On Predicting Recurrence in Early Stage Non-small Cell Lung Cancer

Sameh K. Mohamed[1,2], Brian Walsh[1,2], Mohan Timilsina[1,2], Maria Torrente[6], Fabio Franco[6], Mariano Provencio[6], Adrianna Janik[3], Luca Costabello[3], Pasquale Minervini[4], Pontus Stenetorp[4], Vít Nováček[1,5]

[1]Data Science Institute, NUI Galway, Galway, Ireland
[2]Insight Centre for Data Analytics, NUI Galway, Galway, Ireland
[3]Accenture Labs, Dublin, Ireland
[4]University College London, London, United Kingdom
[5]Faculty of Informatics, Masaryk University, Brno, Czech Republic
[6]Medical Oncology Department, Hospital Universitario Puerta de Hierro Majadahonda, Madrid, Spain

Abstract

Early detection and mitigation of disease recurrence in non-small cell lung cancer (NSCLC) patients is a nontrivial problem that is typically addressed either by rather generic follow-up screening guidelines, self-reporting, simple nomograms, or by models that predict relapse risk in individual patients using statistical analysis of retrospective data. We posit that machine learning models trained on patient data can provide an alternative approach that allows for more efficient development of many complementary models at once, superior accuracy, less dependency on the data collection protocols and increased support for explainability of the predictions. In this preliminary study, we describe an experimental suite of various machine learning models applied on a patient cohort of 2442 early stage NSCLC patients. We discuss the promising results achieved, as well as the lessons we learned while developing this baseline for further, more advanced studies in this area.

Introduction

Lung cancer is the most common cause of cancer related mortality in the world. Globally in 2018, more than 2 million new cases were diagnosed and more than 1.8 million deaths due to lung cancer occurred [(1)], where the main risk factor for the development of the disease is tobacco. In recent years, while mortality in men has decreased slightly due to smoking cessation, mortality in women has increased and has practically doubled due to its later incorporation into smoking habit[(2)]. Non-small cell lung cancer (NSCLC) accounts for approximately 80% of all lung malignancies. The diagnosis of lung cancer in early stages remains a challenge and often occurs incidentally in the study of other diseases [(3)]. Approximately 70% of patients are diagnosed in locally advanced stages (stage III) or metastatic disease (stage IV) [(4)], which contributes to low survival rates. Of note, survival declines progressively with increasing clinical stage. Overall survival at 5 years in NSCLC is around 10–15% [(5)]. Early stage NSCLC (stage I-II) patients are typically treated with complete surgical resection of the tumor. However, even after the entire resection of the tumor, 30–55% of patients will develop disease recurrence within the first 5 years of surgery [(6)]. This encouraged research into applying survival analysis methods and machine learning models to identify patients with high risk for recurrence to help in personalizing surveillance plans for these patients.

Predictive approaches for identifying tumor recurrence in NSCLC patients utilize different types of techniques such as statistical analysis of patient Electronic Health Records (EHRs), machine learning models and patient self-assessment feed-backs. The simplest form of these methods is the self-assessment techniques which rely on continuous patient self-evaluation and reporting after successful treatment[(7)]. This technique traditionally requires patients to perform self-assessment in terms of symptoms, physical abilities and other metrics, and report these metrics to hospitals through previously specified means *e.g.* mobile applications, and online forms[(8)]. It then notifies the physician when certain patterns appear in the patient recorded metrics[(8)]. Due to the simplicity of the approach, it cannot provide early predictions of the tumor recurrence, and it relies on patients compliance to reporting which is not guaranteed[(9)].

On the other hand, statistical survival analysis methods and supervised machine learning models are able to provide early predictions for the tumor recurrence based on the patient characteristics and previous treatments history. One of the simplest forms of the statistical analysis tools is nomograms[(10)], which are a graphical representations of equations that predict medical outcome. Nomograms use a points-based system whereby a patient accumulates points based on levels of his or her risk factors, where the cumulative points total is associated with a prediction. More complex techniques such as the Cox hazard model can provide a time-based analysis of the patient hazards *e.g.* tumor recurrence.

The Cox hazard model is one of the most widely used techniques in predicting recurrence and other prognosis factors in NSCLC patient [(11–13)]. Despite its popularity, however, the Cox hazard model operate strictly with multiple assumptions *e.g.* feature independence, linear relations, etc, which are not always guaranteed. It also cannot consume heterogeneous data such as medical imagery, genetic arrays, etc. These two limitations are addressed in supervised machine learning models such random forests, support vector machines, neural networks, etc, which operate using different techniques that makes them suitable for modeling complex relations on heterogeneous data[(14, 15)].

