SUPPLEMENTARY MATERIALS:
A novel analytical framework for risk stratification of real-world data using machine learning: a small cell lung cancer (SCLC) study

Luca Marzano*, Adam S. Darwicha, Salomon Tendlerb, Asaf Danb, Rolf Lewensohnb, Luigi De Petrisb, Jayanth Raghothamaa, Sebastiaan Meijera

*Corresponding author: Luca Marzano, +4672844279, lmarzano@kth.se

aDivision of Health Informatics and Logistics, School of Engineering Sciences in Chemistry, Biotechnology and Health (CBH), KTH Royal Institute of Technology, Huddinge, Sweden
bDept. of Oncology-Pathology, Karolinska Institutet and the Thoracic Oncology Center, Karolinska University hospital, Stockholm, Sweden

S1. Code implementation

The model was implemented in RStudio environment [1]. The following libraries were used: missForest (missing imputation), survival (Cox regression), ranger (Random Survival Forest), and cluster (Gower distance and Partition Around Medoids).

S2. Multiple Missing Imputation

The entire cohort before the patients’ selection for the study showed 8% rate of missing values for clinical covariates. The majority of missing values were chemical and blood values (35% of missingness within Albumin, LD, HB, CRP and Na). Within the rest of other covariates, the missingness was low (1%). This was handled with missing forest algorithm [2]. This algorithm exploits the mechanism of Random Forest in a missing multiple imputation task. The observed part is used to fit a random forest and then predict the missing values. The model is subsequently trained from previous step results in order to perform missing imputation on better quality data. The process is made until
a stopping criterion is reached. In detail, missing forest algorithm works in
the following way: dataset is initialized; to each missing variable is assigned
the mean, if numeric, or the mode, if categorical, of non-missing observations;
a random forest is fitted with observed values; missing values are predicted
with the trained model, the dataset is updated with the new imputed val-
ues, the process is repeated until difference between imputed variables of two
subsequent iterations begin to increase.
We chose this approach among other missing imputation methods because
this algorithm works with mixed data and for the high predictive power of
random forest. Indeed, several studies demonstrated that missing forest out-
performed other standard algorithms [2, 3, 4, 5].
This algorithm showed better performance with respect to MICE algorithm
in a previous study with the majority of patients of this cohort (n=377) [6].
Missing forest does not require assumptions about the random distribution
of missing values [7], thus handling missing not at random (MNAR) values.
Another assumption not required is relationship between features because is
a non-parametric method (e.g., MICE algorithm assumes linearity) and its
capability of work with non-linear effects.
We found as best configuration a random forest trained with 100 decision
trees sampling 4 categorical variables and 5 numeric variables at each split.
Out-of-bag error estimation was used to assess the best configuration. Nor-
malized root mean squared error (NRMSE) for numerical variables and pro-
portion of falsely classified (PFC) for categorical variables were computed.
The configuration adopted yielded NRMSE= 0.06 and PFC= 0.15.

S3. Clinical baselines
Input covariates of the cohort that were used to cluster the patients. T8,
N8, M8 and ST8 covariates refers to the 8th version of the TNM staging
descriptors [8]. According to medical knowledge, VALSG LD (ED) staging
 correspond to I-III (IV) TNM staging. Therefore, the studied cohort encom-
passed n=185 LD and n=451 ED patients.

