CMST 16(1) 5-11 (2010)

DOI:10.12921/cmst.2010.16.01.05-11

COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY **16(1)**, 5-11 (2010)

An Exploratory Clustering Study of Rare Adverse Events in Drug Eluting Stents Patients

B. Bychowiec[1], J. Piskorski[2], K. Stanislawska[1], M. Dziarmaga[1], A. Minczykowski[1]
A. Wykretowicz[1], H. Wysocki[1]

[1]*Department of Cardiology – Intensive Therapy and Internal Diseases, Poznań University of Medical Sciences*
ul. Przybyszewskiego 49, Poznań, Poland
e-mail: bartek.bychowiec@gmail.com, kasia.stanislawska@onet.eu,
mdziarmaga@poczta.onet.pl, aminczykowski@gmail.com, awykreto@ptkardio.pl, hwysocki@plusnet.pl

[2]*Institute of Physics, University of Zielona Góra, ul. Szafrana 4a, Zielona Góra, Poland*
e-mail: jaropis@zg.home.pl

(Received: 17 January; revised: 16 April 2010; accepted: 23 April; published online: 30 June 2010)

Abstract: We present an exploratory study of a group of patients (150) who have undergone precutaneous transluminal coronary angioplasty with the use of drug eluting stents. We concentrate on the most often studied rare adverse event, i.e. death, as well as the still unexplored cancer variable. The aim of the study is to identify possible statistical hypotheses for a subsequent, large sample study. The results of this study may lead to a change in the therapy administered after precutaneous coronary interventions which will reduce the mortality rate. To achieve this, we use clustering techniques such as hierarchical cluster analysis, principal components methods and interactive brushing. We show that death cases cluster in the space defined by the available variables, while the cancer cases do not seem to cluster.

Key words: data mining, medical visualization, PTCA, computationally intensive statistical analysis

I. INTRODUCTION

Precutaneous transluminal coronary angioplasty (PTCA) is a minimally invasive procedure performed to mechanically widen narrowed or obstructed vessels. These conditions are a result of building up of cholesterol-laden plaques due to atherosclerosis. The vessels are widened by passing non-inflated balloons to the narrowed locations and pumping them with water solution of contrast medium under pressure which is 70-500 times larger than that found in human vessels under normal conditions (6 to 20 atmospheres). This procedure is called predilation for stenting. After that in the majority of cases the narrowed location is fortified with a metal tube called a *stent* – a folded stent is introduced into the narrowed spot and then it is expanded to its intended size by the expanding balloon [1].

Precutaneous intervention disrupts the endothelium and atheromatous plaque, so it obviously causes vessel trauma. This trauma leads to localized thrombosis and consequently impairs blood flow, precipitates vessel occlusion and/or causes distal embolisation. All this makes the drug inhibition of thrombous formation absolutely necessary, with the complicating factors of the risk of both systemic bleeding and bleeding at the access site. The antiplatelet drugs are administered orally [1-3].

Two main coronary types of stent which are now in use are the *bare metal stent* (or BMS) and the *drug eluting stent* (DES). BMS stents were the first ones licensed for use in cardiac vessels – they are constructed as mesh-like tubes of wire and they have no coating. The DES are newer and differ from BMS by having a coating that slowly releases a cell-proliferation blocking drug to prevent fibrosis and clots which could lead to restenosis. The present paper concentrates on DES only [1, 3].

The complications after the implantation of DES are very rare – for example, in 24-month observation of a large group the death rate in DES patients was 4.3%, myocardial infarction rate 5.7% and target vessel revascularisation 7.4% [4]. Furthermore, in informal settings some other adverse events, such as cancer, are mentioned.

To study rare events large groups of uniform subjects are necessary. Indeed, rare events imply small rates (proportions) which entail (under the binomial distribution assumption, which is reasonable for uniform samples) asymmetric distributions with one side (the one reaching toward the less extreme values) extending far beyond the proportion, resulting in a wide, asymmetric confidence interval (CI). The results are usually compared with another group with similarly rare rates, and it is difficult to obtain statistical significance [5].