Supervised models were used in different studies to predict NSCLC recurrence from patients EHRs, where different studies have shown that they provide superior predictive accuracy compared to classical survival analysis methods[(16)] such as the Cox hazard model. However, these studies traditionally only examine a limited set of supervised machine learning approaches, and they also rely on differently structured patient records due to the different data collection protocols. These two factors make it difficult to perform a more representative comparison between the survival analysis methods and the different supervised machine learning models. Therefore, it is essential to establish rigorous means for comparison between the different supervised predictive models and statistical survival analysis techniques on unified patient EHRs structure to understand the differences between these techniques, and establish informative baselines for the task of predicting NSCLC recurrence.

In this work, we study the problem of predicting recurrence in NSCLC patient where we discuss our first-steps towards building an efficient and accurate recurrence prediction approach within the CLARIFY project — a European project focused on monitoring health status and quality of life after the cancer treatment (cf. `https://www.clarify2020.eu/`). Based on the previously discussed developments and challenges in relation to predicting recurrence in NSCLC patients, we started our development by assessing the performance of the basic supervised machine learning models in NSCLC recurrence prediction to establish a baseline benchmark.

In this study, we provide the outcomes of our assessment of the basic supervised machine learning methods on patient cohort of 2442 early stage NSCLC patients, where we evaluate these models on a binary classification task with the objective of classifying patients with successful treatments into one of two categories: tumor recurrence or disease-free survival. We also discuss the challenges associated to the application of these methods in terms of model accuracy, data quality, and utilized evaluation protocols.

Related Work

In the following, we discuss other works related to our study where we we categorize these works into two categories: survival analysis methods and other machine learning supervised and unsupervised methods.

• *Survival analysis methods.* Survival analysis methods such as the Kaplan–Meier and Nelson–Aalen estimators, the proportional hazards models, etc, are the most common approach for predicting prognosis, relapse and other clinical outcomes of lung cancer patients. For example, **(author?)** [17] used the Cox hazard to predict long-term mortality/survival after lung cancer resection in patients older than 65 years on Medicare data for lung cancer resections. These models were also used in more specific survival analysis studies which focused on patients' survival in relation to specific bio-markers[(12)] and chemotherapy drug configurations[(11)]. Similarly, study conducted by Wen et.al. [(18)] in 393 North American patients with NSCLC used Cox proportional hazard model to investigate associations between functional genetic variants of autophagy-related genes and radiation pneumonitis as well as clinical outcomes after definitive radiotherapy. The multi-variable Cox model enabled them to predict radiation pneumonitis, local recurrence-free survival, progression-free survival, and overall survival of NSLC patients.

On a similar note, **(author?)** [19] used survival analysis methods in their study to examine the causes of death of long-term survivors of lung cancer where they utilized the Surveillance, Epidemiology and End Results (SEER) database[1] (1988–2008). They conducted a survival analysis using a Kaplan–Meier estimator and conducted a multivariate analysis using a Cox proportional hazard where the study concluded that cardiac as well as non-malignant pulmonary condition contributes considerable proportion of deaths in long-tern lung cancer survivors.

• *Supervised machine learning models.* Multiple studies of non-small cell lung cancer (NSCLC) postoperative recurrence have utilized supervised machine learning model to predict the probability of tumour recurrence and other

[1] http://seer.cancer.gov/seerstat

Characteristic		Recurrence	Survival	Total
		1275 (52.2%)	**1167 (47.8%)**	**2442 (100.0%)**
Age	Mean (range)	65.9 (25-89)	65.4 (26-117)	65.6 (25-117)
Gender	Male	989 (77.6)	861 (73.8%)	1850 (75.8%)
	Female	286 (22.4%)	306 (26.2%)	592 (24.2%)
Smoking history	Current/Previous	1115 (87.5%)	1000 (85.7%)	2115 (86.6%)
	Non smoker	160 (12.5%)	167 (14.3%)	327 (13.4%)
Cancer stage	I	9 (0.7%)	31 (2.7%)	40 (1.6%)
	IA	289 (22.7%)	353 (30.2%)	642 (26.3%)
	IB	348 (27.3%)	303 (26.0%)	651 (26.7%)
	II	18 (1.4%)	8 (0.7%)	26 (1.1%)
	IIA	246 (19.3%)	177 (15.2%)	423 (17.3%)
	IIB	365 (28.6%)	295 (25.3%)	660 (27.0%)
T stage	T1	333 (26.1%)	387 (33.2%)	720 (29.5%)
	T2	610 (47.8%)	475 (40.7%)	1085 (44.4%)
	T3 / T4	255 (20.0%)	175 (15.0%)	430 (17.6%)
N stage	N0	973 (76.3%)	891 (76.3%)	1864 (76.3%)
	N1	223 (17.5%)	141 (12.1%)	364 (14.9%)
	N2 / N3	13 (1.0%)	8 (0.7%)	21 (0.9%)
M stage	M0	1207 (94.7%)	1039 (89.0%)	2246 (92.0%)
	M1	8 (0.6%)	1 (0.1%)	9 (0.4%)
Tumor size	Mean (range)	36.0 (0.8-110.0)	33.2 (1.5-110.0	34.6 (0.8-110.0)
ECOG status	0	623 (48.9%)	765 (65.6%)	1388 (56.8%)
	1	558 (43.8%)	362 (31.0%)	920 (37.7%)
	2 / 3 / 4	91 (7.1%)	37 (3.2%)	128 (5.2%)
Tumor Differentiation	Poorly	98 (7.7%)	79 (6.8%)	177 (7.2%)
	Moderately	140 (11.0%)	171 (14.7%)	311 (12.7%)
	Well	102 (8.0%)	94 (8.1%)	196 (8.0%)

