Analysis of Death Causes in the STULONG Data Set

Jan Burian, Jan Rauch

EuroMISE – Cardio, University of Economics Prague,
W. Churchilla 4, 130 67 Prague, Czech Republic
burian@vse.cz, rauch@vse.cz

Abstract. The associations between attributes of the Entry data matrix and death causes are analysed. Attributes of the Entry data matrix are divided into three groups. Association rules of four types are mined for each group of attributes and for particular causes. Several potentially interesting association rules relations were found.

1. Introduction

The aim of this analysis is to look for associations between medical characteristics of patients and death causes. We are looking for associations between attributes of the *Entry* and *Death* data matrices from the STULONG data set. We deal only with the attribute *Cause of death* of the data matrix *Death* see Tab. 1. For more detailed information about STULONG project and description of attributes of data matrix *Entry* look at http://euromise.vse.cz/challenge2003/data/.

Values of attribute *Cause of death*			
Code	**Cause of death**	**No. of patients**	**%**
05	myocardial infarction	80	20.6
06	coronary heart disease	33	8.5
07	stroke	30	7.7
08	other causes	79	20.3
09	sudden death	23	5.9
10	cause of the death unknown	8	2.0
16	tumorous disease	114	29.3
17	general atherosclerosis	22	5.7
Total		**389**	**100.0**

Tab. 1. – Attribute *Cause of death*

An overview of analysed attributes of the data matrix *Entry* is in section 2. We search for association in the form of association rules of various types that are introduced in section 3. We solve several analytic tasks related to three groups of attributes, see section 4. The overview of results is in section 5 and some concluding remarks are in section 6.

2. Attributes of *Entry* Data Matrix

We used many attributes of the data matrix *Entry*. We divided them into three groups – General characteristics, Investigations and Vices, see Table 2:

General characteristics	Examinations	Vices
Activity after work	Chest pain	Alcohol
Traffic	Breathlesness	Former smoker
Marital status	Cholesterol	Tea
Physical activity in work	Urine	Sugar
Education	Subscapular	Time of smoking
Responsibility	Triceps	Coffe
Age		Smoking
Weight		Liquors
Height		Liquors – how much
		Beer 10
		Beer 12
		Beer – how much
		Wine
		Wine – how much

Tab. 2. – Analysed attributes of data matrix *Entry*

3. Association rules

We analyse data matrix *ED* that is result of joining of data matrices *Entry* and *Death*, see figure 1. Each row of data matrix *ED* corresponds to one patient and each column corresponds to one attribute.

Patient	General characteristics		Examinations		Vices		Cause of death
	Activity after work	…	Chest pain	…	Alcohol	…	
1	moderate activity		not present		no		stroke
2	great activity		non-ischaemic		occasionally		myocardial infarction
...	...	...	...	...	...	...	...
389	he mainly sits		other pains		regularly		tumorous disease

Fig. 1. – Data matrix *ED*

We use the procedure 4ftMiner that is a part of the academic system LISp-Miner (for details see http://lispminer.vse.cz). The procedure 4ft-Miner mines for association rules of the form

antecedent ≈ succedent.

Both antecedent and succedent are Boolean attributes automatically derived from columns of the analysed data matrix *ED*. The symbol ≈ is called 4ft quantifier. It corresponds to a condition concerning four-fold contingency table of antecedent and succedent in analysed data matrix se table 3.

ED	succedent	¬ succedent
antecedent	a	b
¬ antecedent	c	d

Tab. 3. – Four fold contingency table of antecedent and succedent in *M*

Here a is the number of rows of the analysed data matrix *ED* satisfying both antecedent and succedent, b is the number of rows satisfying antecedent and not satisfying succedent, c is the number of rows not satisfying antecedent and satisfying succedent, and d is the number of rows satisfying neither antecedent nor succedent.

