reviewed database: summary of patients offered surgery by MDT and those eligible for surgery according to SIGN guideline

	Offered surgery by MDT <i>n</i> = 54	SIGN indicates surgery, but not offered by MDT <i>n</i> = 52	<i>p</i>
Age	67.7	71.8	0.025
Mean (95% CI)	(64.8–70.5)	(69.6–74.0)	
FEV1 (litres)	1.98	1.71	0.023
Mean (95% CI)	(1.78–2.19)	(1.54–1.89)	
FEV1 (% predicted)	75.7%	70.4%	0.22
Mean (95% CI)	(69.5–81.9)	(63.6–77.2)	
Estimated post op			
FEV1 (% predicted)	50.6%	48.2%	0.11
Mean (95% CI)	(46.4–54.8)	(43.7–52.7)	
Estimated post op			
FEV1 (absolute volume, litres)	1.32	1.18	0.42
Mean (95% CI)	(1.20–1.44)	(1.06–1.30)	
% significant co-morbidity	58.7%	82.7%	0.001
% COPD	11.1%	28.8%	0.022
% cardiovascular disease	27.8%	38.5%	0.24
Performance status 0/1	98.1%	86.5%	0.024

example, they might decide whether a patient is likely to survive an operation based on a number of factors which they know about the patient (some of which may be recorded in the dataset). It would of course be very helpful for *trainee clinicians* if all the attributes used to make decisions are made explicit.

### Guidelines

Based on this study it is apparent that the current SIGN lung cancer guideline, whilst highly sensitive, lacks specificity and requires clinical judgement to make the final decision (about surgical resection) in about 50 per cent of patients. Currently the objective SIGN guideline is still reliant on the ‘art’ of the clinician with its inherent variability; future guidelines should attempt to increase the specificity.

As a result of this study, we have some general recommendations for the development of guidelines. The last two points consider computational approaches to ensure that the guidelines created are made more consistent.

- Guidelines should contain definitions, with appropriate examples, of the descriptors to be collected. (These will be of considerable value for audit as well as computer-based analyses.)
- A significant finding of our study is that this MDT, although aware of the SIGN guideline, did not offer treatment to all patients ‘covered’ by the guideline as they had a number of valid clinical reasons why they thought the treatment would not be appropriate. As we reported earlier, co-morbidities are a major factor influencing clinicians not to proceed with some treatments. Thus it would seem sensible to include, in future guidelines, co-morbidities and associated severity which suggest that a particular treatment should *not* be offered.

- Indeed, we believe there is a case for guideline committees to publish a package of material which includes the guidelines as well as the associated clinical protocol(s) which would include this additional patient management information; such protocols should be sufficiently detailed to support decision making in a range of clinical situations.
- At present a guideline's rules and procedures are described in natural language text; we are proposing that these inherent rules should be described additionally in an unambiguous rule format (for example, SWRL, <http://www.w3.org/Submission/SWRL>) which will make their interpretation less subjective.
- Further, we propose that guidelines be published with a set of typical cases in an appropriate machine readable form. The dataset could then be evaluated against the rule set (previous point) to determine whether any of the cases would be misclassified by the suggested rule set.

### Further work

Given the shortcomings noted above, we are planning to run a trial at three Scottish centres which addresses the points identified. Additionally, this study will use several of the data analysis tools already developed at Aberdeen, including the ACHE<sup>29</sup> and INSIGHT<sup>30</sup> systems.

### Acknowledgements

We gratefully acknowledge support from the EPSRC AKT IRC for Laura Moss (Studentship) and Dr Elias Gyftodimos. Further, Professor John Kinsella (Glasgow) and Professor Jeremy Wyatt (Dundee) have provided us with very helpful feedback on an earlier version of this article.

### References

1. Scottish Executive Health Department. *Cancer scenarios: an aid to planning cancer services in Scotland in the next decade*. Edinburgh: Scottish Executive, 2001.
2. Coleman MP, et al. Trends and socioeconomic inequalities in cancer survival in England and Wales up to 2001. *Br J Cancer* 2004; 90: 1367–1373.
3. Sant M, et al. EUROCARE-3 survival of cancer patients diagnosed 1990–94: results and commentary. *Ann Oncol* 2003; 14 suppl 5: 61–118.
4. Jack RH, et al. Geographic inequalities in lung cancer management and survival in South East England: evidence of variation in access to oncology services. *Br J Cancer* 2003; 88: 1025–1031.
5. NHS Executive. *National performance indicators for the NHS*. London, 2000.
6. Mitchell T. *Machine learning*. McGraw Hill, 1997.
7. Scottish Intercollegiate Guidelines Network (SIGN). Management of patients with lung cancer 2005. <http://www.sign.ac.uk/pdf/sign80.pdf> (accessed April 2009).
8. Ten Teije A, Miksch S and Lucas P (eds). *Computer-based medical guidelines and protocols: a primer and current trends*. Amsterdam: IOS Press, 2008.
9. Shahar Y, Miksch S and Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artific Intell Med* 1998; 14 (1–2): 29–51.
10. Peleg M, Boxwala AA, Ogunyemi O, et al. Glif3: the evolution of a guideline representation format. In: *Proceedings of the American Medical Informatics Association Annual Symposium 2000*, 2000, p.645–649.
11. Fox J, Johns N, Rahmanzadeh A, Thomson R. Disseminating medical knowledge: the PROforma approach. *Artific Intell Med* 1998; 14 (1–2): 157–182.