Table 1: Analysis of the characteristics of the patient cohort examined in our study in relation to tumor recurrence. The summation of the total percentages of the sub characteristics can be less than 100 for some of the examined features due to missing data values of this type of data for some patients EHRs.

related factors(14). For example, **(author?)** [20] have studied the associations between age, comorbidity, and other patient factors and treatment of postoperative NSCLC recurrence where they used a logistic regression to predict the postoperative prescribed treatment type – active or palliative. Supervised methods is able to operate on different types of data such as images, sequence and structured tabular data, therefore, they were utilized in different studies to consume heterogeneous data types. For example, they have been used for survival prediction of NSCLC patients from a combination of clinical and micro-array data (15). They were also used to model PET/CT images in order to predict earl stage NSCLC patients survival (21).

Patient Cohort Analysis

Our study uses electronic health records (EHRs) of patients which are collected and stored by Spanish Lung Cancer Group (SLCG) which is a multi-disciplinary group focused on finding better treatments for lung cancer. In the following, we discuss the patient cohort examined in our study and the features and characteristics extracted from the patient EHRs.

• *Patient cohort.* Our study is conducted on a cohort of 2442 early-stage (stage I or II) NSCLC patients where 1275 (52.2%) of these patient had tumor recurrence after successful treatment. The mean age of the patients in our dataset is 65.6 years with little difference between recurrence (65.9) and disease-free survival (65.4) patients. Males represent around 74% of the patient cohort while 85% of this cohort are current or previous smokers. All the examined patients are diagnosed with stage I and stage II NSCLC where patient are split between the two stages with a 54.6% for stage I and 45.4% for stage II. In terms of the TNM staging system, the vast majority of patients have are diagnosed with M stage and N stage of 0 (92% and 76.3% respectively) while the T stage diagnoses has a more even split. The mean tumor size was 34.6mm, where patients who suffered a recurrence had a slightly larger mean tumor size (36.0mm) compared to disease-free survival surviving patients who have a mean tumor size of 33.2mm with a similar range for both. The ECOG performance status of the patient cohort is mostly divided between the statuses 0 and 1 with approximately 57% and 38% respectively while other ECOG statuses 2,3 and 4 are associated with approximately 5% of the patients only. Table 1 provides a more detailed view of the characteristics of our patient cohort in relation to patient tumor recurrence and disease free survival.

• *Feature analysis.* The electronic health records of our patient cohort contain a different set of patient information and features such as the patient's demographics, diagnosis, bio-markers, treatments and follow up information. Patient history includes gender, smoking habits, familial cancer history and comorbidities. Diagnosis features detail information on the tumour classification, histology, etc, at time of diagnosis along with the symptoms the patient suffered. The bio-marker features record the results of any bio-marker analysis performed on the patient during the course of their treatments. The treatment records of a patient include the details of any chemotherapy, radiotherapy or surgical procedures the patient underwent during their treatment. Chemotherapy the features include the drugs used (including any maintenance drugs), the start and end dates of the treatment and the patients response to the chemotherapy. Radiotherapy features include the area radiated, the dose and the fractioning. Surgery features contain the type of surgical procedure, resection degree, the response and the post-surgeric TNM staging values.

• *Patient labelling.* In order to label patients as having relapsed (i.e., positive cases for training the machine learning models) we rely on 2 feature groups in the SLCG dataset. The primary source for labelling comes from the patients' progression records. If a patient has a progression record with status "Progression" or "Relapse" they are labelled positive. A secondary source for labelling comes from the patients' follow-up records. If a patient has a follow-up status of "Alive with disease" or a follow-up status of "Dead" with cause of death "Lung cancer" they are also labelled as positive. All other patients are considered negative examples for the purposes of training our models.

• *Patient consent.* The patient data used in this study was collected under the provisions of the Law 14/2007 on Biomedical Research at all times along with the confidentiality of the data of patients according to the requirements of the law 15/9 [54] on Protection of Personal Data and Implications of EU General Data Protection Regulation 2016/679 (GDPR), applicable from 25 May 2018, Directive 95/46/EC.