We use the following 4ft quantifiers with parameters $0 < p \leq 1$ and $Base > 0$:

Founded implication $\Rightarrow_{p;Base}$ that is associated to the condition

$$\frac{a}{a+b} \geq p \wedge a \geq Base$$

The association rule antecedent $\Rightarrow_{p;Base}$ succedent can be interpreted as "*100p per cent of patients satisfying* antecedent *satisfy also* succedent" or "antecedent *implies* succedent *on the level 100p per cent*".

Above average $\Rightarrow^{+}_{p;Base}$ that is associated to the condition

$$\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq Base$$

The association rule antecedent $\Rightarrow^{+}_{p;Base}$ succedent can be interpreted as "*among patients satisfying* antecedent *there is the relative frequency of objects satisfying* succedent *at least 100p per cent higher than the relative frequency of objects satisfying* succedent *among all the objects in the whole data matrix*".

Double founded implication $\Leftrightarrow_{p;Base}$ that is associated to the condition

$$\frac{a}{a+b+c} \geq p \wedge a \geq Base$$

The association rule antecedent $\Leftrightarrow_{p;Base}$ succedent can be interpreted as "*100*p per cent of patients satisfying* antecedent *or* succedent *satisfy both* antecedent *and* succedent" or "antecedent and succedent" *implies* "antecedent or succedent" *on the level 100*p per cent*".

Founded equivalency $\equiv_{p;Base}$ that is associated to the condition

$$\frac{a}{a+b+c} \geq p \wedge a \geq Base$$

The association rule antecedent $\equiv_{p;Base}$ succedent can be interpreted as "1*00*p per cent of patients have the same value (i.e. "true" or" false") both for* antecedent *and* succedent".

4. Tasks

An example of association rule concerning data matrix *ED* is the expression [1]:

Education(university) & Height<176-180> $\Rightarrow_{1.1,16}$ Death cause (tumour) [1]

Here the antecedent Education(university) & Height<176-180> is Boolean attribute that is true for the patient (i.e. row in the data matrix *ED*) if both the value of the attribute Education is university and the value of the attribute Height is in the interval <176-180> of cm.

Analogously the succedent Death cause (tumorous disease) is the Boolean attribute that is true if the value of the attribute Death cause is tumorous disease.

The four-fold table of antecedent Education(university) & Height<176-180> and succedent Death cause (tumorous disease) in data matrix *ED* is in table 4.

ED	Death cause (tumorous disease)	¬ Death cause (tumorous disease)
Education(university) & Height<176-180>	16 (=a)	10 (=b)
¬ (Education(university) & Height<176-180>)	98 (=c)	265 (=d)

Tab. 4. – Four fold contingency table of association rule [1] in data matrix *ED*

The association rule [1] is true because of both the relative frequency 0.615 (i.e. 16/(16 + 10)) of patients satisfying succedent Death cause (stroke) among patients satisfying antecedent Education(university) & Height<176-180> is 110 (i.e. 1.1*100) per cent higher than the relative frequency 0.293 (i.e. (16 + 98) / 389) of patients satisfying Death cause (stroke) among all the observed patients and $a \geq 16$ *(i.e.* $a \geq Base$*).*

The antecedent Education(university) & Height<176-180> is the conjunction of basic Boolean attributes Education(university) and Height<176-180>. Generally we deal with basic Boolean attributes of the form Attribute(value) where Attribute correspond to the column of the analysed data matrix and value is one of possible values or interval of possible values of Attribute.

We search for true association rules of the form

$$\text{antecedent} \approx \text{Death cause (?)}$$

where antecedent is basic Boolean attribute Attribute(value) or conjunction

$$\text{Attribute}_1(\text{value}_1) \,\&\, \text{Attribute}_2(\text{value}_2)$$

of basic Boolean attributes $\text{Attribute}_1(\text{value}_1)$ and $\text{Attribute}_2(\text{value}_2)$ and where Death cause (?) is basic Boolean attribute with arbitrary value except „other cause“, „sudden death“ and „cause of the death unknown“. Further ≈ is one of 4ft quantifiers $\Rightarrow_{p;Base}$, $\Rightarrow^{+}_{p;Base}$, $\Leftrightarrow_{p;Base}$ $\equiv_{p;Base}$ see section 3.