S4. Covariate selection: Cox Proportional Hazards regression
Cox regression is a semi-parametric survival analysis method. The main
assumption is that the hazard function h(t), the probability that a patient
|                  | No treatment | CT + RT | CT          |
|------------------|--------------|---------|-------------|
|                  | (N=89)       | (N=72)  | (N=389)     |
| PCI              |              |         |             |
| age              |              |         |             |
| Mean (SD)        | 72.9 (7.94)  | 68.9 (9.74) | 64.6 (8.43) |
| Median [Min, Max]| 74.0 [51.0, 80.0] | 70.0 [47.0, 81.0] | 65.0 [37.0, 81.0] |
| gender           |              |         |             |
| F                | 47 (52.8%)   | 42 (58.3%) | 218 (56.0%) |
| M                | 42 (47.2%)   | 30 (41.7%) | 171 (44.0%) |
| ECOG             |              |         |             |
| 0                | 2 (2.2%)     | 7 (58.3%) | 45 (62.5%)  |
| 1                | 18 (20.2%)   | 2 (16.7%) | 26 (36.1%)  |
| 2                | 25 (28.1%)   | 3 (25.0%) | 1 (1.4%)    |
| 3                | 44 (49.5%)   | 0 (0%)   | 57 (14.7%)  |
| T8 tumor size    |              |         |             |
| T0               | 0 (0%)       | 0 (0%)   | 1 (1.4%)    |
| T1a              | 0 (0%)       | 0 (0%)   | 1 (1.4%)    |
| T1b              | 3 (3.4%)     | 1 (8.3%) | 8 (11.1%)   |
| T1c              | 5 (5.6%)     | 4 (33.3%) | 7 (9.7%)    |
| T2a              | 1 (1.1%)     | 0 (0%)   | 5 (6.9%)    |
| T2b              | 3 (3.4%)     | 1 (8.3%) | 3 (4.2%)    |
| T3               | 18 (20.2%)   | 2 (16.7%) | 9 (12.5%)   |
| T4               | 59 (66.3%)   | 4 (33.3%) | 37 (54.4%)  |
| N8 nodules       |              |         |             |
| N0               | 11 (12.4%)   | 5 (6.9%) | 37 (9.5%)   |
| N1               | 3 (3.4%)     | 1 (8.3%) | 9 (2.3%)    |
| N2               | 31 (34.8%)   | 33 (45.8%) | 138 (35.5%) |
| M8 metastasis    |              |         |             |
| M0               | 12 (13.5%)   | 67 (93.1%) | 71 (18.3%)  |
| M1A              | 9 (10.4%)    | 30 (7.7%) | 10 (2.6%)   |
| M1B              | 4 (4.5%)     | 3 (4.2%) | 13 (3.1%)   |
| M1C              | 64 (71.9%)   | 265 (62.7%) | 275 (70.7%) |
| ST8 cancer stage |              |         |             |
| IIIA             | 1 (1.1%)     | 21 (29.2%) | 23 (5.9%)   |
| IIIB             | 4 (4.5%)     | 29 (40.3%) | 16 (4.1%)   |
| IIIC             | 7 (7.9%)     | 16 (22.2%) | 32 (8.2%)   |
| IVA              | 13 (14.6%)   | 141 (31.5%) | 141 (31.5%) |
| IVB              | 64 (71.9%)   | 274 (70.4%) | 32 (43.2%)  |
| LDH (log) µkat/L |              |         |             |
| Mean (SD)        | 0.919 (0.315) | 0.550 (0.129) | 0.550 (0.129) |
| Median [Min, Max]| 0.903 [0.301, 1.89] | 0.477 [0.301, 1.32] | 0.477 [0.301, 1.32] |
| CRP (log) mg/L   |              |         |             |
| Mean (SD)        | 1.33 [0.243] | 1.07 [0.466] | 1.22 [0.574] |
| Median [Min, Max]| 1.02 [0.243] | 1.20 [0.243] | 1.20 [0.243] |
| Albumin g/L      |              |         |             |
| Mean (SD)        | 29.8 (6.02)  | 34.3 (2.98) | 34.3 (2.98) |
| Median [Min, Max]| 34.5 [30.0, 38.0] | 34.0 [22.0, 42.0] | 34.0 [22.0, 42.0] |
| HB g/L           |              |         |             |
| Mean (SD)        | 131 (17.8)   | 127 (15.9) | 128 (15.4)  |
| Median [Min, Max]| 129 [10.0, 15.0] | 125 [7.00, 15.0] | 125 [7.00, 15.0] |
| Na mmol/L        |              |         |             |
| Mean (SD)        | 137 (5.65)   | 137 (3.41) | 137 (4.72)  |
| Median [Min, Max]| 137 [116, 151] | 137 [115, 147] | 137 [115, 147] |
| PET CT           |              |         |             |
| No               | 81 (91.0%)   | 33 (33.3%) | 11 (25.5%)  |
| Yes              | 8 (9.0%)     | 26 (66.7%) | 32 (74.5%)  |
| Brain CT         |              |         |             |
| No               | 57 (64.0%)   | 7 (58.3%) | 47 (65.3%)  |
| Yes              | 32 (36.0%)   | 25 (41.7%) | 24 (34.7%)  |

Table S1: Clinical baselines collected before treatment decision.\(^a\). In brackets: percentage of patients.