The data collected was identified with a code. Access to patients' personal information was restricted to the study doctors/investigators, the ethical clinical research committee of Puerta de Hierro-Majadahonda University Hospital and specialized authorized personnel to verify the study data and procedures, always keeping the confidentiality of the latter in accordance with current legislation. Only clinical data regarding the disease were used for subsequent analysis by the technical partners of the consortium, no personal data that could reveal the patients' identity was used. The signature of the informed consent for entry into the study is revocable since the patients can withdraw their consent at any time so that we stop using their personal data or their own samples for this research; this revocation can be done simply by notifying the investigator. The entry in this study was totally voluntary so in case of revocation or refusal to enter the study the patient would continue to be clinically treated by their doctor whether or not participating in the study.

Methods

In this section we discuss the data preparation and preprocessing step, analysis of the features included in the investigated patients EHRs and the learning models that we use in our study.

• *Data preprocessing.* We have applied a set of data preprocessing operations on the patients' EHRs to be consumable by our learning models and to decrease feature sparsity. We have fixed some corrupted date values where the date contains an invalid day or month value by replacing these values with a a placeholder value 1. We have also filled some

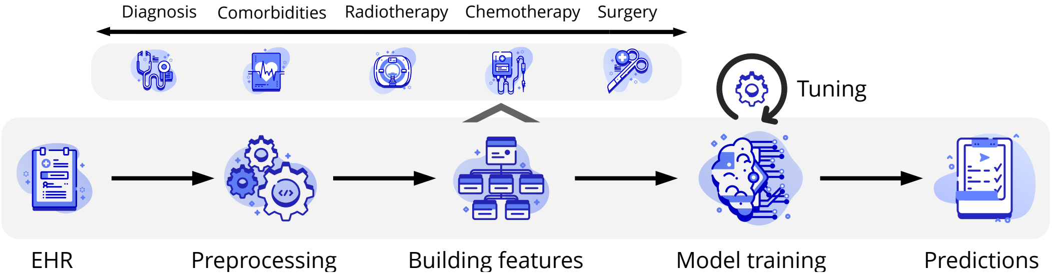

Figure 1: An illustration of the data processing and model training pipeline that we executed to apply the predictive models on the patient EHR data.

empty boolean fields with expected default values. For example, empty comorbidities annotations were assumed absent. This have been executed carefully with manual expert support to enrich the patient features extracted from their EHRs.

We have then formulated the patient records in a tabular dataset form, where we convert each patient record into a row of features. In this step, we have one included numerical and categorical feature values and we have excluded free text annotations and date values. We have attempted to preserve the order events in the patient row values by using different indexed columns for the same event which are indexed based on their order.

• *Data censorship.* In our study, we use the patient records to the full of their length to decide whether the patient had a tumour recurrence where we use all the available follow-up data points to detect recurrence. On the other hand, when build the patient features, we censor the patient data features using two criteria defined by experts to ensure that the learning models are only trained on features which are extracted from events successful treatment. First, if the patient had chemotherapy and radiotherapy treatments and did not have a surgery, we only consider the patient's features prior to the first successful chemotherapy (the chemotherapy which results in absence of tumor). Secondly, when the patient have a successful surgery, we only consider features prior to this surgery and including the surgery detail.

• *Learning models.* In this study, we use a set of four supervised machine learning models: Random Forest (RF), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) and Logistic Regression (LR). In the following, we provide a short description of each of these models and how they work:

- ⋆ Random Forests are supervised machine learning models which uses a large number of small decision trees called estimators to produce predictions, where aggregates the prediction of these estimators to produce a more robust overall prediction. Random Forests can provide explanations for their predictions based on its decision trees (estimators) feature weights.

- ⋆ Multi-Layer Perceptrons are artificial neural networks composed of multiple neural layers, where each layer is composed by an affine transformation followed by an element-wise non-linear operation. In our implementation, MLPs use stochastic gradient descent to optimize the log-loss on the training data.

- ⋆ Support Vector Machines are supervised machine learning models which performs binary classification by learning a linear decision boundary by optimizing a margin-based loss function.

- ⋆ Logistic Regression is a machine learning model which assumes a linear relation between the input features and the output predictions. The predictions of the logistic regression model are easily interpreted in terms of the coefficients corresponding to each of the input features contributing to the final model prediction.

• *Model training pipeline.* We trained the previously mentioned models on the patient EHRs using a multiphase-procedure as illustrated in Fig. 1, where we perform the data preprocessing and censorship as previously discussed. We then train the models using a k-fold cross-validation strategy, where models' hyperparameter tuning process is executed using the hyperparameters optimization framework optuna (22) across at least 10 trials per model class. Further discussion on the model training and evaluation process is included in the Results section.

