Annales UMCS Informatica AI XIV, 3 (2014) 21–31
DOI: 10.2478/umcsinfo-2014-0019

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI
http://www.annales.umcs.lublin.pl/

# On the capacity of survival analysis with the R language

Andrzej Krajka[1*], Paweł Krawczak[2†], Radosław Mlak[3‡]

[1] *Institute of Computer Science, Maria Curie-Sklodowska University Akademicka 9, 20-033 Lublin, Poland*
[2] *Department of Pneumology, Oncology and Allergology, Medical University of Lublin, Jaczewskiego 8, 20-090 Lublin, Poland*
[3] *Department of Human Physiology, Medical University of Lublin, Radziwiłłowska 11, 20-080 Lublin, Poland*

**Abstract** – In order to make the big data mining analysis we meet the limit of computer capacity. We concentrate here on such a situation. We describe the problem, test the key fragment of the algorithm and conclude on the possibilities of similar computations.

## 1 Introduction

In this paper we present the use of computer methods in order to investigate the influence of genes and other demographic and clinical characteristics on the clinical outcome of non-small cell lung cancer (NSCLC) patients treated with first line chemotherapy based on platinum compounds. The investigations were carried out on the sample of 201 persons observed in the Department of Pneumology, Oncology and Allergology of Public Hospital No 1 in Lublin. The investigations concerned the influence made by 13 genes and such characteristics as: Age, Sex, Smoking, Comorbidities, History of chemotherapy, Complications after chemotherapy and others on lung cancer treatment. The main difficulty lies in the possible interactions between the investigated attributes, although a number of persons was not large, but taking into account the possible interactions, the number of possible investigations was $2^{numgen} - 1$ where *numgen* is the quantity of considered attributes. Computation was done in R - programistic language, well fitted to statistical and data mining modelling.

*akrajka@gmail.com
†krapa@poczta.onet.pl
‡radoslaw.mlak@gmail.com

Although the medical remarks seem to be very interesting, we restrict our attention (except in the first section) to the computer science side of fragment of these computations.

In the next subsection we will briefly sketch the biological significance of these investigations and the applied statistical tools.

### 1.1 Biological background

Lung cancer is the most frequent malicious tumour in Poland and in the world (in 2012 around 1.8 million new diseases and over 1.6 million deaths were reported due to it). Its most prevalent histological subtype is non-small cell lung cancer (NSCLC) **[1, 2]**. In Poland for different reasons, the majority of patients have their lung cancer diagnosed in its advanced stage. As a result, the overwhelming majority of patients $(80-85\%)$ can be qualified only for chemotherapy or radiotherapy, which are, however, of moderate effectiveness. For those patients who are treated with standard chemotherapy of line 1 the median of progression free survival is $6-8$ months, the median of overall survival is $8-10$ months, and the percentage of 5-year survival is within the $10-15\%$ range **[3, 4]**. Currently the highest effectiveness of treatment of advanced inoperable lung cancer can be achieved in specially selected groups of patients who are given drugs which are molecularly targeted (e.g., Erlotinib, Crizotinib). These drugs, however, are reserved only for a very small group of patients who have specific molecular aberrations (around 10% for Caucasians). This is the reason why the overwhelming majority of patients $(80-85\%)$ get standard, double-drug chemotherapy. As is shown by research, also this group can be treated taking into account genetic predispositions (e.g., polymorphisms of genes coding proteins crucial for DNA repair) so as to choose potentially the most effective scheme of chemotherapy **[5, 6]**. Many previous studies proved that marking particular SNPs (single nucleotide polymorphisms) in genes coding DNA repair proteins may be useful for qualification of the most appropriate (and the most effective) chemotherapy scheme. Due to the fact that DNA repair is a complex and multi-stage process, and the genes which are responsible for it are characterised by "low penetration", which is a significant yet limited effect of a single change on the evaluated process (in this case, metabolism of cytostatic agents), in order to find strong relationships there is a need to carry out comprehensive research encompassing many different SNPs located in the line of genes belonging to a few DNA repair mechanisms. Finding the right combination of variants of particular genes will make it possible to qualify tumour patients for proper treatment in the most effective way. Owing to the fact that the current state of knowledge about genetic disturbances is extremely wide, the implementation of advanced bioinformatics tools in the process of searching for clinically-relevant changes seems to be absolutely essential.