We solve four particular tasks:

- **General characteristics** $\approx$ **Death cause (?)** – in this task we search for rules antecedent $\approx$ Death cause (?) where antecedent is automatically generated from the group of attributes "General characteristics", see Tab. 2
- **Examinations** $\approx$ **Death cause (?)** – in this task we search for rules antecedent $\approx$ Death cause (?) where antecedent is automatically generated from the group of attributes "Examinations", see Tab. 2
- **Vices** $\approx$ **Death cause (?)** – in this task we search for rules antecedent $\approx$ Death cause (?) where antecedent is automatically generated from the group of attributes "Vices", see Tab. 2
- **Combinations** $\approx$ **Death cause (?)** – in this task we search for rules antecedent $\approx$ Death cause (?) where antecedent is automatically generated conjunction of **two or three** basic Boolean attributes such that they are not from the same group of attributes, see Tab. 2

For every 4ft quantifier we start the task with very weak parameters (p = 0.5, Base = 15 for 4ft quantifiers $\Rightarrow_{p;Base}$ and $\Rightarrow^{+}_{p;Base}$ and p = 0.8, Base = 20 for 4ft quantifiers $\Leftrightarrow_{p;Base}$ and $\equiv_{p;Base}$). When we get zero or only few true rules we stop the analysis. When we get too many rules we use more strong parameters to get a reasonable number of association rules.

5. Overview of Results

We have not found any interesting results for the quantifier $\Leftrightarrow_{p;Base}$ of double founded implication and for the quantifier $\equiv_{p;Base}$ of founded equivalence. It means that there are not interesting combinations of attributes of the Entry data matrix equivalent to particular death causes. This fact is not surprising in our opinion. However there are several potentially interesting association rules with quantifier $\Rightarrow_{p;Base}$ of founded implication and with quantifier $\Rightarrow^{+}_{p;Base}$ of above average relation.

5.1 General characteristics

There are three rules for the **founded implication quantifier** $\Rightarrow_{p;Base}$ with parameters **p = 0.5** and **Base = 15**:

1) Education(university) & Height<176-180>
$\Rightarrow_{0.62;16}$ Death cause (tumouros disease)

It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.

2) Physical activity in work(he mainly sits) & Height<176-180>

$\Rightarrow_{0.52;24}$ Death cause (tumouros disease)

It means that on tumorous disease have died 24 i.e. 52% of patients that mainly sit in the work and whose height is 176-180 cm.

3) Education(university) & Responsibility(managerial worker)

$\Rightarrow_{0.50;15}$ Death cause (tumouros disease)

It means that on tumorous disease have died 15 i.e. 50% of patients with university education and with managerial responsibility.

There are three rules for the ***above average*** **quantifier** $\Rightarrow^{+}_{p;Base}$ with parameters **p = 0.75** and **Base = 15**:

1) Education(university) & Height<176-180>

$\Rightarrow^{+}_{1.1;16}$ Death cause (tumouros disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on tumorous disease among patients with university education and with height 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients
- 16 of patients with university education and with height 176-180 cm died on tumorous disease

2) Physical activity in work(he mainly sits) & Height<176-180> ¨

$\Rightarrow^{+}_{0.78;24}$ Death cause (tumouros disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on tumorous disease among patients who mainly sits in work and whose height is 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients
- 24 of patients who mainly sits in work and whose height is 176-180 cm died on tumorous disease.

3) Age(≥65) $\Rightarrow^{+}_{0.76;15}$ Death cause (general atherosclerosis)

It means that the two following assertions are true:

- the relative frequency of patients who died on general atherosclerosis among patients with age 65 or older is 76 per cent higher than the relative frequency of patients who died on general atherosclerosis among the 389 observed patients
- 15 of patients with age 65 or older died on general atherosclerosis.