\(^a\)SD: standard deviation
of the cohort have not experienced an event (death or censoring) at a time \( t \) has an event at that time, has the following form:

\[
h(t) = h_0(t) e^{(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)} = h_0(t) e^{(\sum \beta_i X_i)}
\]  

(1)

Where coefficients \( \beta_1, \ldots, \beta_p \) measure the impact of \( X_1, \ldots, X_n \) covariates. \( h_0(t) \) represents baseline hazard: the value of all the covariates is equal to zero. Cox model could be formulated as a multivariate linear regression of the logarithm of hazard. Since \( h_0(t) \) is unspecified, Cox hazard model is a semi-parametric method. Values of \( \beta \) greater than zero are associated to a bad prognostic factor. \( \beta \) that tends to zero, means that the covariates does not affect the hazard. Given two patients \( j \) and \( k \), the hazard ratio of i-th covariate is:

\[
\frac{h_0(t) e^{(\beta_i X_{ij})}}{h_0(t) e^{(\beta_i X_{ik})}} = e^{\beta_i (X_{ij} - X_{ik})}
\]  

(2)

As we can see, hazard ratios are independent from baseline hazard, and so, independent of time.

For categorical variables, in any group of patients with same value, \( X_i \) the hazard is computed comparing other groups with different values for \( X_i \). A value of \( X_i \) is chosen as reference to compute hazard ratios respect other values that variable can assume. Then, an hazard ratio greater than one means that its associated level is associated to bad prognosis in comparison with the reference level. When \( X_i \) is a continuous variable, the hazard ratio reports the change in the risk of death if the parameter in question rises. Therefore, a hazard ratios greater than 1 means that as \( X_i \) increase, the probability of event’s occurrence increase with that hazard ratios’ rate.

The statistical confidence level of hazard ratios to select covariates was quantified by Wald test. Furthermore, we computed the concordance index with Harrell method \([9]\).

S5. Covariate importance: Random Survival Forest

Parameters that categorise the algorithm are number of trees, \( m \) number of sampled features to built the forest’s trees and \( d_0 \) number of patients in the terminal node. For the analysis number of trees was fixed ad 100 and \( m \) and \( d_0 \) were chosen with a grid search that minimize out-of-bag error (concordance index \([10]\)). Parameters explore were \( m = [2, n_{Cox} - 1] \), where \( n_{Cox} \) is the
number of covariates selected by Cox regression, and $d_0 = [1, 3, 5, 10]$. The optimal was with $m = 3$ and $d_0 = 5$.

Feature importance proposed by Fisher et al. [11] was computed to measure the relative difference in prognosis impact between the covariates. Given the $j-th$ covariate, its feature importance is computed randomly shuffling its values in the data, computing the performance of the model with this corrupted value and then measuring the difference without the shuffling of the variable. The operation is repeated $K$ times. So, variable importance $VI$ of $j-th$ feature is

$$VI_j = C - \frac{1}{K} \sum_{k=1}^{K} C_{k,j}$$

where $C$ is the concordance index proposed by Harrell et al. [10] of the model without permutation, and $C_{k,j}$ is the concordance of $k-th$ shuffling of the $j-th$ feature.

We performed 100 times the analysis to compute robust values of feature importance. We indicated the mean of the feature importance of the $j-th$ covariate among these iterations with $\bar{VI}_j$.