Indeed, if in a group of 50 we get 2 cases, the proportion is 4%, but the Wald-type 95% CI in (0.1-15%). In other words, the different proportions which are possible to observed is virtually the rate of 0% and proportions above 15%.

This observation leads to study design dilemmas – the collection of a reasonable sample is expensive, so it should be used to study as many phenomena and relations as possible. On the other hand, the calculation of many parameters and conducting many statistical analyses weakens the power of the reached conclusion. To overcome this problem a training sample can be used for exploratory analysis and hypothesis formation. The techniques of data mining proposed by Tukey [6] are designed, among others, for this propose. In exploratory data analysis no or hardly any statistical hypotheses are tested – its main purpose is to see what the structure of the data is and to form hypotheses which will be tested in future, larger-sample study.

The purpose of this paper is as follows. We have a sample of 150 patients who have undergone PTCA. In this sample some rare adverse events (death and cancer) may be observed, but the rates of these events are too small and the CIs too wide for us to form and test statistical hypotheses. Therefore, in what follows we will use clustering exploratory techniques to see whether the above-mentioned events are related to the many clinical and other variables collected in the course of the study. Basing on the structure of the data we will form hypotheses, but we will not test them in this group.

The techniques used in this paper are hierarchical cluster analysis, principal component analysis and principal component brushing. For the analyses the R [7] statistical system and programming language together with the `ggobi` interactive environment and `rggobi` package [16], have been used.

II. GROUP DESCRIPTION

The studied group consisted of 150 patients who underwent the PTCA procedure in the Department of Cardiology – Intensive Therapy and Internal Diseases at Poznań University of Medical Sciences, for various reasons. 36 of the subjects were women, the average age was (60.33 ± 8.72) (mean ± standard deviation). The group was followed for the median of 11.5 months with IQR (interquartile range) of 13 months. 28 of the patients suffered from diabetes, 76 from hypertension, 143 from hyperlipidemia and 66 suffered a myocardial infarction before they entered the study.

There were 5 deaths and 6 cases of cancer in the group.

Other recorded variables included: the presence of antiplatelet therapy, TIMI before and after the inclusion procedure (PTCA), types of medication, types of stents used (PES or SES), blood pressure, indication for the inclusion procedure, kidney failure, the number of dilated arteries and the number of stents used.

As can be seen, if the rare events of interest were divided into groups and subgroups according to the values of the other variables, then in most cases the groups would have no or extremely few adverse events, which would entail a total impossibility to draw any statistical conclusions. Furthermore, even if the group were much more numerous, the testing of all interesting hypotheses would weaken the conclusions considerably.

As described in the introduction, to deal with this problem we will perform exploratory analysis of the available data and form hypotheses for a future study without testing any statistical hypotheses as we move along.

III. EXPLORATORY ANALYSIS

In this part of the paper we will perform the exploratory data mining on the studied group. We expect that the variables describing the group are strongly related. Therefore, to see if the rare adverse cases cluster in a meaningful pattern, we will perform the hierarchical cluster analysis, and to see what the potential candidate variables as explanatory variables for the adverse events are, we will use Principal Components analysis and brushing.

The set of variables used in the analyses is the following: age, sex, the admission mode (emergency/planned), previous myocardial infarction, hypertension, hyperlipidemia, diabetes, TIMI before the procedure, heart failure and blood pressure. As can be seen, some of the variables are categorical/binomial. In an analysis whose aim is hypothesis testing, the various kinds of distributions would have to be addressed while selecting critical and full regions, for example for test construction. However, in the present paper our only aim is to establish proximity and clustering of variables.

III.1. Hierarchical cluster analysis

The hierarchal cluster analysis is a multivariate technique for discovering groups (clusters) which are homogeneous and separated from other. Points are assigned distances according to some metrics, and the main aim is to obtain a hierarchal representation of the points with the use of this metrics, as a two-dimensional diagram. These methods have mainly an exploratory use, as the way of calculating the metrics is quite arbitrary – the findings obtained with this method need to be confirmed with other methods.