5.2 Examinations

There is one rule for the **founded implication quantifier** $\Rightarrow_{p;Base}$ with parameters **p = 0.5** and **Base = 15**

Cholesterol<244;265> & Subscapular<14;21)
$\Rightarrow_{0.54;15}$ Death cause (tumouros disease)

It means that on tumorous disease have died 15, i.e. 54% of patients with cholesterol in interval <244;265> mg% and with the subscapular skinfold in the interval <14;21) mm.

There are four rules for the ***above average*** **quantifier** $\Rightarrow^{+}_{p;Base}$ with parameters **p = 0.50** and **Base = 15**

1) Cholesterol<244;265> & Subscapular<14;21)
$\Rightarrow^{+}_{0.83;15}$ Death cause (tumouros disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on tumorous disease among patients with cholesterol in the interval <244;265> of mg% and with the subscapular skinfold in the interval <14;21) mm is 83 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients
- 15 of patients of patients with cholesterol in the interval <244;265> mg% and with the subscapular skinfold in the interval <14;21) mm died on tumorous disease

2) Chest pain(not present) & Subscapular<14;21)
$\Rightarrow^{+}_{0.65;15}$ Death cause (coronary heart disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on coronary heart disease among patients with not present chest pain and with the subscapular skinfold in the interval <14;21) mm is 65 per cent higher than the relative frequency of patients who died on coronary heart disease among all the 389 observed patients,
- 15 of patients of patients with not present chest pain and with the subscapular skinfold in the interval <14;21) mm died on coronary heart disease.

3) Subscapular<14;21) $\Rightarrow^{+}_{0.53;18}$ Death cause (coronary heart disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on coronary heart disease among patients with the subscapular skinfold in the interval <14;21) mm is 53 per cent higher than the relative frequency of patients who died on coronary heart disease among all the 389 observed patients,

- 18 of patients with the subscapular skinfold in the interval <14;21) mm died on coronary heart disease.

4) Chest pain(not present) & Subscapular<7;14)
$\Rightarrow^{+}_{0.52;19}$ Death cause (myocardial infarction)

It means that the two following assertions are true:
- the relative frequency of patients who died on myocardial infarction among patients with not present chest pain and with the subscapular skinfold in the interval <14;21) mm is 52 per cent higher than the relative frequency of patients who died on myocardial infarction among all the 389 observed patients,
- 19 of patients with not present chest pain and with the subscapular skinfold in the interval <14;21) mm died on myocardial infarction.

5.3 Vices

There are ten rules for the ***above average*** **quantifier** $\Rightarrow^{+}_{p;Base}$ with parameters **p = 0.50** and **Base = 15:**

1) Beer 12(yes) & Wine(yes) $\Rightarrow^{+}_{0.66;17}$ Death cause (tumouros disease)

It means that the two following assertions are true:
- the relative frequency of patients who died on tumouros disease among patients who drink both beer 12° and wine is 66 per cent higher than the relative frequency of patients who died on tumouros disease among all the 389 observed patients,
- 17 of patients of patients who drink both beer 12 and wine died on tumouros disease.

2) Beer 12(yes) & Wine – amount per day(up to half a litre)
$\Rightarrow^{+}_{0.65;16}$ Death cause (tumouros disease)

It means that the two following assertions are true:
- the relative frequency of patients who died on tumouros disease among patients who drink both beer 12° and up to half a litre of wine per day is 65 per cent higher than the relative frequency of patients who died on tumouros disease among all the 389 observed patients,
- 16 of patients who drink both beer 12° and up to half a litre of wine per day died on tumouros disease.

3) Coffee(3 and more cups) & Beer - how much (up to 1 litre per day)
$\Rightarrow^{+}_{0.63;20}$ Death cause (tumouros disease)

It means that the two following assertions are true:
- the relative frequency of patients who died on tumouros disease among patients who drink both 3 and more cups of coffee and up to 1 litre of bier per

day is 63 per cent higher than the relative frequency of patients who died on tumouros disease among all the 389 observed patients,
- 20 of patients who drink both 3 and more cups of coffee and up to 1 litre of bier per day died on tumouros disease.

