**Thoracic Surgery, binary survival – data description**

**Source:**

Creators: Marek Lubicz (1), Konrad Pawelczyk (2), Adam Rzechonek (2), Jerzy Kolodziej (2)  
-- (1) Wroclaw University of Technology, wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland  
-- (2) Wroclaw Medical University, wybrzeze L. Pasteura 1, 50-367 Wroclaw, Poland  
  
Donor: Maciej Zieba (maciej.zieba **'@'** pwr.wroc.pl), Jakub M. Tomczak (jakub.tomczak **'@'** pwr.wroc.pl), (+48) 71 320 44 53  
  
Date: November, 2013

**Data Set Information:**

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007â€“2011. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

**Attribute Information:**

1. DGN: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)  
2. PRE4: Forced vital capacity - FVC (numeric)  
3. PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)  
4. PRE6: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)  
5. PRE7: Pain before surgery (T,F)  
6. PRE8: Haemoptysis before surgery (T,F)  
7. PRE9: Dyspnoea before surgery (T,F)  
8. PRE10: Cough before surgery (T,F)  
9. PRE11: Weakness before surgery (T,F)  
10. PRE14: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)  
11. PRE17: Type 2 DM - diabetes mellitus (T,F)  
12. PRE19: MI up to 6 months (T,F)  
13. PRE25: PAD - peripheral arterial diseases (T,F)  
14. PRE30: Smoking (T,F)  
15. PRE32: Asthma (T,F)  
16. AGE: Age at surgery (numeric)  
17. Risk1Y: 1 year survival period - (T)rue value if died (T,F)  
  
Class Distribution: the class value (Risk1Y) is binary valued.  
Risk1Y Value: Number of Instances:  
T 70  
N 400  
  
Summary Statistics:  
  
Binary Attributes Distribution:  
PRE7 Value: Number of Instances:  
T 31  
N 439  
PRE8 Value: Number of Instances:  
T 68  
N 402  
PRE9 Value: Number of Instances:  
T 31  
N 439  
PRE10 Value: Number of Instances:  
T 323  
N 147  
PRE11 Value: Number of Instances:  
T 78  
N 392  
PRE17 Value: Number of Instances:  
T 35  
N 435  
PRE19 Value: Number of Instances:  
T 2  
N 468  
PRE25 Value: Number of Instances:  
T 8  
N 462  
PRE30 Value: Number of Instances:  
T 386  
N 84  
PRE32 Value: Number of Instances:  
T 368  
N 2  
  
Nominal Attributes Distribution:  
DGN Value: Number of Instances:  
DGN3 349  
DGN2 52  
DGN4 47  
DGN6 4  
DGN5 15  
DGN8 2  
DGN1 1  
PRE6 Value: Number of Instances:  
PRZ2 27  
PRZ1 313  
PRZ0 130  
PRE14 Value: Number of Instances:  
OC11 177  
OC14 17  
OC12 257  
OC13 19  
  
Numeric Attributes Statistics:  
Min Max Mean SD  
PRE4: 1.4 6.3 3.3 0.9  
PRE5: 0.96 86.3 4.6 11.8  
AGE: 21 87 52.5 8.7

**Relevant Papers:**

ZiÄ™ba, M., Tomczak, J. M., Lubicz, M., & ÅšwiÄ…tek, J. (2013). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing. [[Web Link]](http://dx.doi.org/10.1016/j.bbr.2011.03.031)  
- Results:  
-- Boosted SVM for for imbalanced data gained the Gmean value equal 0.657,  
-- Decision rules induced using Boosted SVM as an oracle gained the Gmean value equal 0.648.