We have constructed a dendrogram using the set of variables given above. Using the `hclust` algorithm from the `R` statistical package with the single linkage we have obtained the following dendrogram (Fig. 1).

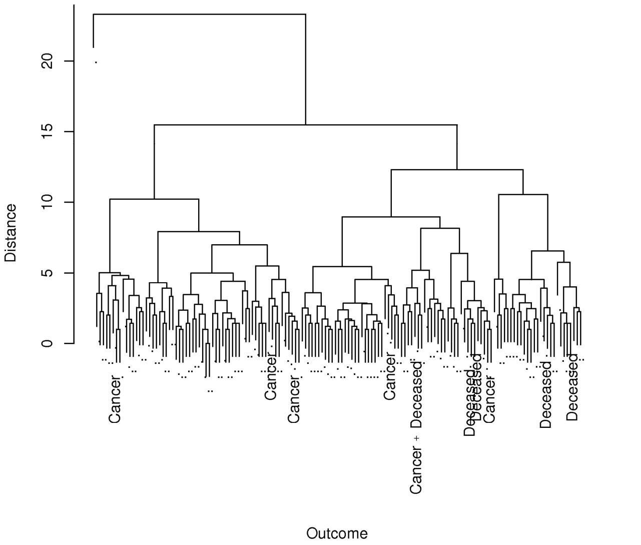

Fig. 1. The single linkage cluster dendrogram representing relations in the studied dataset. Some clustering of the death cases is visible, while no such clustering is evident for the cancer cases. There is a case of death as a result of cancer clearly marked in the label

If other algorithms are used in this data, a similar clustering occurs – we have established it for average linkage, median linkage and complete linkage. It is easy to notice that the death cases cluster, except for one case that is the result of cancer. It actually belongs to both groups. We can conclude that the variables describing the studied group are correlated in such a way that the deaths have some combinations in common, but there is no relation in cancer cases reflected in the used set of data. Perhaps the cancer cases do form a cluster with respect to other variables, not represented here (like radiation exposition time), but this information is unavailable.

III.2. Principal components analysis

The main aim of the Principal Component Analysis (PCA) is to describe the variation of correlated data in a set of transformed, uncorrelated variables. Linear combination of existing variables are created so as to obtain a set of eigenvalues of the correlation matrix or the covariance matrix. These, when selected as basis vectors, provide a set of uncorrelated basis in which each point (e.g. a subject) may be represented by a set of projections on the eigenvector directions rather than the values of the original, correlated variables [9].

The hope when performing PCA is that a few eigenvectors will account for a substantial portion of the variability in the dataset. By analyzing the scree plot and the transformation matrix it is possible to hypothesize what combinations of the original variables account for the dataset variability – sometimes it is also possible to assign an interpretation to a set of variables which has the biggest influence on a basis vector [9].

PCA results can be used to explain variability (as explained above), forming summarizing indices (e.g. finding a combination of variables which has a clinical predictive value) or exploratory analysis. In the last case the PCA is an aim in itself [10] – the most immediate result of the procedure, that is the coordinates of the datapoints in the transformed orthogonal coordinate system is interpreted. It is to be shown below.