S6. Unsupervised Machine Learning: Gower distance and Partition Around Medoids

For all patients we computed Gower distance proposed by Gower [12]. Given two patients $i$ and $j$ described by $n$ covariates, Gower distance $d(i, j)$ is defined as follows:

$$d(i, j) = \sum_{k=1}^{n} \frac{w_k d_k(i, j)}{w_k}$$

where $d_k(i, j)$ is the contribute of $k-th$ covariate and $w_k$ its associated weight. Given $X_{ik}$ and $X_{jk}$, $k-th$ covariates of $i-th$ and $j-th$ patient respectively, if the interested covariate is continuous:

$$d_k(i, j) = \frac{|X_{ik} - X_{jk}|}{R_k}$$

where $R_k$ is the range of $k-th$ covariate. If the variable is categorical and
not-ordered:

\[ d_k(i,j) = \begin{cases} 
0 & \text{if } X_{ik} = X_{jk}; \\
1 & \text{otherwise.} 
\end{cases} \]  

and finally, if the variable is categorical and ordered, \( X_{ik} \) and \( X_{jk} \) are replaced by their respective order integer \( r_{ik} \) and \( r_{jk} \) and \( d_k(i,j) \) is computed as continuous contribute with equation 5.

For each covariate, the weight \( w_k \) was assigned from the mean of Random Survival Forest average feature importance:

\[ w_k = \bar{VI}_k \]  

\( d(i,j) \) is a normalized, it can assume values from 0 to 1 where a pairwise score of 0 indicates completely similarity between two patients.

PAM clustering steps are i) \( k \) number of clusters is choosen; ii) \( k \) random medoids are selected; iii) associate to each observation the closest medoid; iv) for each medoid \( m \) and non-medoid \( p \), the cost function \( S \) is computed for both configuration with \( m \) medoid (\( C_i \)) and \( p \) new medoid (\( C_f \)). The cost function is defined as the sum of distance from points to their medoids; v) if \( S_f - S_i < 0 \), \( p \) is set as new medoid; vi) points ii)-v) are repeated until there is no more exchange of medoids.

We used three indexes to estimate the optimal number of clusters \( k \) in terms of compactness and goodness of partition: average silhouette width, Dunn coefficient and the within sum of square distances [13]. Silhouette coefficients can assume values in \([-1, 1]\); larger values indicates good cluster partitions. High values of Dunn index indicates compacted and well separated clusters. Optimal \( k \) from within sum of squares is usually found with “elbow” method.

S7. Optimal number of clusters

In Figure S1 are reported Average Silhouette between clusters, Dunn coefficient and within cluster sum of squares obtained with \( k \) number of clusters from 2 to 15. The optimal configuration was found with \( k = 7 \).

S7.1. comparison with other clustering methods

In Figure S2 we compared our unsupervised approach (PAM-cov.imp) with other strategies: hierarchical clustering with RSF covariate impact
(hier-cov.imp), traditional partition around medoids and hierarchical clustering without RSF covariate impact (PAM and hier). Our approach provided better clustering performance.

S8. Cluster survival analysis: Kaplan-Meier curves

In Figure S3 we reported the Kaplan-Meier curves of patients with OS < 2 years (n=546). We set this limit to plot curves in the same time range. The patients with OS > years were mostly in cluster7 (n=39) and cluster2 (n=29). The rest of patients not included in the evaluation of Kaplan-Meier curves were in cluster1 (n=8), cluster5 (n=12) and cluster6 (n=2). The curves were statistically significant according to log-rank test.

S9. Clinical and treatment patterns

Clinical and treatment patterns of the detected clusters are shown in Table S2.

Kruskal-Wallis showed that LDH (p < 2.2e^{-16}) and CRP (p < 0.002) were statistical significant in the clusters. These blood chemistry measures were associated with poor prognosis.

S9.1. Chemotherapy cycles

In table S3 we reported patterns regarding chemotherapy lines across the clusters. The majority of patients received only one line of PE. Re-challenge with PE (PE-PE) was administrated mostly in cluster1, cluster2 and cluster7. PE-IP was administrated in same clusters of PE-PE and cluster5. PE-monotherapy was administrated in all subgroups except cluster3. Other chemotherapy strategies were found in all clusters except cluster6 and cluster7. In table S4 we compared treatment outcomes of platinum re-challenge for cluster1, cluster2, cluster5 and cluster7. We indicated with PFS 1L the progress free survival of the first line cycle (the interval between the start of the therapy and the earliest documented clinical progress, or death). Instead, PFS 2L indicated the progress free survival of the second cycle (the interval between the start of the second cycle and the earliest documented clinical progress, or death). PE-PE outperformed PE-IP in terms of overall survival and PFS 1L. However, PFS 2L median was higher for cluster1 and cluster2 and comparable for cluster5 and cluster7. The few patients receiving the third line were scattered across cluster1, cluster2, cluster5 and cluster7.
Table S2: treatment decision and selected covariates of the clusters

