JOURNAL OF APPLIED
COMPUTER SCIENCE
Vol. 21 No. 2 (2013), pp. 209-219

On Statistical Analysis Methods Improving Epidemiological Studies

Agnieszka Wosiak

Institute of Information Technology
Lodz University of Technology
Wólczańska 215, 90-924 Łódź
agnieszka.wosiak@p.lodz.pl

Abstract. *Nowadays almost all the principles of diagnosis and treatment are determined by the statistical analysis of the observations made by practitioners. The analysis of medical data is therefore one of the most important elements that affect the level of modern medical care. The research described in this paper aims to determine the characteristics of cardiac parameters in healthy children and in children with a diagnosis in arterial hypertension. Studies include children in the Lodz region. The purpose of the analysis is to determine risk factors of arterial hypertension and thus early diagnosis of children and as quickly as possible the inclusion of an appropriate treatment or observation. The analysis applies a number of methods, including descriptive analysis, grouping and statistical inference. The choice of methods is consistent with the requirements of the USMLE (The United States Medical Licensing Examination) and depends on the evaluated parameters. The research is carried out using professional statistical packages and the computer system designed and developed for this study. The use this system allows for the continuous process of research staging of new cases, which is a definite advantage when conducting epidemiological studies.*

Keywords: *statistical analysis, medical data analysis, epidemiological system, statistical inference.*

1. Introduction

Nowadays the medical progress makes possible to obtain the increasing amount of data for practitioners and clinics. The availability of these data causes the expectation that medical staff will make use of them to improve patient care, develop new therapies and improve existing ones. Almost all the principles of diagnosis and treatment are determined by the statistically developed results of medical observations. At the same time the problem of receiving the results of the analysis as quick as possible and correctly is growing up. Consequently, it is very important to build effective tools that are available to support the process of data analysis. The analysis of medical data is therefore one of the most important elements that affect the level of modern medical care.

The numerous data in medicine challenges the computer scientist to provide solutions for effective data storage, processing and exchange. A major challenge is to support the knowledge discovery process, which means extraction as much as possible useful knowledge from data. There are two main paths of applications for data analysis: (a) descriptive analysis to extract summaries of important features, such as grouping patients with similar syndromes and find important characteristics of each disease; and (b) predictive analysis to derive classification rules, such as identifying rules which predict the future presence or the course of a disease. Both of them can be led with many different computer techniques.

The selection of appropriate methods of analysis is a key factor in the research process and the results of this selection may have a significant impact on the further work on the development of new medical diagnostic methods. In most of the cases the principles of diagnosis and treatments are determined by the statistical analysis of medical data.

The aim of this study is to develop a computer system to support epidemiological research based on statistical data analysis. Using the techniques of data analysis in the process of conducting epidemiological studies, in particular in the field of cardiology, allows for early diagnosis and improvement of the conditions of treatment [1, 2, 3].

The paper is organized as follows. Section 2 presents relevant background information. Next section concerns epidemiological studies and the role of statistical analysis in these research. In Section 4 we describe the methodology of the research. We discuss selected, most appropriate for our data, techniques of statistical analysis and present the results of our studies. Finally, in Section 5 we discuss what we draw the conclusions.

2. Related work

Now, almost all the principles of diagnosis and treatment are determined by the statistical analysis of medical data. Therefore the literature survey reveals many results on statistical analysis [4, 5, 6, 7].

In [8] a statistical inference of heart rate and blood pressure was examined. The data were obtained form adults by measurements with the invasive method in the radial artery. Three different approaches were tested. First one was based on correlation between raw data. Than, since the measurements could be corrupted by noise, the signals were also correlated after performing a filtration procedure. Another approach was based on least squares approximation. The results of these three different methods were similar. The observed correlation coefficient seemed a random number and unpredictable, however the short-term correlation was relatively large.