### 1.2 Mathematical background

We are interested in the survival function, conventionally denoted S, which is defined as $S(t) = P[T > t]$, where $t$ is some time, $T$ is a random variable denoting the time of

death. That is, the survival function is the probability that the time of death is later than some specified time $t$. The survival function is also called the survivor function or the survivorship function in problems of biological survival, and the reliability function in mechanical survival problems. All attributes are used to categorize the set of persons. For example the people who are $0-18$ years old and have a gene on loci XPD2 written by GG (two guanine) have the survival function $S_1$, the other group $19-30$ years old and a gene of loci XPD2 written as AG (adenine and guanine) have the survival function $S_2$. When the attributes *Age* and *Gen XPD2* influenced the stage of illness of lung cancer, then the survival functions $S_1$ and $S_2$ should be "essentially" different. Because $1-S_i$, $i=1,2$, are the distribution functions thus the Kolmorov's-Smirnov or Chi-square tests can be applied but because due to discrete observations, $S_i$, $i=1,2$, are the discrete functions, thus only the Chi-square test can be used. This test has restrictions on the number of observations, it should be greater than 5. The groups which contain fewer than 5 observations are deleted at the moment of this analysis only.

## 2 Data and tools

### 2.1 Tools

In R language ([**7, 8, 9, 10**]) the tools for the survival analysis can be found in library **survival** ([**11**]) where there is the object **Surv(time, time2, event, type)** with the parameters: TIME1 - the vector with observations in months, TIME1$= c(0.25, 8, 2.5, 4, 2, 4, 7, 3.5, 0.1, ...)$ TIME2 - the vector with the type of finishing observation (1 death observed, 0 censored) TIME2$= c(1, 1, 0, 1, 1, 1, 1, 1, 0, ...)$ EVENT, TYPE - used if the time given is not exact but is e.g. in interval.

The R constructs the Kaplan-Meier estimate of the survival function, $S(t)$, corresponds to the non-parametric MLE estimate of $S(t)$. The resulting estimate is a step function that has jumps at the observed event times, $t_i, 1 \le i \le n$. In general, it is assumed that the $t_i$ are ordered: $0 < t_1 < t_2 < \ldots < t_n$. If the number of individuals with an observed event time $t_i$ is $d_i$, and the number of individuals at risk (i.e. those who have not experienced the event) at the time before $t_i$ is $Y_i$, then the Kaplan-Meier estimate of the survival function and its estimated variance is given by

$$\begin{aligned} \widehat{S}(t) &= \begin{cases} 1, & \text{if } t < t_1, \\ \prod_{t_i \le t}[1-\frac{d_i}{Y_i}], & \text{if } t_1 \le t, \end{cases} \\ Var[\widehat{S}(t)] &= (\widehat{S}(t))^2 \sum_{t_i \le t} \frac{d_i}{Y_i(Y_i-d_i)}. \end{aligned} \tag{1}$$

### 2.2 Data

We prepare the data for the MySQL database (for further applications, cf. [**12**]) but for the analysed fragment of program usually a CSV format file is taken as input. The investigation comprised 201 persons with different stages of lung cancer. They are described by almost 100 attributes which can be grouped into:

**Identifying attributes::** Name, Sex, Age, Address, Personal Identification Number, ...

**Data describing the person::** Smoking, Intensity of smoking, Other illnesses, Cancer cases in the family, ...

**Data describing the stage of illness:** Stage of illness, Decrease of weight, Anaemia, Illness complications and the results of the medical test Complete blood count, Biochemistry of blood, Serology, Smear test,...