Model	Accuracy	Precision	Recall	F1 Score	Avg. Precision	AUC-ROC
Random Forest	**0.69 (0.032)**	**0.71 (0.028)**	**0.74 (0.059)**	**0.72 (0.034)**	**0.67 (0.023)**	**0.68 (0.031)**
SVM	0.66 (0.015)	0.7 (0.015)	0.69 (0.027)	0.69 (0.016)	0.65 (0.011)	0.66 (0.015)
Neural Net. (MLP)	0.64 (0.019)	0.67 (0.019)	0.7 (0.026)	0.69 (0.018)	0.64 (0.014)	0.64 (0.019)
Logistic Regression	0.65 (0.019)	0.68 (0.018)	0.68 (0.026)	0.68 (0.019)	0.64 (0.014)	0.64 (0.019)

Table 2: A Comparison between four supervised learning models on 5-fold cross validation to predict recurrence in NSCLC patients as a binary classification task.

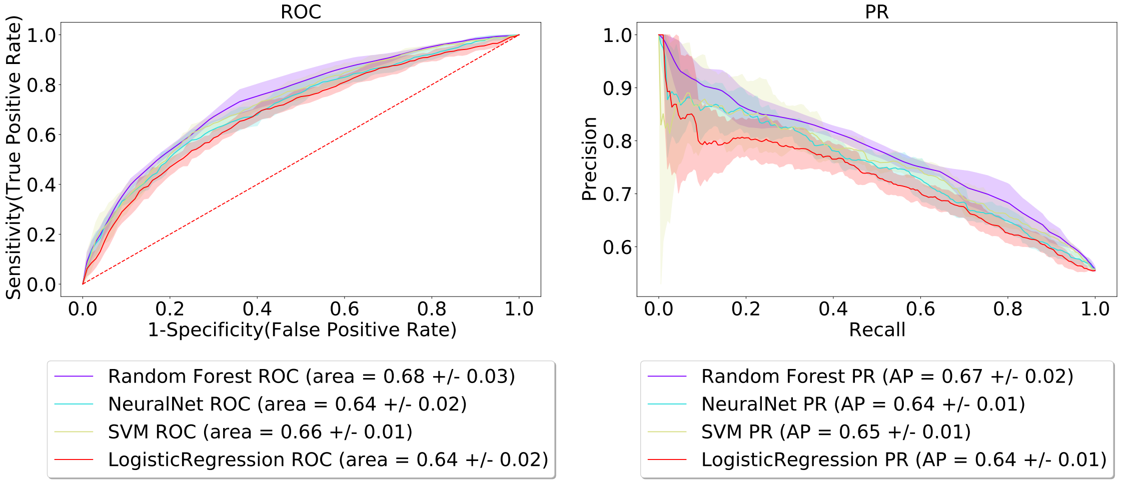

Figure 2: (a) Receiver operating curve (ROC) and (b) Precision-recall (PR) curve of the best performing models across 10 trials.

Results

In this section, we discuss the setup of our experiments where we describe the details of our evaluation protocol, metrics and the outcomes of our evaluation.

• *Evaluation protocol.* We cast the problem of assessing risk of tumour recurrence for a patient as a binary classification task which answers the question: whether a patient will have a recurrence or not. We execute our task on a cohort of 2242 patients where we use a 5-fold cross validation strategy. The 5-fold cross validation strategy is an evaluation technique where we split the patient cohort into 5 different random splits and evaluated our learning model on the five different splits independently. In each split evaluation, we use the other splits as training data and we evaluated our model on the specified splits. This technique produces a generalized view of the model predictive accuracy on the whole patient cohort with no data leakage between the training and testing cohorts.

• *Metrics.* We report the evaluation results of our examined models using a range of conventional binary classification metrics such as accuracy, precision, recall and the F1 score. We also use ranking metrics such as the average precision (i.e. the area under the precision-recall curve) and the are under the ROC curve(AUC-ROC).

• *Model hyperparameter tuning.* Our experiments were focused on four models as described in the methods section : Random Forest (RF), Logistic Regression (LR), Multi Layer Perceptron (MLP), and Support Vector Machine (SVM). The hyperparameters for these models were chosen using a hyperparameter search procedure executed using the hyperparameter optimization framework optuna [(22)]. For each model, we execute a set of 10 trials to find the optimal hyperparameter configuration where each trial corresponding to a single hyperparameter configuration chosen from a predefined search space. We then choose hyperparameters corresponding to the best performing configuration as the