The authors of [9] discussed the challenges of leading medical research for treatment of cancer. They distinguished three major steps of clinical development: phase I studies to find the maximum tolerated dose of a drug, phase II trials to identify the cancer types in which the drugs showed some degree of biological anti-cancer activity, and phase III for comparative studies of drugs to establish therapies. The important role of statistical analysts is pointed to work on new methodologies for early development of targeted agents and for the exploration and validation of imaging markers.

In [10] the challenges faced in handling data from biomarker studies were introduced and methods for the appropriate analysis and interpretation of these data were described. The research showed that basic statistical methods, i.e. testing distribution assumptions, testing for trend in proportions, testing for linear-by-linear association and Spearman's correlation, when properly applied, were sufficient. However some of statistical techniques, such as the maximum likelihood approach for dealing with non-detectable values in the analysis of biomarker data, were pointed as particularly useful in dealing with biomarker data.

3. Epidemiological Studies

One of the main tasks of epidemiology is to search and test activities for the modification of factors affecting human health. The essence of epidemiological studies is to determine the causal relations between the state of health of patients

and the object of measurement. The process of conducting epidemiological studies typically consists of the following four steps [11]:

- test preparation,
- collection of observations for study and control groups,
- preparation of research results
- analysis of the results of observation.

The first two stages are the domain of the medical staff, although knowledge of statistical methods makes the process of planning more appropriate and enables achieving assumed goals.

The third stage, the preparation of research results is based on an appropriate and systematic comparison of research results. During this stage the individual data of all cases are subjected to the classification and grouping. As the result working tables are obtained with aggregated data from which the process of analysis can be led in the next stage.

Analysis of the results of observation, is the most important step in the process of conducting a medical research, because, depending on the conclusions arising from this analysis, there will be the possibility of practical application of research results. This phase includes the development of descriptive statistics for the collected data (measures of central tendency and dispersion position), the definition of statistical distributions and statistical inference (estimation and verification).

4. Methodology and Experimental Results

The process of medical research supported by statistical analysis can be described by the system architecture shown in Fig.1.

The process of leading medical research starts with data acquisition. The data are collected by performing observations, examinations, laboratory measurements and interviews.

The first stage of the medical data analysis is the selection of appropriate subset of the available features. Medical classification analysis applications generally take one of the following two feature selection approaches: including use of automatic feature selection mechanisms or expert judgment. In our approach we use expert judgment to obtain a set of attributes for further analysis. This approach is widely

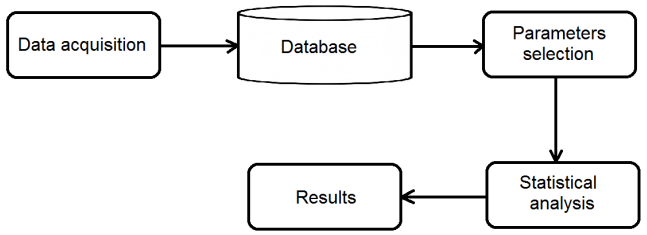

Figure 1. System architecture for statistical analysis solution

used in medical classification analysis. The literature study [12] shows that for the cardiovascular diseases the feature subsets selected by experts improve the sensitivity of the analysis.

The analysis of the obtained results usually begins with an assessment of measures of descriptive statistics, i.e. the characteristic of variables. The descriptive analysis also allows, in addition to raising the cognitive, detecting errors that were not identified during the preparation phase of the obtained results. The basic descriptors of descriptive statistics, for which the evaluation is indicated, include measures of central tendency (arithmetic mean, median and modal), measures of dispersion (range and standard deviation).

The next step is to carry out data analysis of statistical inference using a suitable test. The selection of the test is made on the basis of the kind and the structure of the analyzed data. A list of commonly performed tests is presented in Table 1. Their selection is in accordance with the requirements of the USMLE (The United States Medical Licensing Examination) [13].

This research involves the following kinds of test listed in Table 1:

- One sample *t* test,
- Kolmogorov–Smirnov test,
- Wilcoxon signed-rank test,
- Unpaired two-sample Student's *t*-test,
- Mann–Whitney U test.