**Data describing the type of used treatment:** Chemotherapy, Complications after chemotherapy, RTG-therapy, ...

**Gens::** The investigations of SNPs (conducted on DNA isolated from peripheral blood of NDRP patients) encompassed 9 genes, within which the following 13 polymorphic places were selected for research:

Table 1. Polymorphic places (genes)

| Gene | SNP | Reference sequence | Our code |
|---|---|---|---|
| ERCC1 | $19007C > T$ | rs11615 | G1 |
| ERCC1 | $8092C > A$ | rs3212986 | G2 |
| XPD/ERCC2 | $2251A > C$ | rs13181 | G3 |
| XPD/ERCC2 | $934G > A$ | rs1799793 | G4 |
| XPA | $-4A > G$ | rs1800975 | G5 |
| XPC | $1385C > T$ | rs2228000 | G6 |
| XPC | $2704C > A$ | rs2228001 | G7 |
| XRCC1 | $580C > T$ | rs1799782 | G8 |
| XRCC1 | $1196A > G$ | rs25487 | G9 |
| XPG/ERCC5 | $3310C > G$ | rs17655 | G10 |
| RRM1 | $-37C > A$ | rs12806698 | G11 |
| RRM1 | $-524C > T$ | rs11030918 | G12 |
| STMN1 | $-2166T > C$ | rs182455 | G13 |

**Survival analysis data::** Time of life, Fixed-right censoring where Time of life is given in months.

The genetic results were obtained by the amplifications PCR method and analysed by the **GeneMapperSoftware Version 4.0** produced by Applied Biosystems. Figure 1 presents the analysis of 6 SNPs of genes involved in DNA repair (ERCC1, XPC, XPD, XPA, XPG, XRCC1), genotypes from the left: heterozygote GA, homozygous

GG homozygous AA, homozygous GG homozygous CC homozygous AA. This is an example of a peak electropherogram, obtained by the separation of the reaction of products SNaPshot PCR (capillary electrophoresis was performed on a 3130 Genetic Analyzer, Applied Biosystems).

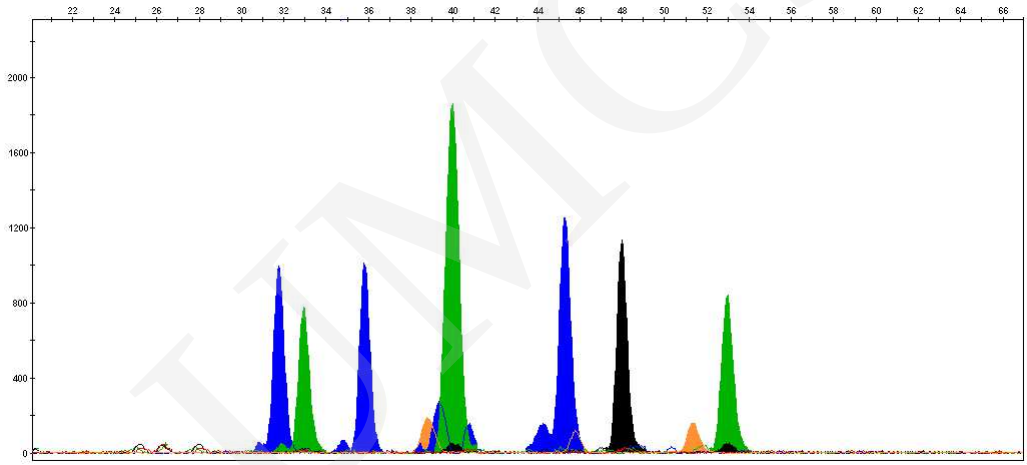


Fig. 1. A sample peak electropherogram.

As can be seen in Figure 1 a similar analysis was made for all persons and later written in databases (for simplicity coded as $0, 1, 2$ where 0 and 2 are homozygosity and 1 means heterozygosity). Such prepared data were written in the MySQL table.

## 3 Programs in R

The parameters work and the introductory was set as follows:

Listing 1. Introduction

```
chi_level=6
p_level=0.05
result="result.txt"
catalog="C:/Workspace/cancer"
datas="gen.csv"
from_column=6
numgen=16

library(survival)
setwd(catalog)
datar=read.table(datas, sep=";", header=T)
file.create(result)
cat(c("set","chi^2","degree","p-value","rho","Comb","Number"),
    "\n", file=result, append=T)
```

where `chi_level` is the minimal number of observations for which the $chi^2$ test is doing, `p_level` is the level of statistical significance, `from_column` is the number of column from which there are gene's (or may be other) attributes of persons and `numgen` is the number of genes (or may be other) attributes of investigated persons. The ASCII file for output (`result`) is prepared with the header defined in lines $13 - 14$.