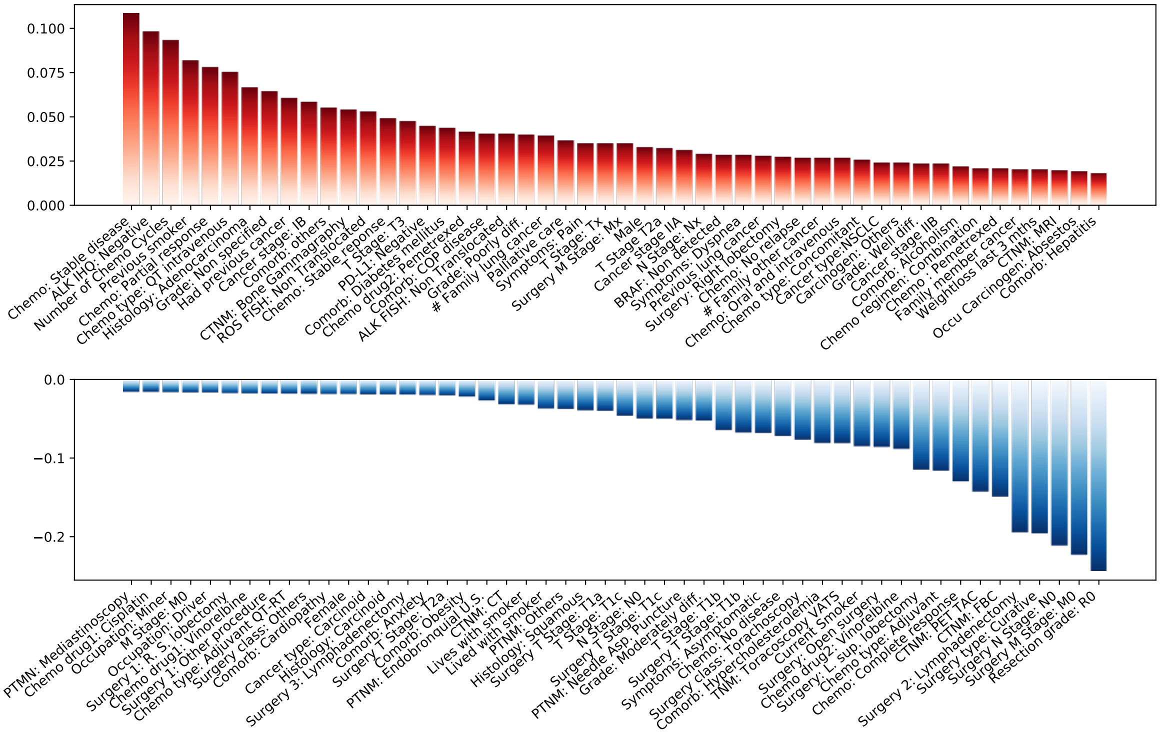

Figure 3: One of the best performing Logistic Regression model's coefficients. As a linear model, LR is interpretable through its coefficients that are tuned to give weights to the input features. Top chart presents features yielded by model as relevant (coefficients > 0) which contributes the most to recurrence prediction, while the bottom one non-relevant for prediction (coefs <= 0) which contributes the most to predicting disease-free survival.

model best hyperparameters for each of the examined models.

• *Evaluation results.* Table 2 shows the results of our computational experimental evaluation of the four examined models using a 5-fold cross validation evaluation strategy. The results show that the random forest model achieved the best results in terms of all the used evaluation metrics where it achieves 0.69, 0.71, 0.74, 0.72, 0.67 and 0.68 scores in terms of accuracy, precision, recall, F1 score, average precision and the are under the ROC curve metrics respectively. However, the results also show that the margin between the metric scores of the random forest model and the other examined models in not significantly large, where the standard deviation of the metric scores of all the models is 0.02, 0.02, 0.03, 0.02, 0.01 and 0.02 in respect to the accuracy, precision, recall, F1 score, average precision and the are under the ROC curve metrics scores respectively. The results also show that all the examined model have consistent results through the different 5-fold cross validation runs where the reported standard deviation of the their reported metrics is always less than 0.05.

In the Fig 2, we demonstrate plots of the ROC and precision recall curves of the examined models. The ROC curve shows that all the examined models a better predictive accuracy than random the random baseline with marginal differences between the performance of all the models. On the other hand, the precision-recall curve demonstrate some disparities between the model performance especially in terms of the models precision in relation to changes of the model's recall. For example, the logistic regression and SVM model suffer from a significant decrease of precision when the model's recall is less than 0.2 in comparison to other models.

• *Feature importance analysis.* In Fig. 3, we provide an example of model feature prioritization, where demonstrate the highest and lowest coefficient values learnt by the logistic regression model and their corresponding features. These coefficients represent the weight of contribution of their corresponding features towards the recurrence/survival predictions. The top part of figure demonstrates the features with the highest coefficient which contributes the most to predicting that a patient would have a recurrence. For example, the highest contributing feature to predicting recurrence is have a stable disease at the end of the last successful chemotherapy treatment. Other major contributing features include smoking, previous cancers, comorbidities such as hepatitis and COB disease, and large tumor size at diagnosis represented by the T stage. The bottom part of Fig. 3, however, demonstrates the features with the lowest coefficient (negative values) which contributes the most to predicting that a patient would have a disease-free survival. The figure shows that the most significant features in this category are outcomes of the last successful surgery *e.g.* TNM stages and resection grade, and the final response to last successful chemotherapy treatments.