4) Smoking(non-smoker) & Alcohol(occasionally)

$$\Rightarrow^{+}_{0.6;15} \text{Death cause (tumouros disease)}$$

It means that the two following assertions are true:
- the relative frequency of patients who died on tumouros disease among non-smokers who drink occasionally alcohol is 60 per cent higher than the relative frequency of patients who died on tumouros disease among all the 389 observed patients,
- 15 of patients - nonsmokers who drink occasionally alcohol died on tumouros disease.

5) Alcohol(occasionally) & Coffee(3 and more cups)

$$\Rightarrow^{+}_{0.55;20} \text{Death cause (tumouros disease)}$$

It means that the two following assertions are true:
- the relative frequency of patients who died on tumouros disease among patients who drink occasionally alcohol and drink 3 and more cups of coffee per day is 55per cent higher than the relative frequency of patients who died on tumouros disease among all the 389 observed patients,
- 20 of patiens who drink occasionally alcohol and drink 3 and more cups of coffee per day died on tumouros disease.

6) Coffee(3 and more cups) & Beer 10(yes)

$$\Rightarrow^{+}_{0.53;17} \text{Death cause (tumouros disease)}$$

It means that the two following assertions are true:
- the relative frequency of patients who died on tumouros disease among patients who drink both 3 and more cups of coffee and beer 10° is 53 per cent higher than the relative frequency of patients who died on tumouros disease among all the 389 observed patients,
- 17 of patiens who drink both 3 and more cups of coffee and beer 10° died on tumouros disease.

7) Tea(he does not drink) & Smoking(21 and more cig./day)

$$\Rightarrow^{+}_{0.59;17} \text{Death cause (myocardial infarction)}$$

It means that the two following assertions are true:
- the relative frequency of patients who died on myocardial infarction among patients who does not drink tea and smoke 21 and more cigarettes per day is 59 per cent higher than the relative frequency of patients who died on myocardial infarction among all the 389 observed patients,
- 17 of patients who does not drink tea and smoke 21 and more cigarettes per day died on myocardial infarction.

8) Tea(he does not drink) & Wine – amount per day(up to half a litre)

$$\Rightarrow^{+}_{0.53;23}$$ Death cause (myocardial infarction)

It means that the two following assertions are true:

- the relative frequency of patients who died on myocardial infarction among patients who does not drink tea and drink up to half a litre of wine per day is 53 per cent higher than the relative frequency of patients who died on myocardial infarction among all the 389 observed patients,
- 23 of patients who does not drink tea and drink up to half a litre of wine per day died on myocardial infarction.

9) Tea(he does not drink) & Wine(yes)

$$\Rightarrow^{+}_{0.52;24}$$ Death cause (myocardial infarction)

It means that the two following assertions are true:

- the relative frequency of patients who died on myocardial infarction among patients who does not drink tea and drink wine is 52 per cent higher than the relative frequency of patients who died on myocardial infarction among all the 389 observed patients,
- 24 of patients who does not drink tea and drink wine died on myocardial infarction.

10) Smoking(21 and more cig./day) &

& Wine – amount per day(up to half a litre)

$$\Rightarrow^{+}_{0.52;19}$$ Death cause (myocardial infarction)

It means that the two following assertions are true:

- the relative frequency of patients who died on myocardial infarction among patients who smoke 21 and more cigarettes per day and drink up to half a litre of wine per day is 52 per cent higher than the relative frequency of patients who died on myocardial infarction among all the 389 observed patients,
- 19 of patients who smoke 21 and more cigarettes per day and drink up to half a litre of wine per day died on myocardial infarction.