### 3.1 Algorithm

The essential algorithm reads as follows:

Listing 2. Algorithm

```
for (k in 1:numgen-1)   {
  a=combn(numgen,k)
  for (kx in 1:length(a)/k)   {
    u=a[,kx]
    x=c(rep("",length(daner[,1])))
    for (e in u) x=paste(x,daner[,from_column+e], sep="")
    for (i in 1:length(x))  {
      if (length(grep("NA",x[i],fixed=TRUE))>=1) {x[i]=NA}
                              }
    sl=levels(factor(x))
    sx=summary(factor(x))
    for (i in 1:length(x))  {
      if (!is.na(x[i])) {if (sx[x[i]]<chi_level) {x[i]=NA }
                         }
    if (length(levels(factor(x)))>1)  {
```

```
SObj=Surv(daner$time_live, daner$cenzored)~x
test=survdiff(SObj, rho=0)
chival=test$chisq
degval=length(test$n)-1
cat(c(toString(u),chival,degval,1-pchisq(chival,degval),
  0,toString(sl),toString(sx)),"\n",file=result,append=T)
test=survdiff(SObj, rho=1)
chival=test$chisq
degval=length(test$n)-1
cat(c(toString(u),chival,degval,1-pchisq(chival,degval),
  1,toString(sl),toString(sx)),"\n",file=result,append=T)
                                    }
                              }
                        }
```

The key instruction $a = combn(numgen, k)$ effect is that the matrix $a$ size $\binom{numgen}{k} \times k$ has all $k$ combinations of columns from 1 to $numgen$:

$$a = combn(8,5) = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 6 \\ 1 & 2 & 3 & 4 & 7 \\ \vdots & & & & \vdots \\ 4 & 5 & 6 & 7 & 8 \end{bmatrix}^T, \tag{2}$$

where $A^T$ denotes the transposition of the matrix $A$. The instruction in row 6 paste together the choice in $a$ attributes, in rows 7-9 all pasted codes, in which there is at least one NA value change the whole value on NA, in row 10 there are fixed (in the vector `sl`) numbers of class, in row 11 there are computed quantities of each class whereas in rows 12-14 there are excluded classes with the quantities lower than `chi_level`. The next instructions (15-27) are done if only there exist at least two classes. In this case two tests of difference of survival function were done: one with $\rho = 0$ and the other with $\rho = 1$. The parameter $\rho$ belonging to the interval $[0, 1]$ allows to weigh times such that every moment of death is multiplied by $S(t)^\rho$. For $\rho = 0$ we obtain usually a log-rank test.

### 3.2 Optimization

Each fragment of code 1 was verified with respect to possible optimization. We show two such attempts. Consider the fragment of $12 - 14$ of algorithm 1:

Listing 3. Optimalisation 1

```
# Algorithm A
system.time(replicate(10000,{
  x2=x
```

```
  for (i in 1:length(x2))
    {
      if (!is.na(x[i])) {
        if (is.na(sx[x2[i]]) | (sx[x2[i]]<=chipoziom)){
          x2[i]=NA }}
                          }}))

# Algorithm B
system.time(replicate(10000, {
  x1=x
  x1[ is.na(sx[x1]) | (sx[x1]<=chipoziom)]=NA
  }))
```

gives the running times

Table 2. Times of executions of fragments A and B

| part | user | system | elapsed |
|---|---|---|---|
| A | 31.05 | 0.03 | 31.07 |
| B | 0.88 | 0.01 | 0.88. |

We see that the running times of fragment B are significantly better than those of fragment A and in consequence, fragment B replaces fragment A in algorithm 2. It is natural, due to speed operation on indices of big data frames and vector. The attempt of replacing iteration by the instruction `lapply` does not give satisfactory results:

Listing 4. Optimalisation 2

```
# C
system.time(replicate(100, {for (e in u) {
    x=paste(x,daner[,e], sep="") }}))

# D
system.time(replicate(100, x=unlist(lapply(1:length(daner[,1]),
    function(y) paste(daner[y,u], sep="", collapse=""))))
```

with the running times:

Table 3. Times of executions of fragments C and D

| part | user | system | elapsed |
|---|---|---|---|
| C | 0.28 | 0.00 | 0.28 |
| D | 7.33 | 0.00 | 7.33 |