12. <http://cossac.org/groups/oxford/teaching/cogsys/tallis/> (accessed 25 March 2010).
13. Abidi SR and Abidi SSR. Towards the merging of multiple clinical protocols and guidelines via ontology-driven modeling. In: *12th International Conference on Artificial Intelligence in Medicine Proceedings*, 2009, Springer, p.81–85.
14. Bottrighi A, Giordano L, Molino G, Montani S, Terenziani P, Torchio M. Adopting model checking techniques for clinical guidelines verification. *Artific Intell Med* 2010; 48 (1): 1–19.
15. Shiffman R. Representation of clinical practice guidelines in conventional and augmented decision tables. *J Am Med Inform Assoc* 1997; 4: 382–393.
16. Duftschmid G and Miksch S. Knowledge-based verification of clinical guidelines by detection of anomalies. *Artific Intell Med* 2001; 22 (1): 23–41.
17. Bottrighi A, Chesani F, Mello P, et al. A hybrid approach to clinical guideline and to basic medical knowledge conformance. In: *12th International Conference on Artificial Intelligence in Medicine Proceedings*, 2009, Springer, p.91–95.
18. Groot P, Hommersom A, Lucas P, Serban R, ten Teije A, van Harmelen F. The role of model checking in critiquing based on clinical guidelines. In Bellazzi R, Abu-Hanna A, Hunter J (eds), *Proceedings of the AIME 2007 Conference*, Springer, 2007, p.411–420.
19. Advani A, Kinkoi L, Kinkoi LM, Shahar Y. Intention-based critiquing of guideline-oriented medical care: the Asgaard Project at Stanford. In: *Proceedings AMIA Annual Symposium*, 1998, p.483–487.
20. Berka P, Rauch J and Zighed D. *Data mining and medical knowledge management: cases and applications*. Medical Information Science Reference, 2009.
21. Sleeman D, Fluck N, Gyftodimos E, Moss L, Christie G. An intelligent aide for interpreting a patient's dialysis dataset. In Bellazzi R, Abu-Hanna A, Hunter J (eds), *Proceedings of the AIME 2007 Conference*, Springer, 2007, p.57–66.
22. McQuatt A, Andrews PJD, Sleeman D, Corruble V, Jones PA. The analyses of head injury data using decision tree techniques. In Horn W, et al. (eds). *Artificial intelligence in medicine: Proceedings of AIMDM'99 Conference*, Aalborg, June 1999, Springer, p.336–345.
23. Erkurt E, Tunalı C and Erkisi M. Primary therapeutic decision-making in inoperable non-small cell lung cancer. *Int J Rad Oncology* 2000; 46: 439–444.
24. Wigren T, and Kolari P. Evaluation of a decision-support system for inoperable non-small cell lung cancer. *Methods Inform Med* 1994; 33: 397–401.
25. Sanders GD, Nease RF Jr and Owens DK. Design and pilot evaluation of a system to develop computer-based site-specific practice guidelines from decision models. *Med Decision Making* 2000; 20: 145–159.
26. WHO/ECOG Performance status. [www.ecog.org/general/perf\\_stat.html](http://www.ecog.org/general/perf_stat.html) 2006 (accessed June 2009).
27. Mountain CF. A new international staging system for lung cancer. *Chest* 1986; 89 (suppl 4): 225S–233S.
28. Witten IH and Frank E. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
29. Moss L, Sleeman D, Kinsella J, Sim M. ACHE: an architecture for clinical hypothesis examination. In: *Proceedings of 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS 2008)*, Jyväskylä, Finland, p.158–160.
30. Sleeman D, Aiken A, Moss L, Kinsella J, Sim M. A system to detect inconsistencies between a domain expert's different perspectives on (classification) tasks. In: Kacprzyk J (ed.) *Advances in Machine Learning II: Studies in Computational Intelligence*. Springer, 2009, p.293–314.