| Clusters | PCI | No 119 (85.0%) | Yes 21 (15.0%) |
|----------|-----|---------------|---------------|
| Cluster1 | (N=140) | 44 (48.4%) | 3 (3.1%) |
| Cluster2 | (N=90) | 94 (96.9%) | 3 (3.1%) |
| Cluster3 | (N=97) | 84 (96.6%) | 4 (4.2%) |
| Cluster4 | (N=87) | 74 (87.1%) | 11 (12.9%) |
| Cluster5 | (N=85) | 42 (91.3%) | 4 (8.7%) |
| Cluster6 | (N=46) | 57 (63.3%) | 25 (23.9%) |
| Cluster7 | (N=90) | 29 (63.3%) | 14 (30.0%) |

**First line therapy**

| CT + RT | 1 (0.7%) | 27 (29.7%) | 0 (0%) |
|---------|----------|-------------|-------|
| No Treatment | 124 (88.6%) | 61 (67.0%) | 54 (59.3%) |

**ECOG**

| ECOG | 0 | 1 | 2 | 3 |
|------|---|---|---|---|
| No 119 (85.0%) | 0 (0%) | 21 (15.0%) | 1 (0.7%) | 0 (0%) |
| Yes 21 (15.0%) | 44 (48.4%) | 47 (51.6%) | 3 (3.1%) | 1 (0.7%) |

**Tumor size**

| T0 | 1 (0.7%) | 1 (0.7%) | 7 (9.2%) | 3 (3.3%) |
|----|---------|---------|--------|--------|
| No 119 (85.0%) | 0 (0%) | 43 (44.3%) | 3 (3.1%) | 33 (36.7%) |
| Yes 21 (15.0%) | 43 (48.4%) | 3 (3.1%) | 0 (0%) | 57 (63.3%) |

**Metastasis**

| M0 | 68 (74.7%) | 70 (76.9%) | 61 (62.9%) | 44 (50.6%) |
|----|------------|------------|------------|------------|
| No 119 (85.0%) | 3 (3.1%) | 52 (14.0%) | 15 (14.3%) | 12 (11.8%) |
| Yes 21 (15.0%) | 2 (2.2%) | 12 (11.8%) | 3 (3.1%) | 1 (0.7%) |

**Cancer stage**

| T0 | 119 (85.0%) | 90 (69.5%) | 97 (97.2%) | 87 (96.6%) |
|----|-------------|------------|------------|------------|
| No 119 (85.0%) | 83 (22.4%) | 54 (14.6%) | 54 (14.6%) | 54 (14.6%) |
| Yes 21 (15.0%) | 19 (28.6%) | 12 (18.5%) | 12 (18.5%) | 12 (18.5%) |

Two patients receiving the fourth line were in cluster7, while cluster2 and cluster5 had one isolated case.

Table S3: chemotherapy lines across the clusters

```sql
| No treatment | PE (N=89) | PE PE (N=371) | PE IP (N=64) | PE monotherapy (N=66) | Others (N=20) |
|-------------|-----------|---------------|--------------|-----------------------|--------------|
| cluster     |           |               |              |                       |              |
| 1           | 15 (16.9%)| 83 (22.4%)    | 19 (28.6%)   | 5 (25.0%)             | 6 (25.0%)    |
| 2           | 3 (3.4%)  | 52 (14.0%)    | 12 (18.5%)   | 5 (25.0%)             | 4 (15.4%)    |
| 3           | 41 (43.3%)| 49 (13.2%)    | 0 (0%)       | 0 (0%)                | 4 (15.4%)    |
| 4           | 18 (20.2%)| 54 (14.6%)    | 3 (4.7%)     | 3 (15.0%)             | 5 (19.2%)    |
| 5           | 2 (2.2%)  | 46 (12.4%)    | 9 (14.1%)    | 4 (15.4%)             | 6 (23.3%)    |
| 6           | 6 (9.0%)  | 30 (8.1%)     | 5 (7.8%)     | 1 (1.5%)              | 2 (10.0%)    |
| 7           | 0 (0%)    | 57 (15.4%)    | 20 (31.3%)   | 11 (16.7%)            | 1 (1.8%)     |
|             | 89 (89.0%)| 371 (371.0%)  | 64 (64.0%)   | 66 (66.0%)            | 20 (20.0%)   |
```