The profiler Rprof("profiler.out", interval=0.1, memory.profiling=TRUE) applied for numgen in 17:19 gives the following (we present only the results with time higher than 300 sec):

Table 4. Profiler applied for Algorithm 2

| $by.total | total.time | total.pct | mem.total | self.time | self.pct |
|---|---|---|---|---|---|
| factor | 2387.4 | 32.20 | 102046.2 | 140.3 | 1.89 |
| paste | 1963.1 | 26.47 | 80771.7 | 1591.9 | 21.47 |
| sort.list | 1709.2 | 23.05 | 66395.4 | 1458.2 | 19.67 |
| summary | 1437.5 | 19.39 | 61096.7 | 17.0 | 0.23 |
| levels | 1283.4 | 17.31 | 54993.9 | 112.5 | 1.52 |
| grep | 1198.9 | 16.17 | 83705.0 | 1180.3 | 15.92 |
| survdiff | 876.4 | 11.82 | 54447.4 | 22.9 | 0.31 |
| cat | 558.8 | 7.54 | 32641.2 | 62.9 | 0.85 |
| [ | 528.6 | 7.13 | 30592.9 | 64.3 | 0.87 |
| summary.factor | 512.6 | 6.91 | 25373.7 | 10.1 | 0.14 |
| table | 466.9 | 6.30 | 24353.0 | 71.0 | 0.96 |
| eval | 460.2 | 6.21 | 27834.9 | 42.9 | 0.58 |
| [.data.frame | 450.7 | 6.08 | 25963.7 | 170.2 | 2.30 |
| close | 436.4 | 5.89 | 25311.1 | 3.0 | 0.04 |
| close.connection | 433.4 | 5.84 | 25172.8 | 433.4 | 5.84 |
| match.arg | 333.4 | 4.50 | 19573.3 | 39.0 | 0.53 |
| survdiff.fit | 321.0 | 4.33 | 18842.4 | 5.5 | 0.07 |

Thus the conversion factor, operations on strings paste and other operations on factors (summary, levels) are most time-consuming, but **factor type** in R is not so flexible in order to change Algorithm 2.

## 4 Results

We run algorithm 1 in three cases: for $N = 50$ where 151 randomly chosen items were deleted, for $N = 201$ - the original database of investigated persons and with $N = 1000$ where 799 persons were drawn (every attribute was drawn according to the attribute distribution). For the `numgen`$< 13$ we take the randomly chosen genes whereas for `numgen` $> 13$ we take all gene values and additionally randomly chosen attributes. The time (given in seconds) of this running on computer (processor - Intel(R) Core(TM)2 Quad CPU Q8200 2.33 GHz 2.34 GHz, RAM 4GB, 64 bit, system Debian, R version 2.15.2) is as follows:

Table 5. Times of executions of algorithm 2

| num- | N=50 | | | N=201 | | | N=1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| gen | user | system | elapsed | user | system | elapsed | user | system | elapsed |
| 8 | 1.3 | 0.2 | 1.9 | 2.8 | 0.3 | 3.8 | 10.5 | 0.3 | 12.5 |
| 9 | 2.2 | 0.1 | 3.0 | 5.3 | 0.3 | 7.0 | 22.0 | 0.7 | 26.5 |
| 10 | 4.6 | 0.3 | 5.9 | 10.4 | 0.7 | 13.7 | 44.3 | 1.5 | 52.3 |
| 11 | 7.7 | 0.3 | 9.7 | 19.1 | 1.3 | 25.0 | 95.0 | 2.9 | 112.0 |
| 12 | 17.3 | 0.7 | 21.3 | 38.4 | 2.2 | 49.4 | 203.8 | 5.6 | 236.6 |
| 13 | 30.6 | 0.7 | 35.7 | 66.5 | 2.5 | 80.2 | 399.3 | 9.5 | 454.0 |
| 14 | 61.8 | 1.4 | 68.9 | 141.6 | 5.3 | 166.6 | 743.9 | 15.8 | 832.0 |
| 15 | 130.4 | 2.0 | 141.5 | 296.1 | 8.9 | 340.3 | 1534.4 | 30.4 | 1700.4 |
| 16 | 279.3 | 3.0 | 299.0 | 593.4 | 16.7 | 668.9 | 3087.8 | 55.5 | 3375.2 |
| 17 | 524.9 | 5.8 | 558.9 | 1173.3 | 26.6 | 1302.8 | 6454.6 | 104.0 | 7009.6 |
| 18 | 1019.5 | 6.1 | 1050.2 | 2233.6 | 34.4 | 2408.9 | 12550.0 | 179.1 | 13627.6 |
| 19 | 2103.0 | 6.7 | 2136.9 | 4529.3 | 38.4 | 4711.5 | 20395.4 | 342.2 | 21276.2 |
| 20 | 4303.9 | 12.3 | 4374.6 | 9300.0 | 67.9 | 9616.9 | 41432.3 | 512.5 | 43422.1 |
| 21 | 8817.2 | 13.3 | 8896.3 | 19656.1 | 69.6 | 24626.2 | 80139.9 | 698.1 | 85838.3 |