Discussion

The main goal of this work has been to assess the suitability of off-the-shelf machine learning models to the task of predicting relapse in early stage non-small cell lung cancer patients. To that end, we have used data on a comprehensive patient cohort from the Medical Oncology Department of Hospital Universitario Puerta de Hierro Majadahonda in Madrid, Spain. While the standard methods used in this context (such as models based on Cox proportional hazard analysis) were reported to produce competitive results, they do have their limitations.

First, they use purely statistical, local features based on frequency for the predictions, and thus they cannot, by definition, use complex, long-range relationships between the particular data elements. This is an aspect some readily-available binary classification machine learning models such as neural networks may address better, as they can naturally model non-linear boundaries between the classes of relapsing and non-relapsing patients.

The Cox models also cannot work with heterogeneous data (e.g. patient data together with biomedical database and publication data), even though such data are more likely to provide a truly comprehensive representation of the patients and their disease. This is an aspect models based on extensible and semantically integrated machine learning data sets, which we plan to incorporate in the next stages of this work, can address more naturally than purely statistical models.

Furthermore, using predictive models based on machine learning allows for rapid prototyping and easy experimentation with a broad range of possibly complementary models, as shown in this preliminary work.

While a rigorous clinical validation of the model predictions is a part of our future work, our experiments have demonstrated the feasibility of the machine learning-based approach in the context of relapse prediction. All explored models significantly outperformed a naive baseline in standard machine learning metrics, and thus established a solid non-trivial stepping stone for further experiments in this area.

Perhaps the most important lesson learned was the crucial role of clinical data preprocessing with direct involvement of the oncologists, who were instrumental in defining the scope of features applicable for training the machine learning models. They were also essential for devising specific strategy for labelling positive and negative patient examples that is a must for optimal and clinically-relevant model performance. We believe the strategy is generally applicable across similar datasets and clinical needs. However, there is certainly lot of space for improvement in terms of data cleansing and conversion of qualitative values of patient features into quantitative data that are more suited for the machine learning models. This aspect would require a dedicated publication on its own and is a part of our future work.

Another critical aspect we intend to address in future research is the clinical validation of the predictive models. One would ideally like to compare the performance of the models with a human baseline, which is, however, possible only in a prospective study we are currently performing. In these settings, the models will be trained on retrospective data and tested on prospective patients, which will let us assess the capability of the models to predict high risk scores of patients who did indeed relapse, and low scores of patients who did not, in the follow-up period. This will allow for much more realistic assessment of the machine learning models, and overcome the main limitation of the presented preliminary work.

Funding

This paper is part of the CLARIFY project that has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 875160. It has also been funded by the Insight Centre for Data Analytics at National University of Ireland Galway (supported by the Science Foundation Ireland grant 12/RC/2289_2).

References

1. Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

2. Yaakov Tolwin, Roni Gillis, and Nir Peled. Gender and lung cancer—seer-based analysis. *Annals of Epidemiology*, 46:14–19, 2020.

3. Hidetaka Uramoto and Fumihiro Tanaka. Recurrence after surgery in patients with nsclc. *Translational lung cancer research*, 3(4):242, 2014.

4. Rodrigo Arriagada, Ariane Dunant, Jean-Pierre Pignon, Bengt Bergman, Mariusz Chabowski, Dominique Grunenwald, Miroslaw Kozlowski, Cécile Le Péchoux, Robert Pirker, Maria-Izabel Pinel, et al. Long-term results of the international adjuvant lung cancer trial evaluating adjuvant cisplatin-based chemotherapy in resected lung cancer. *J Clin Oncol*, 28(1):35–42, 2010.

5. Sarah Burdett, Jean Pierre Pignon, Jayne Tierney, Helene Tribodet, Lesley Stewart, Cecile Le Pechoux, Anne Aupérin, Thierry Le Chevalier, Richard J Stephens, Rodrigo Arriagada, et al. Adjuvant chemotherapy for resected early-stage non-small cell lung cancer. *Cochrane Database of Systematic Reviews*, (3), 2015.

6. Anne Aupérin, Cecile Le Péchoux, Estelle Rolland, Walter J Curran, Kiyoyuki Furuse, Pierre Fournel, Jose Belderbos, Gerald Clamon, Hakki Cuneyt Ulutin, Rebecca Paulus, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet]*, 2010.

7. Riad N Younes, Jefferson L Gross, and Daniel Deheinzelin. Follow-up in lung cancer: how often and for what purpose? *Chest*, 115(6):1494–1499, 1999.