5.3 Combinations

There are five rules for the **founded implication quantifier** $\Rightarrow_{p;Base}$ with parameters **p = 0.5** and **Base = 15:**

1) Cholesterol<250;273> & Coffee(3 and more cups)

$$\Rightarrow_{0.63;15}$$ Death cause (tumorous disease)

It means that on tumorous disease have died 15, i.e. 63% of patients with cholesterol in interval <25;265> mg% who drink 3 and more cups of coffee per day.

2) Subscapular<18;21> & Wine - how much(up to half a litre)

$\Rightarrow_{0.56;15}$ Death cause (tumorous disease)

It means that on tumorous disease have died 15, i.e. 56% of patients with subscapular skinfold in the interval <18;21> mm who drink up to half a litre of wine per day.

3) Subscapular<16;19> & Wine - how much(up to half a litre)

$\Rightarrow_{0.53;18}$ Death cause (tumorous disease)

It means that on tumorous disease have died 15, i.e. 56% of patients with subscapular skinfold in the interval <16;19> mm who drink up to half a litre of wine per day.

4) Activity after work(moderate activity) & Subscapular<16;19> &
& Wine - how much(up to half a litre)

$\Rightarrow_{0.52;15}$ Death cause (tumorous disease)

It means that on tumorous disease have died 15, i.e. 52% of patients with moderate activity after work and with subscapular skinfold in the interval <16;19> mm who drink up to half a litre of wine per day.

5) Activity after work(moderate activity) & Subscapular<16;19> &
& Tea(he does not drink)

$\Rightarrow_{0.50;16}$ Death cause (tumorous disease)

It means that on tumorous disease have died 16, i.e. 50% of patients with moderate activity after work and with subscapular skinfold in the interval <16;19> mm who does not drink tea.

There are three rules for the ***above average*** **quantifier** $\Rightarrow^{+}_{p;Base}$ with parameters **p = 0.8** and **Base = 15:**

1) Cholesterol<250;273> & Coffee(3 and more cups)

$\Rightarrow^{+}_{1.13;15}$ Death cause (tumorous disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on tumorous disease among patients with cholesterol in the interval <250;273> mg% who drink 3 and more cups of coffee per day is 113 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients
- 15 of patients with cholesterol in the interval <250;273> of mg% who drink 3 and more cups of coffee per day died on tumorous disease

2) Subscapular<18;21> & Wine - how much (up to half a litre)

$\Rightarrow^{+}_{0.9;15}$ Death cause (tumorous disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on tumorous disease among patients with subscapular skinfold in the interval <18;21> mm who drink up to half a litre of wine per day is 90 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients
- 15 of patients with subscapular skinfold in the interval <18;21> mm who drink up to half a litre of wine per day died on tumorous disease.

3) Subscapular<16;19> & Wine - how much (up to half a litre)
$\Rightarrow^{+}_{0.81;18}$ Death cause (tumorous disease)

It means that the two following assertions are true:

- the relative frequency of patients who died on tumorous disease among patients with subscapular skinfold in the interval <16;19> mm who drink up to half a litre of wine per day is 81 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients
- 15 of patients with subscapular skinfold in the interval <16;19> mm who drink up to half a litre of wine per day died on tumorous disease

6. Concluding remarks

There have been found potentially interesting association rules concerning attributes of the *E*ntry data matrix and the following causes of death:

- myocardial infarction,
- tumorous disease,
- coronary heart disease
- general atherosclerosis.

Further analysis concerning *Control* data matrix and death causes is supposed to be done after consultations with the medical expert.

References

1. The homepage of PKDD Discovery Challenge 2003 http://euromise.vse.cz/challenge2003/
2. Rauch, J. – Šimůnek, M.: Alternative Approach to Mining Association Rules (in FDM 2002, The Foundation of Data Mining and Knowledge Discovery, The Proceedings of the Workshop of ICDM02, ISBN: 4-947717-02-6, pages 157-162), Japan, December 2002
3. The homepage of the LISp-Miner system http://lispminer.vse.cz/

This paper has been supported by project LN00B107 of the Mininstry of Education of the Czech Republic