On the basis of Table 5, using the regression methods in R, we find the following formula on time (in sec.) of execution of algorithm 2:

$$time = 0,000183837 \times \exp\{0,6823 \times numgen\} \times N^{0,80421}. \tag{3}$$

thus if we allow the computer to work for $10min, 1h, 24h$ or 1 week for the given quantity of persons $N$ we may take the maximum number of attributes as in Table 5

Table 6. Maximum number of attributes for $N$ persons

| | Time | | | |
|---|---|---|---|---|
| N | 10 min | 1h | 24h | 1week |
| 40 | 17 | 20 | 24 | 27 |
| 100 | 16 | 19 | 23 | 26 |
| 200 | 15 | 18 | 23 | 25 |
| 500 | 14 | 17 | 21 | 24 |
| 1000 | 13 | 16 | 21 | 23 |

Neglecting very interesting medical conclusions we remark that

- Language R is the fastest among the data mining and statistical tools (compared with STATISTICA, SPSS and others),
- From the mathematical viewpoint it would be better to investigate all attributes of persons but informatics restrictions described in Tables 5 and 6 make this often impossible
- Everyone planning similar computations should be aware of the above mentioned restrictions.

## References

[1] Dela Cruz C.S., Tanoue L.T., Matthay R.A., Lung cancer: epidemiology, etiology, and prevention, Clin Chest Med. 32 (2011): 605.

[2] NSCLC Meta-Analysis Collaborative Group. Chemotherapy in addition to supportive care improves survival in advanced non-small-cell lung cancer. A systemic review and meta-analysis of individual patient data from 16 randomized controlled trials, J. Clin Oncol. 26 (2008): 4617.

[3] Simon G.R., Ismail-Khan R., Bepler G., Nuclear excision repair-based personalized therapy for non-small cell lung cancer: from hypothesis to reality, Int. J. Biochem. Cell. Biol. 39 (2007): 1318.

[4] Jordheim L.P., Sève P, Trédan O., et al., The ribonucleotide reductase large subunit (RRM1) as a predictive factor in patients with cancer, Lancet Oncol. 12 (2011): 693.

[5] Stewart D.J., Tumor and host factors that may limit efficacy of chemotherapy in non-small cell and small cell lung cancer, Crit. Rev. Oncol. Hematol, 75 (2010): 173.

[6] Vilmar A., Sorensen J.B., Excision repair cross-complementation group 1 (ERCC1) in platinum-based treatment of non-small cell lung cancer with special emphasis on carboplatin: a review of current literature, Lung Cancer. 64 (2009): 131.

[7] Biecek P., Przewodnik po pakiecie R, Oficyna Wydaw. GIS, Wrocław (2008).

[8] The home page of language R; http://cran.r-project.org/

[9] Statistics in R; http://zoonek2.free.fr/UNIX/48_R/all.html

[10] Walesiak M., Gatnar E., Statystyczna analiza danych z wykorzystaniem programu R, Wydaw. Nauk. PWN SA., Warszawa (2009).

[11] The documentation of library `survival`;
http://cran.r-project.org/web/packages/survival/ survival.pdf

[12] The RMySQL manual; http://cran.r-project.org/web/packages/RMySQL/RMySQL.pdf