Table 1. Statistical tests selection

Number of experimental groups	**Scale type**	**Kind of hypothesis**	**Test name**
1	Ratio/interval scale	Hypothesis of a population mean	One sample *t* test
1	Ratio/interval scale	Hypothesis of a normality of the distribution	Kolmogorov–Smirnov test Shapiro–Wilk test Lilliefors test
1	Ordinal scale	Hypothesis about a population median	Wilcoxon signed-rank test
2 (Independent variables)	Ratio/interval scale	Hypothesis of the difference between two means	Unpaired two-sample Student's *t*-test Welch's *t*-test
2 (Independent variables)	Ratio/interval scale	Hypothesis about the equality of variances	Fisher–Snedecor test Levene's test Brown—Forsythe test
2 (Dependent variables)	Ratio/interval scale	Hypothesis of the difference between two means	Paired two-sample Student's *t*-test
2 (Independent variables)	Ratio/interval scale	Hypothesis of the difference between the medians	Mann–Whitney U test

Each of the different tests are assessed at significance level 5% unless stated otherwise.

The impact of one variable measured in an interval or ratio scale to another variable in the same scale can be expressed using the Pearson's correlation coefficient r_P as follows:

$$r_P = \frac{cov(X, Y)}{\sigma_X \sigma_Y}, \tag{1}$$

where the covariance of the two variables X and Y is divided by the product of their standard deviation.

In the case where one or both of the variables are measured with an ordinal scale, or variables are expressed as an interval scale, but the relationship is not a linear relationship, the Spearman's correlation test should be used:

$$r_S = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}, \tag{2}$$

where variables X_i and Y_i are converted to ranks x_i and y_i.

The research aims to determine the characteristics of cardiac parameters in healthy children and in children with a diagnosis in arterial hypertension. The studies include children in the Lodz region. The purpose of the analysis is to determine risk factors of arterial hypertension and thus early diagnosis of children and as quickly as possible the inclusion of an appropriate treatment or observation. The computer analysis applies a number of methods, including descriptive analysis, grouping and statistical inference. The choice of methods is consistent with the requirements of the USMLE (The United States Medical Licensing Examination) and depends on the evaluated parameters.

The analytical research are carried out using professional statistical packages (Statistica 10 by StatSoft) and a computer system designed and developed for this study (based on Microsoft SQL Server database management system, procedural language T-SQL and PHP language). The use of this own solution allows for the continuous process of research staging of new cases, which is a definite advantage when conducting epidemiological studies.

The data on the health status of children are collected from medical records and by conducting environmental surveys. All the patients undergo physical examination, manual measurements of arterial pressure, ambulatory blood pressure monitoring, echocardiographic examination with cardiac function evaluation with the use of standard parameters and tissue Doppler examination.

Figure 2. System panel for results of statistical analysis

Based on the derived statistical analysis as well as set out rules and relationships, the system reports the results with visualization in the form of charts, tables, and the aggregate values.

In order to obtain the results of the process of statistical inference, the patient data are grouped using such attributes as age, sex, date of a disease and the date of the examination. The application process includes identifying measures of descriptive statistics, comparative analysis in groups of healthy and study children, and between different categories. Above all, the correlation analysis is performed, in order to determine the factors of family, maternal and environmental conditions that may affect the incidence of children associated with the cardiovascular system: hypertension and hypotrophy.

A typical query on the results of aggregate data collected in the system can be described as follows:

Table 2. Sample descriptions derived from the system

Number	**Description**
1	There are no statistically significant differences between the groups in the distributions of *[parameter name]*
2	A significantly higher *[parameter name]* is found in patients from *[control/study]* group
3	A significantly lower mean *[parameter name]* was found in patients from *[control/study]* group
4	There are *[no]* statistically significant differences between groups in measurements of *[parameter names]*.
5	There is a statistically significant *[negative]* correlation between *1st parameter name* and *2nd parameter name*.

```
SELECT
count(case_id),
age,
disease_age,
diagnose
FROM patient_data
WHERE age BETWEEN ... AND ... AND sex = ....
AND environmental_factors = TRUE
AND maternal_factors = TRUE
GROUP BY age, diagnose
ORDER BY age, diagnose;
```

The sample panel with the results of analysis based on aggregate data is shown in Figure 2.