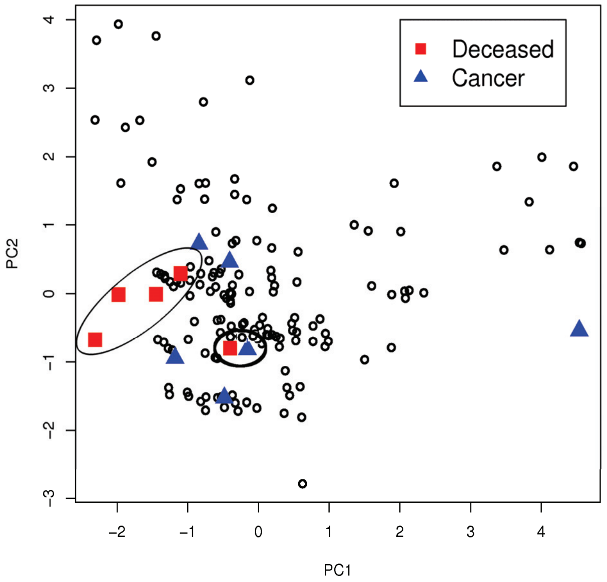

Fig. 2. PCA – the second component versus the first component. A clustering of the deceased cases is observed – one case of death as a result of cancer is marked with a bold elipse

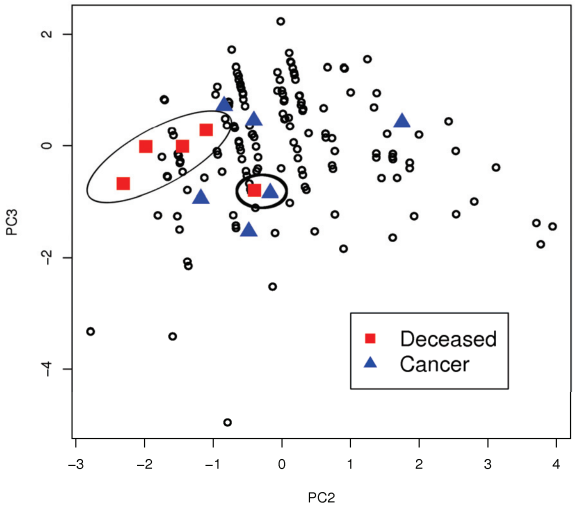

Fig. 3. PCA – the third component versus the second component. A clustering of the deceased cases is observed – one case of death as a result of cancer is marked with a bold elipse

We have performed PCA using the variables enumerated above. It has been found that the death cases clearly cluster whereas the cancer cases do not. This may be observed in Figures 2 and 3. By clustering we mean here a topographical proximity of cases in the trans formed space. As before, the single death because of cancer does not belong to the cluster of deceased patients. This analysis confirms the conclusions reached in Section III.1.

The clustering is best visible in Components 1, 2, 3 and 5 – by analyzing the loadings (see Table 1) we can hypothesize that the most important variables influencing the clustering are TIMI before the procedure, the admission mode, sex, age and heart failure. These variables would be of most interest in future, larger study.

Table 2 shows the proportion of variance explained by the principal components. It can be seen that the first three are responsible for 40% of the overall variance.

Table 1. The loadings for the PCA analysis. *Ad. mode* stands for *Admission mode*, *MI* for *Myocardial infarction* and *SBP* for *systolic blood pressure* – the rest of the variables should be self-explanatory

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7
Age	– 0.246	0.109	– 0.533	– 0.238		– 0.461	0.537
Sex	– 0.132	– 0.428	– 0.354		– 0.517		
Ad. mode	0.550	0.183		– 0.255	0.209	0.311	
MI	0.293	– 0.367	– 0.192	0.387	– 0.265	– 0.229	– 0.247
Hyp.tens.	0.139		– 0.435	– 0.655		0.412	– 0.393
Hyp.lip.		0.215	0.428	– 0.412	– 0.544	– 0.440	– 0.257
Diab.		– 0.508	0.311	– 0.356		0.142	0.472
TIMI	– 0.552	– 0.209		0.108	0.215		– 0.261
H.Fail	– 0.225	0.536	– 0.194	0.171	– 0.170	0.217	
SBP	0.393		– 0.178	– 0.122	0.471	– 0.510	– 0.204

Table 2. Proportion of variance explained by the respective Principal Components as well as the cumulative proportion for successive components

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Proportion of Variance	0.16	0.14	0.10	0.09	0.08	0.08
Cumulative Proportion	0.16	0.3	0.4	0.49	0.58	0.650
	Comp. 7	Comp. 8	Comp. 9	Comp. 10	Comp. 11	Comp. 12
Proportion of Variance	0.07	0.06	0.06	0.059	0.046	0.04
Cumulative Proportion	0.73	0.79	0.85	0.91	0.96	1.00000000

III.3. Brushing

In this subsection we use the dynamic procedure of brushing. This is one of the most recent exploratory techniques used to uncover the interrelations between data.