8
Table S4: Platinum re-challenge outcomes for clusters1, cluster2, cluster5 and cluster7.

S9.2. Subcohorts-competitive therapies

In cluster2 and cluster7 sub-cohorts CT+RT provided better prognosis compared to CT alone. Hazard ratios of patients treated with CT were 2.84 [1.66, 4.87] for cluster2 and 4.81 [2.91, 7.95] for cluster7. In cluster2 most of patients received CT rather than CT+RT. Indeed, most of them had stage IIIC and IVA. However, also patients having IIIA and IIIB received CT most probably because of the size of the tumor. Some IIIC patients received CT+RT. In cluster7 IIIA and IIIB were the the majority of patients receiving CT+RT, while IIIC was the dominant stage for CT. However, there were IIIC receiving CT+RT also in this cluster.

The hazard ratios of patients treated with CT alone was statistically significantly different compared to patients who did not received treatment in cluster3 and cluster4: (0.33 [0.21, 0.52] for cluster3 and 0.10 [0.05, 0.20] for cluster4). In cluster3 and cluster4 were not found marked differences in TNM stages.
References

[1] RStudio Team, RStudio: Integrated Development Environment for R, RStudio, PBC., Boston, MA, 2020. URL: http://www.rstudio.com/.

[2] D. J. Stekhoven, P. Bühlmann, Missforest-Non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2012) 112–118. doi:10.1093/bioinformatics/btr597.

[3] F. Tang, H. Ishwaran, Random forest missing data algorithms, Statistical Analysis and Data Mining: The ASA Data Science Journal 10 (2017) 363–377.

[4] A. Pantanowitz, T. Marwala, Missing data imputation through the use of the random forest algorithm, in: Advances in Computational Intelligence, Springer, 2009, pp. 53–62.

[5] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, H. Hemingway, Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study, American journal of epidemiology 179 (2014) 764–774.

[6] Luca Marzano, Adam Darwich, Salomon Tendler, Asaf Dan, Rolf Lewensohn, Luigi De Petris, Jayanth Raghothama, Sebastiaan Meijer, Role of tree-based machine learning algorithms in prognostic pattern detection: an Extensive-Disease Small Cell Lung Cancer (ED-SCLC) study, 2021. URL: https://www.page-meeting.org/default.asp?abstract=9742.

[7] D. B. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592.

[8] W. Lim, C. A. Ridge, A. G. Nicholson, S. Mirdadraee, The 8th lung cancer TNM classification and clinical staging system: review of the changes and clinical implications, Quantitative Imaging in Medicine and Surgery 8 (2018) 709. doi:10.21037/QIMS.2018.08.02.

[9] N. E. Breslow, Analysis of Survival Data under the Proportional Hazards Model, International Statistical Review / Revue Internationale de Statistique 43 (1975) 45. doi:10.2307/1402659.
[10] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the Yield of Medical Tests, JAMA 247 (1982) 2543–2546. doi:10.1001/JAMA.1982.03320430047030.

[11] A. Fisher, C. Rudin, F. Dominici, All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously, Journal of Machine Learning Research 20 (2019) 1–81.

[12] J. C. Gower, A general coefficient of similarity and some of its properties, Biometrics (1971) 857–871.

[13] R. Xu, D. C. Wunsch, Clustering algorithms in biomedical research: A review, IEEE Reviews in Biomedical Engineering 3 (2010) 120–154. doi:10.1109/RBME.2010.2083647.