The results of the statistical analysis, in addition to tabular form, are presented in the form of descriptions. The sample descriptions are presented in the Table 2. This form of results can be a prerequisite for further research to scientific medical staff.

5. Conclusions

The process of determining the correlation between the results according to different kinds of medical research is essential for the development of disease diagnosis and improving the process of detection of diseases and their treatment. According to the consultations with the scientific medical staff , the solution proposed in this paper can significantly support medical research, particularly epidemiological studies. Therefore there is a strong need to implement this kind of computer applications in scientific and medical institutions.

In the future, a very useful part of the system would be to implement the possibility of classifying and aggregating not only on the basis of static metadata predefined in the system, but also to enable dynamic input full range of analytical parameters, so that the potential of a system for epidemiological could significantly increase.

References

[1] Feber, J. and Ahmed, M., *Hypertension in children: new trends and challenges*, Clinical Science, Vol. 119, 2010, pp. 151–161.

[2] Orlowska-Wlodarczyk, B., *Znaczenie badan epidemiologicznych dla rozwoju kardiologii prewencyjnej*, Polski Przeglad Kardiologiczny, Vol. 5, No. 1, 2003, pp. 85–89.

[3] Zamojska, J., Niewiadomska-Jarosik, K., Wosiak, A., and Stanczyk, J., *Ocena funkcji skurczowej lewej komory z wykorzystaniem metody doplera tkankowego u dzieci z nadcisnieniem tetniczym pierwotnym*, Polski Przeglad Kardiologiczny, Vol. 14, No. 2, 2012, pp. 95–100.

[4] Chang, M., *Modern Issues and Methods in Biostatistics*, Springer, 2011.

[5] Gurka, M. and Edwards, L., *Mixed Models for Medical Statistics*, In: Essential Statistical Methods for Medical Statistics, edited by C. Rao, J. Miller, and R. D.C., Elsevier, 2011.

[6] Hilbe, J. and Greene, W., *Count Response Regression Models*, In: Essential Statistical Methods for Medical Statistics, edited by C. Rao, J. Miller, and R. D.C., Elsevier, 2011.

[7] Prentice, R., *Statistical Methods and Challenges in Epidemiology and Biomedical Research*, In: Essential Statistical Methods for Medical Statistics, edited by C. Rao, J. Miller, and R. D.C., Elsevier, 2011.

[8] Polinski, A., Kot, J., and Meresta, A., *Analysis of Correlation Between Heart Rate and Blood Pressure*, In: Proceedings of the Federated Conference on Computer Science and Information Systems, 2011, pp. 417–420.

[9] Collette, L., Bogaerts, J., Suciu, S., Fortpied, C., Gorlia, T., Coens, C., and et, a., *Statistical methodology for personalized medicine: New developments at EORTC Headquarters since turn of the 21st Century*, European Journal of Cancer Suplements, Vol. 10, No. 1, 2012, pp. 13–19.

[10] Looney, S. and Hagan, J., *Statistical Methods for Assessing Biomarkers and Analyzing Biomarkers Data*, In: Essential Statistical Methods for Medical Statistics, edited by C. Rao, J. Miller, and R. D.C., Elsevier, 2011.

[11] Stanisz, A., *Przystepny kurs statystyki z zastosowaniem STATISTICA PL na przykladach z medycyny*, Wydawnictwo StatSoft, Krakow, 2006.

[12] Cheng, T., Wei, C., and Tseng, V., *Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches*, In: Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems, CBMS'06.

[13] Moczko, J. and Breborowicz, G., *Nie sama biostatyka*, Osrodek Wydawnictw Naukowych, Poznan, 2010.