Brushing is an interactive method which allows one to identify the position of points or subsets of points in a plot (target plot) with the use of another plot (reference plot). Usually, the interpretation on the reference is clear and the interpretation or clustering on the other plot is to be established. Points, which are selected on the reference plot are highlighted in the target plot which allows for visual assessment and subsequent interpretation [16].

For brushing we use the PCA described in the previous section, and the procedure is illustrated with the multimedia file accompanying this paper. The `ggobi` package has been used in this part of the paper.

We have brushed the PCA representation of the dataset using the variables which did *not* enter the PCA, most importantly the binomial variables indicating whether a subject is a death/cancer case or not and whether the subject was on antiplatelet medication.

The brushing procedure for this data is available as an `.avi` file from [14].

The animation shows one of the many brushing analyses performed for the data described in the paper. It shows the first three Principal Components of the data without the ones to be brushed (death and antiplatelet medication). The cardiac death cases are highlighted (brushed) using the barchart at the bottom. This group is highlighted in the 3D PC scatterplot and a clustering is visible. The brushing also has an effect on the second barchart – it can be seen that the cardiac death cases cluster in the no-antiplatelet medication bar.

Then the no-antiplatelet group is brushed. From the 3D PC scatterplot it is clear that these cases also group in the same region of PC space as the deaths. This suggests that this may be the region of high risk of death.

Thus, we have found that the death cases are a subset of a larger cluster containing subjects which did not have the antiplatelet medication at the time of death (see the multimedia file). Medically, this conclusion is very interesting and interpretable.

The role of platelets in arterial thrombosis leading to acute myocardial infarction, stroke and coronary syndrome is crucial. The thrombotic may occur after aggregation, platelet adhesion and activation following endothelial disruption. Antiplatelet drugs are intended to reduce the occurrence of arterial thrombosis [11-13]. The present standards of the European Society of Cardiology [15] say that after the DES implantation a 12-month antiplatelet therapy is required. The potential result observed here may indicate that this period may have to be extended. Furthermore, not all patients conform to the prescribed medication, so another area for improvement is better monitoring of this aspect.

Therefore, the brushing technique has yielded another hypothesis which should be tested in a larger study: *the non-cancer deaths are related to the absence of antiplatelet medication*.

No similar conclusion was reached about the cancer group.

IV. SUMMARY

In the present paper we have performed an exploratory analysis of a group of subjects who have undergone the PTCA procedure. Our main interest was the formation of hypotheses on the rare adverse events in this group, namely death and cancer.

To perform the exploratory analysis we have used clustering dendrograms, principal component analysis and dynamic brushing.

We have found that the death cases cluster, while the cancer cases do not, which is visible in all analyses. The reason may be that the death cases are somehow related by other variables, which are mainly cardiological variables, while cancer cases are not. There could be other variables with respect to which the cancer cases cluster, but these were not available in the present study.

The clustering of the death cases is likely to be related to the age, sex, and admission mode – this is visible in the loadings obtained in the PCA.

The brushing technique has led to a very clear hypothesis about the clustering of the death cases, which is not directly accessible from the PCA. The death cases cluster is a sub-cluster of the subjects who were not on the antiplatelet medications. This is the strongest conclusion reached in the present paper and the hypothesis about the relation between death and the absence of antiplatelet medication is worth pursuing in a larger-scale study.

References

[1] E.D. Grech, *ABC of Interventional Cardiology.* BMJ Publishing Group, London (2004).

[2] D.S. Lippincott, *Grossman's Cardiac Catheterization.* Angiography & Intervention, 7-ed., Baim, Williams and Wilkins, Philadelphia (2006).

[3] P. Kay, *Cardiac Catheterization and Percutaneous Interventions I*. Taylor and Francis, London (2004).

[4] J. Tu et al., NEJM, 357, 1393-1402 (2007).

[5] A. Agresti, *Categorical Data Analysis*. 2-ed., Wiley-Interscience, Hoboken, New Jersey (2002).

[6] J.W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, Reading (1977).

[7] www.rproject.org

[8] www.ggobi.org

[9] W. Hardle, L. Simar, *Applied Multivariate Statistical*. Analysis 2-ed., Springer, Berlin (2007).

[10] B.S. Everitt, *An R and S-PLUS Companion to Multivariate Analysis*. Springer, Berlin (2005).

[11] T.N. Nguyen, *A Practical Handbook of Advanced Interventional Cardiology*. 2-ed. Futura, Blackwell Publishing, Malden (2003).

[12] E.J. Topol et al., *Textbook of Cardiovascular Medicine*. Lippincott Williams and Wilkins, Philadelphia (2002).

[13] P. Libby, R.O. Bonow, D.P. Zipes, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. 8-ed., Sounder Elsevier, Philadelphia (2008).

[14] http://www.if.uz.zgora.pl/˜jaropis/recBrush.html

[15] S.B. King et. al., J Am Coll Cardiol. 51, 172-209 (2008).

[16] D. Cook, D.F. Swayne, A. Buja, D.T. Lang, *Interactive and Dynamic Graphics for Data Analysis: With R and Ggobi (Use R)*. Springer, New York (2007).

Bartosz Bychowiec, MD, MSc (physics), medical doctor and a physicist. He is an interventional cardiologist working at catheterization laboratory at the Department of Cardiology – Intensive Therapy at the Poznań University of Medical Science. His professional and research interests are interventional cardiology and the physical aspect in interventional procedures and diagnostics.

Jaroslaw Piskorski, PHD, works at the Institute of Physics of the University of Zielona Góra, Poland. His main areas of interest include the application of the methods of statistical physics and nonlinear dynamics to heart rate variability, elementary particle physics (mainly discrete symmetries and their violation) and applied statistics.

Katarzyna Stanislawska, student at the Faculty of Medicine, Poznań University of Medical Science. Her main research interest include cardiology and radioneurology.

Mieczyslaw Dziarmaga, MD, PHD, specialist in internal diseases and an expert in interventional cardiology, head of the catheterization laboratory at the Department of Cardiology – Intensive Therapy at the Poznań University of Medical Science. His research interests include interventional cardiology and new technology in interventional cardiology.

Andrzej Minczykowski, MD, PhD, Associate Professor, Cardiologist, Department of Cardiology – Intensive Therapy, Poznań University of Medical Sciences. Specialist in internal diseases and cardiology, an expert in echocardiography. Main topic of interest: electrocardiography and echocardiography.

Andrzej Wykretowicz is Professor of Medicine at the University School of Medicine, Poznań, Poland. Specialist in internal diseases and cardiology. He is the deputy Head of the Department of Cardiology-Intensive Therapy. He graduated in medicine at the University School of Medicine in Poznań in 1981, and completed postgraduate medical training in internal medicine and cardiology. He was awarded his PhD in 1985. Dr Wykretowicz accomplished his postdoctoral research training in Dalhousie University, Halifax, Canada as a recipient of I.W. Killam Research Fellowship and in Institut fur Pharmazie, Freie University Berlin, Germany as a recipient of European Union fellowship. His research interests include noninvasive cardiology, diabetology as well as hypertension, and is primarily focused on the role of inflammation in human pathology.

Henryk Wysocki, Full Professor of Medicine working at the Poznań University of Medical science, specialist in internal diseases, cardiologist and expert on hypertensiology. He is the province cardiology consultant for the Wielkopolskie Province and the Head of the Department of Cardiology – Intensive Therapy at the Poznań University of Medical Science. He has done research and published extensively in the fields of cardiology, including interventional cardiology, hematology, heart rate variability, thoracic surgery, diabetology and many others.