8. Fabrice Denis, Louise Viger, Alexandre Charron, Eric Voog, Olivier Dupuis, Yoann Pointreau, and Christophe Letellier. Detection of lung cancer relapse using self-reported symptoms transmitted via an internet web-application: pilot study of the sentinel follow-up. *Supportive Care in Cancer*, 22(6):1467–1473, 2014.

9. Timothy J Judson, Antonia V Bennett, Lauren J Rogak, Laura Sit, Allison Barz, Mark G Kris, CliffordA Hudis, Howard I Scher, Paul Sabattini, Deborah Schrag, et al. Feasibility of long-term patient self-reporting of toxicities from home via the internet during routine chemotherapy. *Journal of clinical oncology*, 31(20):2580, 2013.

10. Vinod P Balachandran, Mithat Gonen, J Joshua Smith, and Ronald P DeMatteo. Nomograms in oncology: more than meets the eye. *The lancet oncology*, 16(4):e173–e180, 2015.

11. Chee Khoon Lee, Lucy Davies, Yi-Long Wu, Tetsuya Mitsudomi, Akira Inoue, Rafael Rosell, Caicun Zhou, Kazuhiko Nakagawa, Sumitra Thongprasert, Masahiro Fukuoka, et al. Gefitinib or erlotinib vs chemotherapy for egfr mutation-positive lung cancer: individual patient data meta-analysis of overall survival. *JNCI: Journal of the National Cancer Institute*, 109(6), 2017.

12. Margarita Kirienko, Luca Cozzi, Lidija Antunovic, Lisa Lozza, Antonella Fogliata, Emanuele Voulaz, Alexia Rossi, Arturo Chiti, and Martina Sollini. Prediction of disease-free survival by the pet/ct radiomic signature in non-small cell lung cancer patients undergoing surgery. *European journal of nuclear medicine and molecular imaging*, 45(2):207–217, 2018.

13. Weiran Zhang, Xuefeng Lin, Xin Li, Meng Wang, Wei Sun, Xingpeng Han, and Daqiang Sun. Survival prediction model for non-small cell lung cancer based on somatic mutations. *The journal of gene medicine*, 22(9):e3206, 2020.

14. Mohamed S Barakat, Matthew Field, Aditya Ghose, David Stirling, Lois Holloway, Shalini Vinod, Andre Dekker, and David Thwaites. The effect of imputing missing clinical attribute values on training lung cancer survival prediction model performance. *Health information science and systems*, 5(1):1–11, 2017.

15. Yu-Heng Lai, Wei-Ning Chen, Te-Cheng Hsu, Che Lin, Yu Tsao, and Semon Wu. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific reports*, 10(1):1–11, 2020.

16. Bora Lee, Sang Hoon Chun, Ji Hyung Hong, In Sook Woo, Seoree Kim, Joon Won Jeong, Jae Jun Kim, Hyun Woo Lee, Sae Jung Na, Kyongmin Sarah Beck, et al. Deepbts: prediction of recurrence-free survival of non-small cell lung cancer using a time-binned deep neural network. *Scientific reports*, 10(1):1–10, 2020.

17. Eisuke Kato, Noboru Takayanagi, Yotaro Takaku, Naho Kagiyama, Tetsu Kanauchi, Takashi Ishiguro, and Yutaka Sugita. Incidence and predictive factors of lung cancer in patients with idiopathic pulmonary fibrosis. *ERJ open research*, 4(1), 2018.

18. Juyi Wen, Hongliang Liu, Lili Wang, Xiaomeng Wang, Ning Gu, Zhensheng Liu, Ting Xu, Daniel R Gomez, Ritsuko Komaki, Zhongxing Liao, et al. Potentially functional variants of atg16l2 predict radiation pneumonitis and outcomes in patients with non–small cell lung cancer after definitive radiotherapy. *Journal of Thoracic Oncology*, 13(5):660–675, 2018.

19. Omar Abdel-Rahman. Causes of death in long-term lung cancer survivors: a seer database analysis. *Current medical research and opinion*, 33(7):1343–1348, 2017.

20. Melisa L Wong, Timothy L McMurry, George J Stukenborg, Amanda B Francescatti, Carla Amato-Martz, Jessica R Schumacher, George J Chang, Caprice C Greenberg, David P Winchester, Daniel P McKellar, et al. Impact of age and comorbidity on treatment of non-small cell lung cancer recurrence following complete resection: A nationally representative cohort study. *Lung Cancer*, 102:108–117, 2016.

21. Mehdi Astaraki, Chunliang Wang, Giulia Buizza, Iuliana Toma-Dasu, Marta Lazzeroni, and Örjan Smedby. Early survival prediction in non-small cell lung cancer from pet/ct images using an intra-tumor partitioning method. *Physica Medica*, 60:58–65, 2019.

22. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework.