EJBI 2013 ISSN 1801 - 5603

European Journal for Biomedical Informatics

# European Journal for Biomedical Informatics

An Official Journal of the European Federation for Medical Informatics

Volume 9 (2013), Issue 3

Special Topic

**Sharing Knowledge and Tools for Decision Support in Biomedicine and Healthcare**

Editors

**John Mantas, Arie Hasman, Jana Zvárová**

www.ejbi.eu

## Aims and Scope

The **European Journal for Biomedical Informatics** reacts on the great European need to share the information in the multilingual and multicultural European area. The journal publishes peer-reviewed papers in English and other European languages simultaneously. This opens new possibilities for faster transfer of scientific-research pieces of knowledge to large international community of biomedical researchers, physicians, other health personnel and citizens. From time to time, the journal publishes articles on particular focus themes as part of a journal's issue.

The generally accepted translations of the English version of the abstract and keywords or full paper are to the European languages, which can be found at `http://www.ejbi.org/en/about/`.


## Editors and Management

**Editor in Chief**
Jana Zvárová, Czech Republic

**Managing Editor**
Petra Přečková, Czech Republic

**Graphic Design**
Anna Schlenker, Czech Republic

**Sales and Marketing Manager**
Karel Zvára, Czech Republic

## Editorial Board: National Members

| | |
|---|---|
| Ammenwerth, Elske | Austria |
| Masic, Izet | Bosnia and Herzegovina |
| Vinarova, Jivka | Bulgaria |
| Kern, Josipa | Croatia |
| Zvárová, Jana | Czech Republic |
| Andersen, Stig Kjaer | Denmark |
| Ruotsalainen, Pekka | Finland |
| Degoulet, Patrice | France |
| Horsch, Alexander | Germany |
| Mantas, John | Greece |
| Surján, György | Hungary |
| Hurl, Gerard | Ireland |
| Reichert, Assa | Israel |
| Mazzoleni, Cristina | Italy |
| Lukosevicius, Arunas | Lithuania |
| Hofdijk, Jacob | Netherlands |
| Moen, Anne | Norway |
| Bobrowski, Leon | Poland |
| da Costa Pereira, Altamiro | Portugal |
| Mihalas, George | Romania |
| Shifrin, Michael | Russian Federation |
| Živčák, Jozef | Slovakia |
| Orel, Andrej | Slovenia |
| Nordberg, Ragnar | Sweden |
| Lovis, Christian | Switzerland |
| Saka, Osman | Turkey |
| Mayorow, Oleg | Ukraine |
| de Lusignan, Simon | United Kingdom |

## Editorial Board: Representatives of Cooperating Journals

| | |
|---|---|
| Mayorow, Oleg | Clinical Informatics and Telemedicine |
| Marolt, Christian | Health IT Management |
| Brumini,Gordana | Hrvatski društvo za medicinsku informatiku |
| Rosina, Jozef | Lékař a technika |
| Svačina, Štěpán | Medicína po promoci |
| Haux, Reinhold | Methods of Information in Medicine |



## Publisher

EuroMISE s.r.o.
Paprsková 330/15
CZ-14000 Praha 4
Czech Republic
EU VAT ID: CZ25666011

## Office

EuroMISE s.r.o.
Paprsková 330/15
CZ-14000 Praha 4
Czech Republic

## Contact

Karel Zvára
zvara@euromise.com


## Instructions to Authors

Please access http://www.ejbi.org/en/instructions/

## EJBI Online

The online version of the full volume is available at http://www.ejbi.org/en/ejbi/.


©2013 EuroMISE s.r.o.

# Instructions to Authors

## General Remarks

This journal follows the guidelines of the International Committee of Medical Journal Editors (http://www.icmje.org/index.html) and the Committee on Publication Ethics (http://www.publicationethics.org).

Authors should especially be aware of the following relevant issues in these guidelines:

### Authorship

All authors should have made

(1) substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data;

(2) drafting the article or revising it critically for important intellectual content; and

(3) final approval of the version to be published.

### Conflicts of interest

All authors must disclose any financial and personal relationships with other people or organizations that could inappropriately influence (bias) their actions.

### Protection of human subjects and animals in research

Authors who submit a manuscript on research involving human subjects should indicate in the manuscript whether the procedures followed were in compliance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects (http://www.wma.net/en/30publications/10policies/b3/).

*European Journal for Biomedical Informatics* does not publish material that has already appeared elsewhere. Submitted manuscripts should not be submitted in parallel to any other journal.

## Manuscript preparation

Authors are kindly requested to carefully follow all instructions on how to write a paper. In cases where the instructions are not followed, the paper will be returned immediately with a request for changes, and the editorial review process will only start when the paper has been resubmitted in the correct style.

Authors are responsible for obtaining permission to reproduce any copyrighted material and this permission should be acknowledged in the paper.

Authors should not use the names of patients. Patients should not be recognizable from photographs unless their written permission has first been obtained. This permission should be acknowledged in the paper.

In general the manuscript text (excluding summary, references, figures, and tables) should not exceed 5 000 words.

Kindly send the final and checked source and PDF files of your paper to `manuscripts@ejbi.org`. You should make sure that the LaTeX and the PDF files are identical and correct and that only one version of your paper is sent. Please note that we do not need the printed paper.

Where appropriate, the paper should be organised into the following sections: *Abstract, Introduction, Objectives, Methods, Results, Discussion, Conclusions, Acknowledgments and References.* Apart from the main headings, subheadings should be used and may be numbered.

Authors are strongly encouraged to use LaTeX $2_\varepsilon$ for the preparation of manuscript. The LaTeX template `ejbi_template.tex` can be downloaded from www.ejbi.org/en/instructions/.

When you are not able to use LaTeX, please use MS Word or OO Writer and send us the unformatted text. Kindly follow just instructions about preparing figures, tables and references. We are going to convert your text into LaTeX instead of you.

If you use LaTeX together with our template file, `ejbi_template.tex`, your text is typeset automatically. Please do *not* change the preset fonts. Do not use your own macros, or styles.

Please use the commands `\label` and `\ref` for cross-references and the commands `\bibitem` and `\cite` for references to the bibliography, to enable us to create hyperlinks at these places.

### Title page

The first page of the article should contain: title of the paper (also the shorter version for running heads), initials and last name of each author, to be followed with their institutional affiliations, the name, address, e-mail address and telephone of the corresponding author.

### Abstract and Keywords

The abstract should summarize the contents of the paper and should not exceed 250 words. Authors are requested to write a structured summary, adhering to the following headings: Background (optional), Objectives, Methods, Results, Conclusions.

At the end of the Abstract, the contents of the paper should be specified by, at most, five keywords. We recommend using MeSH keywords.

### Headings

Headings should be capitalized (i.e. nouns, verbs, and all other words except articles, prepositions, and conjunctions should be set with an initial capital) and should be aligned to the left. Words joined by a hyphen are subject

©2013 EuroMISE s.r.o. 

to a special rule. If the first word can stand alone, the second word should be capitalized.

### Figures and Tables

Attach figures and tables as separate files. Do not integrate them into the text. Do not save your table as an image file or insert a table into your manuscript text document as an image. Figures and tables should be referenced in the manuscript by their numbers.

Annotations belong in a (self-)explanatory legend, do not use headings in the figure, explain abbreviations in the legend. Label all axes. Use a uniform type size (we recommend Arial 10 point), and avoid borders around tables and figures.

Submit graphics as a sharp printout as well as a file. The printout and the file must be identical. Submit the image file with clear labelling (e.g. Fig_1 instead of joint_ap).

Image resolution is the number of dots per width of 1 inch, the "dots per inch" (dpi). Printing images require a resolution of 800 dpi for graphics and 300 dpi for photographics.

Vector graphics have no resolution problems. Some programs produce images not with a limited number of dots but as a vector graphic. Vectorisation eliminates the problem of resolution. However, if halftone images ("photos") are copied into such a program, these images retain their low resolution.

If screenshots are necessary, please make sure that you are happy with the print quality before you send the files.

In the printed volumes, illustrations are generally black and white (halftones), and only in exceptional cases, and if the author is prepared to cover the extra cost for colour reproduction, are coloured pictures accepted. Coloured pictures are welcome in the electronic version free of charge. If you send coloured figures that are to be printed in black and white, please make sure that they really are legible in black and white. Some colours as well as the contrast of converted colours show up very poorly when printed in black and white.

### Formulas

Displayed equations or formulas are centred and set on a separate line (with an extra line or halfline space above and below). Displayed expressions should be numbered for reference. The numbers should be consecutive within each section or within the contribution, with numbers enclosed in parentheses and set on the right margin.

### Footnotes

The superscript numeral used to refer to a footnote appears in the text either directly after the word to be discussed or – in relation to a phrase or a sentence – following the punctuation sign (comma, semicolon, or period). Footnotes should appear at the bottom of the normal text area, with a line of about 2 cm set immediately above them.[1]

### Program Code

Program listings or program commands in the text are normally set in a typewriter font, e.g. CMTT10 or Courier.

### Acknowledgements

Scientific advice, technical assistance, and credit for financial support and materials may be grouped in a section headed Acknowledgements that will appear at the end of the text (immediately after the Conclusions section). The heading should be treated as a subsubsection heading and should not be assigned a number.

In case that a financial support of the paper development (e.g. sponsors, projects) is acknowledged, the fee of 50 EUR will be charged by Publisher. The accepted peer-reviewed papers with an acknowledgement of a financial support, where the fee was not paid, will be published free of charge, but the financial acknowledgement will be withdrawn.

### References

The list of references is headed "References" and is not assigned a number. The list should be set in small print and placed at the end of your contribution, in front of the appendix, if one exists. Please do not insert a pagebreak before the list of references if the page is not completely filled. For citations in the text please use square brackets. In the text number the references consecutively in the order in which they first appear. Use the style, which is based on the formats used by the US National Library of Medicine in MEDLINE (sometimes called the "Vancouver style"). For details see the guidelines from the International Committee of Medical Journal Editors (http://www.nlm.nih.gov/bsd/uniform_requirements.html).

Examples:

[1] Blobel B. Architectural Approach to eHealth for Enabling Paradigm Changes in Health. Methods Inf Med. 2010; 49(2): 123–134.

[2] Kalina J. Robustní analýza obrazu obličeje pro genetické aplikace. EJBI [Internet]. 2010 [cited 2011 Jun 28]; 6(2): cs95–cs102. Available from: http://www.ejbi.eu/articles/201012/47/2.html

[3] van Bemmel JH, Musen M, editors. Handbook of Medical Informatics. Heidelberg: Springer; 1997.

[1] The footnote numeral is set flush left and the text follows with the usual word spacing.


©2013 EuroMISE s.r.o.

[4] Zvarova J, Zvara K. e3Health: Three Main Features of Modern Healthcare. In: Moumtzoglou A, Kastania A. E-Health Systems Quality and Reliability: Models and Standards, Hershey: IGI Global; 2010; 18–27.

### Multilingual Issue

The authors are asked to translate English version of Abstract and Keywords to at least one European languege. The translated versions of Abstract and Keywords should be send to `manuscripts@ejbi.org`

### Checking the PDF File

Kindly assure that the Contact Volume Editor is given the name and email address of the contact author for your paper. The contact author is asked to check through the final PDF files to make sure that no errors have crept in during the transfer or preparation of the files. Only errors introduced during the preparation of the files will be corrected.

If we do not receive a reply from a particular contact author, within the timeframe given, then it is presumed that the author has found no errors in the paper.

### Copyright Transfer Agreement

The copyright form may be downloaded from http://www.ejbi.org/en/downloads/. Please send your signed copyright form to the Contact Volume Editor, either as a scanned pdf or by fax or by courier. One author may sign on behalf of all the other authors of a particular paper. Digital signatures are acceptable.

## EuroMISE Copyright Permission Policy

Written permission is required to reproduce material from EuroMISE s.r.o. publications in other publications, electronic products, or other media. To obtain a copyright permission please contact Jana Zvárová: zvarova@ejbi.org, fax: +420 241471337. You may fax or e-mail your request along with the full citation of the journal in which the paper appears in with volume number and page number(s) as well as what you are requesting to use the material for. Use of copyrighted material always requires proper citation.

Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND): for non-commercial purposes, lets others distribute and copy the article, and to include in a collective work (such as an anthology), as long as credit the author(s) and provided they do not alter or modify the article.

# Contents



 ©2013 EuroMISE s.r.o.

# Sharing Knowledge and Tools for Decision Support in Biomedicine and Healthcare

**John Mantas[1], Arie Hasman[2] and Jana Zvárová[3]**

[1] University of Athens, Greece

[2] University of Amsterdam, The Netherlands

[3] Charles University in Prague, The Czech Republic

The special topic conference Data and Knowledge for Medical Decision Support was held in Prague, April 17-19, 2013 under the auspices of the European Federation of Medical Informatics (EFMI). The EFMI conference focused on the important topic of how advanced information and communication technologies can be used to combine expert knowledge with best practices in the diagnosis, treatment and prevention of diseases. After review of all submissions a scientific program could be established with 49 lectures, 32 posters, 7 workshops and one panel. The conference was attended by more than 160 delegates from 31 countries. Selected papers of the EFMI conference were published by IOS Press [1].

One of the workshops titled Sharing Knowledge and Tools for Decision Support in Biomedicine and Healthcare was held at the 1st Faculty of Medicine of the Charles University in Prague on April 16, 2013. During the workshop 25 Ph.D. students from Charles University in Prague, Masaryk University in Brno, Technical University in Prague and several universities from other European countries presented their work. The workshop was chaired by Prof. John Mantas (University of Athens, Greece), Prof. Arie Hasman (University of Amsterdam, The Netherlands) and Prof. Jana Zvárová (Charles University in Prague, The Czech Republic). Discussions about the presented lectures were at an international level with active involvement of both the chairs and participants. This created an added value for the Ph.D. students.

Based on their presentations during the workshop a number of students wrote full papers. These papers were submitted to different biomedical journals and ten of them to the European Journal for Biomedical Informatics (EJBI), listed among peer-reviewed journals published in the Czech Republic and approved by the Board of Government. The third issue of EJBI publishes in the year 2013 six full papers after a peer-review. We sincerely hope that the active participation at the workshop and the further comments of reviewers about the full papers highly contributed to the students´ knowledge for a successful further development of their Ph.D. theses.

## Reference

[1] Arie Hasman, Bernd Blobel, Jana Zvárová (eds): Data and Knowledge for Medical Decision Support. IOS Press, Amsterdam 2013

©2013 EuroMISE s.r.o.

# Shrinkage Approach for Gene Expression Data Analysis

**Jiří Haman**[1,2], **Zdeněk Valenta**[1]

[1]Institute of Computer Science AS CR, v. v. i., Department of Medical Informatics and Biostatistics, Prague, Czech Republic

[2]First Faculty of Medicine, Charles University in Prague, Czech Republic

## Abstract

**Background**: Microarray technologies are used to measure the simultaneous expression of a certain set of thousands of genes based on ribonucleic acid (RNA) obtained from a biological sample. We are interested in several statistical analyses such as 1) finding differentially expressed genes between or among several experimental groups, 2) finding a small number of genes allowing for the correct classification of a sample in a certain group, and 3) finding relations among genes.
**Objectives**: Gene expression data are high dimensional, and this fact complicates their analysis because we are able to perform only a few samples (e.g. the peripheral blood from a limited number of patients) for a certain set of thousands of genes. The main purpose of this paper is to present the shrinkage estimator and show its application in different statistical analyses.

**Methods**: The shrinkage approach relates to the shift of a certain value of a classic estimator towards a certain value of a specified target estimator. More precisely, the shrinkage estimator is the weighted average of the classic estimator and the target estimator.
**Results**: The benefit of the shrinkage estimator is that it improves the mean squared error (MSE) as compared to a classic estimator. The MSE combines the measure of an estimator's bias away from its true unknown value and the measure of the estimator's variability. The shrinkage estimator is a biased estimator but has a lower variability.
**Conclusions:** The shrinkage estimator can be considered as a promising estimator for analyzing high dimensional gene expression data.

## Keywords

Microarray technology, high dimensional data, mean squared error, James-Stein shrinkage estimator, mutual information

### Correspondence to:

Jiří Haman
Institute of Computer Science AS CR, v. v. i.
Address: Pod Vodárenskou věží 271/2, Prague 8, Czech Republic
E–mail: j.haman@seznam.cz, haman@cs.cas.cz

**EJBI 2013; 9(3):2–8**
received: August 1, 2013
accepted: October 31, 2013
published: November 20, 2013

## 1 Introduction

In this section we describe microarray technology, and mention several types of hypotheses for microarrays together with some problems which arise with microarray data.

### 1.1 Microarray Technology

Microarray technology is an important element in gene expression assessment. Basically, we can describe microarray as a solid wafer which contains from a few thousand to millions of one-stranded segments of deoxyribonucleic acid (DNA) nucleotides in precisely set positions. The positions are often called spots or probes. These spots correspond to individual genes. The mode of microarray production is described e.g. in [1].

Gene expression measurement is conducted by extracting ribonucleic acid (RNA) from a biological sample, e.g. the peripheral blood of a patient or a tuberculous tissue. The extracted RNA sample is marked with a fluorescent dye and spread on a microarray chip. During a process called hybridization the RNA sample binds with the microarray thanks to the complementarity of nucleotides.

The RNA concentration on a microarray is determined using fluorescence. The microarray is scanned using a special scanner and the output has the form of a two-dimensional scanned image. The scanned image is formed by pixels with certain intensities. A group of neighbouring pixels making up the spot corresponding to the fluorescent activity of a specific gene. The spots must be targeted using a specialized image analysis software.


*©2013 EuroMISE s.r.o.*

After performing complicated preprocessing techniques, we obtain summarized values of gene expression for each spot. For more information about the techniques we refer the reader to [2] or [3].

Because we have usually several comparable biological samples, the final data of gene expressions are in the form of matrix. The matrix's rows typically correspond to individual genes and the columns correspond to individual samples (e.g. patients) of gene expression profiles from several microarrays.

### 1.2 Types of Analysis for Microarrays

Several types of statistical inferences are possible in the context of information extracted from microarrays:

(a) detection of differentially expressed genes between/among several experimental sample groups (e.g. comparing gene expression of patients with a tumour and without a tumour),

(b) classification of an unknown sample in a specific group based on the gene expression profile in the sample (e.g. classification of a patient with cancer, i.e. confirmation of the onset of cancer),

(c) discovery of new/unknown groups of genes (e.g. new tumour subtype),

(d) identification of genes which are important for a given group (e.g. for myocardial infarction),

(e) creation of gene networks (description of relations among/between genes in evolution of cancer).

A broad overview of statistical analysis for gene expression data analysis can be found in [3] or [4].

### 1.3 Problem of High Dimension

The problem with certain kinds of microarray technology, e.g. whole-genome microarrays, is high dimensionality due to the fact that there is a greater number of variables (genes) than observations (samples). If we consider $p$ to be the number of variables and $n$ to be the number of observations this means $n \ll p$. Problems associated with high dimensionality include the following, for example:

- the selection of relevant differentially expressed genes, i.e. the problem of multiple testing (see [5]),
- tractability of linear discriminant analysis (LDA) used for classification purposes. LDA requires an estimator of the inverse covariance matrix, which, however, does not exist for the data which we have in this case (see [6]).

## 2 Shrinkage Approach

In this section we present a shrinkage approach which can help deal with the high dimensionality of gene expression data. The approach results in an estimator based on reducing the mean squared error (MSE) compared to a classic estimator.

Firstly, we present the origin of the shrinkage approach, i.e. James-Stein estimator (JSE) which is motivated by the estimation of an unknown mean vector for multivariate normal distribution based on 1 observation from this distribution. Next, we mention the optimal value of the JSE's parameter which leads to the lowest value of the MSE. This optimal value also gives the estimator the shrinkage attribute. Finally, a generalization for both the JSE and its optimal value is mentioned.

### 2.1 Mean Squared Error (MSE)

Let us suppose we have a random sample from a nondegenerated distribution with an unknown one-dimensional parameter $\theta \in \mathbb{R}$. We calculate an estimator $\hat{\theta} \in \mathbb{R}$ on the basis of our data. Then the MSE (or quadratic risk) for the parameter $\hat{\theta}$ is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2. \quad (1)$$

Expression (1) can be rewritten in a different way. We expand $(\hat{\theta} - \theta)^2$ in (1) and "add-subtract" $(\mathrm{E}\hat{\theta})^2$. After using $\mathrm{var}(\hat{\theta}) = \mathrm{E}(\hat{\theta})^2 - (\mathrm{E}\hat{\theta})^2$ and $\theta = \mathrm{E}\theta$ and $\mathrm{E}\theta^2 = (\mathrm{E}\theta)^2$ we get an equivalent form of (1)

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}, \theta))^2. \quad (2)$$

In (2) the term $\mathrm{Bias}(\hat{\theta}, \theta) = \mathrm{E}(\hat{\theta} - \theta)$ represents the accuracy of estimator $\hat{\theta}$ with respect to the unknown parameter $\theta$.

We can see from (2) that (1) for MSE of estimator $\hat{\theta}$ is the sum of its variance and the square of its bias. If $\mathrm{Bias}(\hat{\theta}, \theta) = 0$ we say that estimator $\hat{\theta}$ is the unbiased estimator of parameter $\theta$. If $\mathrm{Bias}(\hat{\theta}, \theta) \neq 0$ we say that estimator $\hat{\theta}$ is the biased estimator of parameter $\theta$.

The idea for introducing the MSE is that allowing a small bias for estimator $\hat{\theta}$ of parameter $\theta$ can substantially improve its variability $\mathrm{var}(\hat{\theta})$. For more details we refer the reader to [7].

### 2.2 James-Stein Estimator (JSE)

The JSE, also known as the Stein estimator or shrinkage estimator, is an estimator which first appeared in [8]. The motivation for this estimator is the following.

Let us assume that $p \in \mathbb{N}$ and we have realization of random vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$. Let vector $X$ follow $p$-dimensional normal distribution with unit covariance matrix $I$ and unknown mean vector $\mu = (\mu_1, \dots, \mu_p) \in \mathbb{R}^p$. The task is to estimate unknown vector $\mu$.

One estimator of $\mu$ is called the maximum likelihood estimator (MLE). We can say it is a classic estimator because it means that we estimate $\mu$ by realization $x$ of $X$. This estimator is in the form

$$\hat{\mu}^{\mathrm{MLE}} = X$$

Notation is in the scalar form, i.e. each component $\mu_i$ of vector $\mu$ is estimated using the value of the respective component $X_i$ of vector $X$, $i = 1, \dots, p$.

Another $\mu$ estimator in which we are particularly interested is called the James-Stein estimator (JSE). This estimator is in the form

$$\hat{\mu}^{\mathrm{JSE}} = (1 - \theta_p(X))X$$

where

$$\theta_p(X) = \frac{p-2}{\|X\|^2}, \qquad \|X\|^2 = \sum_{i=1}^{p} X_i^2.$$

For $p > 2$ we can see that the shrinkage factor $\theta_p(X)$ is positive because the numerator and denominator are always positive.

The reason JSE is called the shrinkage estimator is that we implicitly suppose that $\theta_p(X) \in (0, 1)$. In this case each component $X_i$ of vector $X$ from MLE is proportionally shrunken by the same constant $1 - \theta_p(X)$ closer to respective zero value component of vector $X$.

It can also be observed that in MLE we use "pure" information represented by the individual component while in JSE we borrow information from the individual component together with information contained in all the components.

## 2.3 MSE of JSE

Introducing the shrinkage factor $\theta_p(X)$ results in a lower MSE. More precisely, we have some kind of norm for MSE because we are dealing with estimation of an unknown $p$-dimensional parameter $\mu$. In [8] Stein uses

$$\mathrm{MSE}_p(\hat{\mu}) = \mathrm{E}(\hat{\mu} - \mu)^T(\hat{\mu} - \mu) \qquad (3)$$

where $\mu$ is the true value of the mean and $\hat{\mu}$ is the estimator of $\mu$ based on observed data. Expression (3) for $\mathrm{MSE}_p$ can be further rewritten as

$$\mathrm{MSE}_p(\hat{\mu}) = \mathrm{E}\left(\sum_{i=1}^{p}(\hat{\mu}_i - \mu_i)^2\right). \qquad (4)$$

Thus, MSE in case of estimation for $p$-dimensional parameter $\mu$ is the sum of MSEs for individual components of $\mu$. According to the expression (2) we can rewrite (4) as

$$\mathrm{MSE}_p(\hat{\mu}) = \sum_{i=1}^{p}\left\{\mathrm{var}(\hat{\mu}_i) + (\mathrm{Bias}(\hat{\mu}_i, \mu_i))^2\right\}. \qquad (5)$$

Thus, the MSE of estimator $\hat{\mu}$ equals the sum of variances and squared biases of its individual components. Let's demonstrate the computation of the MSE for MLE and JSE.

1) If $\hat{\mu}$ is MLE then

$$\mathrm{Bias}(\hat{\mu}_i^{\mathrm{MLE}}, \mu_i) = \mathrm{E}(X_i - \mu_i) = 0$$

where we derive benefit from $\mathrm{E}X_i = \mu_i$. Thus, the MLE is unbiased estimator. Similarly, we have

$$\mathrm{var}(\hat{\mu}_i^{\mathrm{MLE}}) = \mathrm{var}(X_i) = 1.$$

This implies according to (5) that $\mathrm{MSE}_p(\hat{\mu}^{\mathrm{MLE}}) = p$.

2) If $\hat{\mu}$ is JSE then

$$\begin{aligned}\mathrm{Bias}(\hat{\mu}_i^{\mathrm{JSE}}, \mu_i) &= \mathrm{E}((1 - \theta_p(X))X_i - \mu_i) \\ &= -\mathrm{E}(\theta_p(X)X_i) \neq 0\end{aligned}$$

where we use $\mathrm{E}X_i = \mu_i$. Thus, JSE is biased estimator in contrast to the unbiased MLE.

Computation of the MSE for JSE is not as straightforward as for the MLE. So we restrict ourselves to stating that

$$\mathrm{MSE}_p(\hat{\mu}^{\mathrm{JSE}}) = p - (p-2)^2\mathrm{E}\left(\frac{1}{\|X\|^2}\right).$$

We can see that for $p > 2$

$$\mathrm{MSE}_p(\hat{\mu}^{\mathrm{JSE}}) < \mathrm{MSE}_p(\hat{\mu}^{\mathrm{MLE}})$$

because $\|X\|^2 > 0$. Especially, for large $p$ we can achieve a great improvement of MSE for JSE in contrast to MSE for MLE. However, the "price" we pay for lower MSE of JSE is bias.

## 2.4 The Optimal Shrinkage Factor

We can also compute the optimal shrinkage factor which guarantees the lowest possible MSE. This is given by maximizing each summand in $\mathrm{MSE}_p(\hat{\mu}^{\mathrm{JSE}})$. If the shrinkage factor $\theta_p(X)$ in $\hat{\mu}^{\mathrm{JSE}}$ is denoted as $\varphi$ then according to (5) we have

$$\sum_{i=1}^{p}\{(1-\varphi)^2\mathrm{var}(X_i) + (\mathrm{E}((1-\varphi)X_i - \mu_i))^2\}$$

which is maximized by taking derivative with respect to shrinkage factor $\varphi$. This leads to the optimal shrinkage factor

$$\varphi^\star = \left(\sum_{i=1}^{p}\mathrm{var}(X_i)\right) \Big/ \left(\sum_{i=1}^{p}\mathrm{E}X_i^2\right)$$

where we use $\mathrm{var}(X_i) + (\mathrm{E}X_i)^2 = \mathrm{E}X_i^2$.

 ©2013 EuroMISE s.r.o.

### 2.5 The General Form of JSE

We have introduced only the basic version of JSE. If we again suppose $p$-dimensional normality for random vector $X$ as in the subsection 2.2, the general form of JSE can be written as

$$\hat{\theta}^{\text{shrink}} = (1-\lambda)\hat{\theta} + \lambda\hat{\theta}^{\text{target}},$$

i.e. the shrinkage estimator $\hat{\theta}^{\text{shrink}}$ is the weighted average of the classic estimator $\hat{\theta}$ and the target estimator $\hat{\theta}^{\text{target}}$ chosen by us. In case of basic JSE $\hat{\theta}$ is $\hat{\mu}^{\text{MLE}}$ and $\hat{\theta}^{\text{target}}$ is the vector of $p$ zeros.

The constant $\lambda \in (0,1)$ is the shrinkage factor, i.e. the weight which is borrowed from the classic estimator and obtained by target estimator. In case of the basic JSE we have $\lambda = \theta_p(X)$.

The estimator $\hat{\theta}^{\text{shrink}}$ is useful because it results in a lower MSE previously represented by (5).

We can compute the optimal shrinkage factor which minimizes the MSE. In case of unbiased classic estimator and nonrandom target estimator we have the optimal shrinkage factor with respect to the MSE in the form

$$\lambda^{\star} = \left\{\sum_{i=1}^{p} \text{var}(\hat{\theta}_i)\right\} \Big/ \left\{\sum_{i=1}^{p} \text{E}(\hat{\theta}_i - \hat{\theta}_i^{\text{target}})^2\right\}. \qquad (6)$$

For the optimal shrinkage factor where the target estimator is random and the classic estimator is biased we refer the reader to [9].

### 2.6 Several Generalizations of JSE

In [10] James and Stein show that JSE has a lower MSE for an arbitrary constant $c \in (0, 2(p-2))$ which is used in the nominator of $\theta_p(X)$. In [11] Baranchik observes that the shrinkage parameter $1-\theta_p(X)$ can be negative and considers positive part of JSE for improving the MSE, i.e.

$$\hat{\mu}^{\text{JSE+}} = (1-\theta_p(X))^{+}X \qquad (7)$$

where $t^{+} = t$ for $t \geq 0$ and $t^{+} = 0$ in other cases.

In [12] Bock expands JSE of $\mu$ for a situation when vector $X$ from $p$-dimensional normal distribution has an arbitrary known or unknown positive definite covariance matrix $V$. According to (7), general JSE can be written in the form

$$\hat{\mu}^{\text{JSE+},V} = \left(1 - \frac{\hat{p}-2}{X^T V^{-1} X}\right)^{+} X \qquad (8)$$

where $V^{-1}$ is inverse of the covariance matrix $V$ and $\hat{p}$ is effective dimension given by the trace of matrix $V$ divided by the maximum eigenvalue of $V$. For $\hat{p} > 2$ the JSE has a lower MSE than the MLE. Bock also shows that substituting of $\hat{p}-2$ in the nominator of the fraction in (8) by an arbitrary constant $\hat{c} \in (0, 2(\hat{p}-2))$ leads to a lower MSE for the JSE than for MLE. For more information about the generalization of the JSE we refer the reader to [7].

## 3 Applications of the Shrinkage Approach

In the subsection 1.2 we described several types of hypotheses in the context of microarrays. In this section we present shrinkage approach as a solution for some of them.

We desribe shrinkage version of the clustering algorithm $K$-means method corresponding to the hypothesis (c) or (d), shrinkage version of the t-statistic corresponding to the hypothesis (a), shrinkage version of the mutual estimation corresponding to the hypothesis (e) and shrinkage version of the covariance matrix corresponding to the hypothesis (b) or (e).

### 3.1 Shrinkage $K$-Means Method

In [13] Gao and Hitchcock introduce a shrinkage version for $K$-means clustering algorithm as an improvement of this when $n \ll p$. The method is applied to Saccharomyces cerevisiae yeast gene expression data. The data contain 78 genes, where each gene is supposed to be differentially expressed in exactly one of 5 groups (5 cell cycle phases). The expression of each gene is measured 18 times at 7-minute intervals.

The shrinkage $K$-means algorithm proceeds as follows. We have $n$ observations divided into $K$ groups, where $K < n$. Each observation has $p$-dimensional normal distribution with mean vector $\mu_i$ and covariance matrix $V_i$, $i = 1, \ldots, K$. We choose randomly $K$ observations which serve as initial estimates for $\mu_i$'s, i.e. group centroids. We compute the overall centroid $\overline{X}$ as the overall mean from all group centroids $\overline{X}_i$. Each centroid $\overline{X}_i$ is then shrunken to the overall centroid $\overline{X}$ as

$$\overline{X}_i^{\text{JSE+},V} = \overline{X} + (1-\theta(\hat{p}, V_i))^{+}(\overline{X}_i - \overline{X})$$

where

$$\theta(\hat{p}, V_i) = \frac{\hat{p}-2}{(\overline{X}_i - \overline{X})^T V_i^{-1} (\overline{X}_i - \overline{X})}$$

and $\hat{p}$ is the effective dimension given similarly as in (8) as the trace of matrix $V_i$ divided by the maximum eigenvalue of $V_i$.

In comparison to the classic $K$-means algorithm the shrinkage $K$-means algorithm has better accuracy as given by the Rand Index. The Rand Index measures concordance between the true underlying clustering structure

©2013 EuroMISE s.r.o. 

and the result produced by a clustering algorithm. For more details we refer the reader to [13].

## 3.2 Shrinkage t-Statistic

In [14] Opgen-Rhein and Strimmer introduce a shrinkage version of t-statistics in case of $n \ll p$. Shrinkage is applied to empirical variances $\nu_1, \ldots, \nu_p$ from gene expressions for each of $p$ genes. Then the median $\nu_{\text{median}}$ from all empirical variances is computed. The shrinkage estimator for $\nu_k$ is proposed in the form

$$\nu_k^\star = (1-\lambda)\nu_k + \lambda\nu_{\text{median}}, \tag{9}$$

$k = 1, \ldots, p$, which is the weighted average of the target estimator for variance ($\nu_{\text{median}}$) and the classic estimator for variance ($\nu_k$). The optimal shrinkage parameter $\hat{\lambda}^\star$ with respect to the MSE is in the form

$$\hat{\lambda}^\star = \min\left(1, \frac{\sum_{k=1}^{p} \widehat{\text{var}}(\nu_k)}{\sum_{k=1}^{p}(\nu_k - \nu_{\text{median}})^2}\right). \tag{10}$$

Estimator (10) differs from estimator (6). Here $\hat{\lambda}^\star$ is composed of a minimum of 1 and the sample estimator of (6). In (6), the numerator and denominator are estimated from data from its sample counterparts. Using the minimum in (10) prevents the shrinkage parameter from "overflowing", i.e. if the estimator of (6) present in (10) is larger than 1 then $\hat{\lambda}^\star = 1$.

Shrinkage t-statistic for comparison of two independent groups of samples is in the form

$$t_k^\star = \frac{\overline{x}_{k1} - \overline{x}_{k2}}{\sqrt{\nu_{k1}^\star/n_1 + \nu_{k2}^\star/n_2}} \tag{11}$$

where $\overline{x}_{k1}$ and $\overline{x}_{k2}$ represent group averages of gene expressions for the $k$-th gene, $\nu_{k1}^\star$ and $\nu_{k2}^\star$ represent the shrunken group variances of gene expressions for $k$-th gene and $n_1$ and $n_2$ represent the number of samples in each group for the $k$-th gene.

The shrinkage t-statistic (11) is a compromise between standard t-statistics ($\lambda = 0$ in (9)) and differences of means t-statistics ($\lambda = 1$ in (9)).

The shrinkage t-statistics (11) and several competing methods, such as moderated t-statistics, Efron t-statistics and Cui t-statistics (see [15]), are performed on three gene expression data with different "setups" for variabilities of individual genes (two Affymetrix spike-in studies and one HIV study). The aim is to find the "true discovery rate" of genes, i.e. genes which are known to be truly differentially expressed among all differentially expressed genes from a statistical point of view. The shrinkage $t$-statistic has the best performance. For more details we refer the reader to [14].

## 3.3 Shrinkage Mutual Information

In [16] Hausser and Strimmer introduce a shrinkage version of mutual information in case of $n \ll p$. The method is applied for constructing association network among genes for Escherichia coli gene expression data. The data consist of 102 known differentially expressed protein coding genes of human superoxid dismutase whose expression is measured at time 0, 8, 15, 22, 45, 68, 90, 150 and 180 minutes after induction by dosage of isopropyl-beta-D-thiogalactopyranoside.

The gene association network is constructed via Algorithm for Reconstruction of Accurate Cellular NEtworks (ARACNE). The algorithm is based on mutual information computed for each pair of genes and model selection is carried out via information processing inequality applied to all gene triplets (see [17]).

Based on whole gene expression data of Escherichia coli, gene expressions for each gene are discretized into $K$ common distinct categories of expression. For each pair of discretized genes we obtain $K \times K$ contingency table.

The mutual information $MI(A, B)$ between discrete random variables $A$ and $B$ (i.e. between discretized expressions of genes) is defined as

$$MI(A, B) = \sum_{i=1}^{K}\sum_{j=1}^{K} \theta_{ij}\left(\ln(\theta_{ij}) - \ln(\theta_i\theta_j)\right) \tag{12}$$

where $\theta_{ij}$ is the joint relative frequency for the $(i, j)$-th combination of row category $i$ for the random variable $A$ and column category $j$ for the random variable $B$ in $K \times K$ contingency table. Relative frequencies $\theta_i$ and $\theta_j$ correspond to marginal relative frequency of $i$-th row category and $j$-th column category, respectively. The task is the estimation of $\theta_{ij}$, $\theta_i$ and $\theta_j$ for (12).

We restrict ourselves to estimation of joint relative frequencies $\theta_{ij}$ and especially with respect to (12) represented by the joint Shannon entropy. The joint Shannon entropy is given by

$$H(A, B) = -\sum_{i=1}^{K}\sum_{j=1}^{K} \theta_{ij}\ln(\theta_{ij}) \tag{13}$$

and measures the uncertainty associated with the discretized random variables $A$ and $B$. When $H(A, B)$ is higher, the uncertainty is also higher.

The classic estimator of $\theta_{ij}$ in (13) is MLE, i.e.

$$\hat{\theta}_{ij}^{\text{MLE}} = \frac{y_{ij}}{n}$$

where $y_{ij}$, $i = 1, \ldots, K$, $j = 1, \ldots, K$ is the absolute frequency of the $(i, j)$-th category in the $K \times K$ contingency table and $n$ is the total sum of absolute

 ©2013 EuroMISE s.r.o.

frequencies from all cells in the contingency table. The form of $\hat{\theta}_{ij}^{\mathrm{MLE}}$ is based on the assumption of multinomial distribution for cell counts in the $K \times K$ contingency table.

The disadvantage of $\hat{\theta}^{\mathrm{MLE}}$ is that it underestimates (13), leading to a biased estimator of (12). The reason is that the $K \times K$ contingency table is sparse, i.e. the majority of cell frequencies are equal to zero. If a cell with zero frequency that represents zero summand in (13).

We can also estimate relative frequencies by shrinkage estimator in the form

$$\hat{\theta}_{ij}^{\mathrm{shrink}} = (1-\lambda)\hat{\theta}_{ij}^{\mathrm{MLE}} + \lambda t_{ij}$$

where $t_{ij} > 0$ is the $(i,j)$-th term from the target distribution $\sum_{i=1}^{K}\sum_{j=1}^{K} t_{ij} = 1$. The role of target distribution is to regularize the contingency table, i.e. zero cell counts are "converted" to nonzero counts and this decreases underestimation of (13). Typically, target distribution is chosen as uniform, i.e. $t_{ij} = 1/L$ where $L$ is the number of cells in the $K \times K$ contingency table.

According to equation (6) the optimal shrinkage intensity $\hat{\lambda}^{\star}$ with respect to the MSE is given by

$$\hat{\lambda}^{\star} = \left(\sum_{i=1}^{K}\sum_{j=1}^{K} \widehat{\mathrm{var}}(\hat{\theta}_{ij}^{\mathrm{MLE}})\right) \Big/ \left(\sum_{i=1}^{K}\sum_{j=1}^{K} (t_{ij} - \hat{\theta}_{ij}^{\mathrm{MLE}})^2\right)$$

where the nominator and denominator are estimated without bias from data.

The shrinkage estimator for estimation of entropy has a performance similar to the Nemenman-Shafee-Bialek (NSB) estimator for entropy (see [18]). However, in contrast to the NSB estimator the shrinkage estimator is computationally much faster and fully analytical. For more information we refer the reader to [16].

## 3.4 The Shrinkage Covariance Matrix

In [9] Schäfer and Strimmer introduce a shrinkage estimator of the population covariance matrix $\Sigma$ in case of $n \ll p$. The method uses the same Escherichia coli data as in the previous part, related to shrinkage estimation of mutual information. In this case we want to establish the gene network among 102 preselected genes.

The construction of the gene association network is based on a $p \times p$ matrix of partial correlations for gene expression data. Partial correlation measures the strength of the relationship between genes which is free of the influence of other genes. If partial correlation is larger than a certain value (e.g. larger than 0.8) then we can suppose there is an association between the genes.

Values of partial correlations can be computed from values of the inverted covariance matrix. For computing of the inverted covariance matrix $\Sigma^{-1}$ we need an estimator of the population covariance matrix $\Sigma$. Two classic estimators for $\Sigma$ are MLE and unbiased estimator. In other words, the elements $\sigma_{ij}$ of $\Sigma$ are estimated by elements $s_{ij}$ from the sample covariance matrix $S$ where

$$s_{ij} = \frac{1}{\mathrm{df}} \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j).$$

Here $x_{ik}$ is the expression of the $i$-th gene in the $k$-th sample, $x_{jk}$ is the expression of the $j$-th gene in the $k$-th sample, $\overline{x}_i$ is the average expression of the $i$-th gene across all samples and $\overline{x}_j$ is the average expression of the $j$-th gene across all samples, $i = 1,\ldots,p$, $j = 1,\ldots,p$, $k = 1,\ldots,n$. If $\mathrm{df} = n-1$ we have an unbiased estimator $S$ for $\Sigma$. In case of $\mathrm{df} = n$ we have MLE for $\Sigma$.

We restrict our attention to the unbiased estimator $S$. The problem of the estimator $S$ of the covariance matrix is that it becomes singular in case of $n \ll p$ as shown in [19]. Thus, it is not possible to make its sample inversion $S^{-1}$.

The singularity of matrix $S$ can be eliminated by an estimator based on JSE. The elements $\sigma_{ij}$ of $\Sigma$ are estimated by sample covariance matrix with the elements

$$s_{ij}^{\star} = (1-\lambda)s_{ij} + \lambda t_{ij} \tag{14}$$

where $i = 1,\ldots,p$, $j = 1,\ldots,p$. In equation (14), $s_{ij}$ is the unbiased estimator of $\sigma_{ij}$, $t_{ij}$ is an element of target matrix $T$ which is regular and of the same dimension as matrix $S$. The shrinkage constant $\lambda$ is supposed to be derived from interval $(0,1)$. An advantage of introducing variant of JSE is not only that it results in a lower MSE but also that it leads to regularization of the unbiased estimator of sample covariance matrix $S$.

The optimal shrinkage intensity $\hat{\lambda}^{\star}$ for the nonrandom target matrix $T = (t_{ij})$ and the singular unbiased estimate of covariance matrix $S = (s_{ij})$ is

$$\hat{\lambda}^{\star} = \left(\sum_{i=1}^{p}\sum_{j=1}^{p} \widehat{\mathrm{var}}(s_{ij})\right) \Big/ \left(\sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{E}(s_{ij} - t_{ij})^2\right)$$

where the nominator and denominator are estimated without bias from the data. Schäfer and Strimmer also examine several types of shrinkage targets. They especially pay attention to diagonal covariance target matrix with unequal variances computed from estimate $S$, i.e. $t_{ij} = s_{ij}$ for $i = j$ otherwise $t_{ij} = 0$. This represents a compromise between simple and complicated estimates. For more information we refer the reader to [9].

In [6] Guo et al. use a shrinkage estimator of covariance matrix for LDA in case of $n \ll p$. The LDA regularized in this way is then combined with the nearest shrunken centroids method (see [20]). Performance of regularized discriminant analysis is tested on nine gene expression data and has for example similar performance as the support vector machines method (see [21]).

## 4 Conclusion

In this paper we present the shrinkage approach. This is a promising approach for improving gene expression data analysis where the number of genes is much higher than the number of samples.

The shrinkage approach leads to the shrinkage estimator, which combines information from the classic estimator and a specified target estimator through a weighted average of these. The advantage of the shrinkage estimator is a lower MSE than for a classic estimator. The shrinkage estimator is biased but has a substantially lower variability than the classic estimator which is unbiased. This is valid especially for high-dimensional problems.

The shrinkage approach is applied to $K$-means algorithm, two-sample $t$-test, estimation of mutual information and estimation of covariance matrix. We can see that the shrinkage estimator is reasonably simple and provides a certain type of regularity. Regularity is in the sense of remedy of covariance matrix (from singular matrix to regular matrix) or sparsity of contingency table (from a higher underestimated value of the true entropy to a less underestimated value of the true entropy).

### Acknowledgements

I would like to thank my supervisor Zdeněk Valenta for his productive advice about writing this text. The paper has been supported by the SVV-2013-266517 project of Charles University in Prague.

## References

[1] Pirrung MC. How to Make a DNA Chip. Angewandte Chemie International Edition. 2002 Apr; 41(8):1276-1289

[2] Quackenbush J. Microarray Data Normalization and Transformation. Nature Genetics Supplement. 2002 Dec; 32:496-501

[3] Smyth GK, Yang YH, Speed T. Statistical Issues in cDNA Microarray Data Analysis. Methods in Molecular Biology. 2003; 224:111-136

[4] Göhlmann H, Talloen W. Gene Expression Studies Using Affymetrix Microarrays. Boca Raton: Chapman & Hall / CRC; 2009

[5] Cui X, Churchill GA. Statistical Tests for Differential Expression in cDNA Microarray Experiments. Genome Biology 2003; 4(4):Article210

[6] Guo Y, Hastie T, Tibshirani T. Regularized Discriminant Analysis and Its Application in Microarray. Biostatistics. 2007 Jan; 8(1):86-100

[7] Lehmann EL, Casella G. Theory of Point Estimation, 2nd ed. New York: Springer-Verlag; 1998

[8] Stein C. Inadmissibility of the Usual Estimator for the Mean of Multivariate Normal Distribution. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability. 1956; 1:197-206

[9] Schäfer J, Strimmer K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology. 2005; 4(1):Article32

[10] James W, Stein C. Estimation with Quadratic Loss. Proceedings of the Fourts Berkeley Symposium on Mathematical Statistics and Probability. 1961; 1:361-379

[11] Baranchik AJ. Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution. Stanford (CA): Department of Statistics, Stanford University. 1964 May. 39 p. Report No.:51

[12] Bock M. Minimax Estimators of the Mean of a Multivariate Normal Distribution. Annals of Statistics. 1975; 3(1):209-218

[13] Gao J, Hitchcock DB. James-Stein Shrinkage to Improve k-Means Cluster Analysis. Computational Statistics and Data Analysis. 2010 Sep; 54(9):2113-2127

[14] Opgen-Rhein R, Strimmer K. Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. Statistical Applications in Genetics and Molecular Biology. 2007; 6(1):Article9

[15] Cui X, Hwang G, Qiu J, Blaxes NJ, Churchill GA. Improved Statistical Test for Differential Gene Expression by Shrinking Variance Components. Biostatistics. 2005 Jan; 6(1):59-75

[16] Hausser J, Strimmer K. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. Journal of Machine Learning Research. 2009 Jul; 10:1469-1484

[17] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics. 2006 Mar; 7(Supplement 1):S7

[18] Nemenman I, Shafee F, Bialek W. Entropy and Inference, Revisited. Advances in Neural Information Processing Systems, 14:471-478

[19] Kwan CCY. An Introduction to Shrinkage Estimation of the Covariance Matrix: A Pedagogic Illustration. Spreadsheets in Education [Internet]. 2011 [cited 2013 Aug 1]; 4(6):Article6. Available from: http://epublications.bond.edu.au/ejsie/vol4/iss3/6/

[20] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(10):6567-6572

[21] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag; 2009


©2013 EuroMISE s.r.o.

# Towards a System of Enhanced Transparency of Medical Curriculum

**Martin Komenda**[1], **Daniel Schwarz**[1], **Ladislav Dušek**[1]

[1] Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

## Abstract

**Introduction:** A correctly compiled and balanced curriculum across medical fields is an essential prerequisite for medical education. Just as in other fields, it is vividly evident in medicine that the overall structure and lesson content are not ideal and it is quite common for the overlap between theoretical and clinical subjects to be either too extensive or, on the contrary, rather insufficient. The issue of curriculum innovation has been addressed in many medical and health care fields by different academic institutions, as proclaimed in the analysis of the current global situation. The need for a computer-aided management system providing a transparent overview of the curriculum at specific faculties has been duly emphasized.
**Aim:** An original methodology for managing the changes during optimization of the medical curriculum in tertiary education is introduced in the paper, together with its conceptual data model.
**Methods:** According to the conclusions of the Bologna Process, which promotes outcome-based education, the theoretical basis is oriented to the new paradigm of teaching and learning.
**Results:** We introduce the all-in-one web-oriented platform covering all elements linked to global curriculum management and providing sufficient information for the curriculum designers.
**Conclusion:** A new powerful and robust system for the management, visualization and analysis of curriculum should pinpoint specific learning imperfections and potential overlaps across the chosen field of study. It brings useful and beneficial information for students, teachers as well as the faculty management.

## Keywords

Medical curriculum, outcome-based approach, computer-aided management system.

**Correspondence to:**

**Martin Komenda**
Institute of Biostatistics and Analyses, Masaryk University
Address: Kamenice 126/3, Brno, Czech Republic
E–mail: komenda@iba.muni.cz

**EJBI 2013; 9(3):9–16**
recieved: July 15, 2013
accepted: October 31, 2013
published: November 20, 2013

## 1 Introduction

Medical students face specific challenges compared to university students of other disciplines. The reason is that their future job does not allow any gaps in the knowledge obtained during the studies and that any error in their medical practice may have fatal consequences. The need for guaranteed and high-quality education involving predefined curricula covering the corresponding scope of input knowledge and skills required in subsequent practice has been gaining momentum. Medical universities compile their curricula so as to ensure that they cover all steps essential for the students to obtain employment later on. The students must fulfil all duties such as the successful completion of compulsory and optional courses and the final state exam, only to prepare for attestation to obtain professional qualification for employment as a physician.

A correctly compiled and balanced curriculum across medical fields is an essential prerequisite for medical doctor education. A suitable combination of theoretically focused courses and a clinical teaching base is certainly the key to a successful draft curriculum.

As in other fields and faculties, it is also evident in medicine that the overview of the lesson structure and content is not ideal and it often happens that the overlap between theoretical and clinical subjects is either too large or, on the contrary, rather insufficient. As modern information technologies have seen rapid development in the last decades, modern internet technologies can be utilized to develop an infrastructure or applications that would not only eliminate the poor transparency of the curricula but also help improve its construction as such.

The introduction of new approaches to learning, the exponential growth of Internet usage and the advent of

©2013 EuroMISE s.r.o. 

the World Wide Web have the potential to change the face of higher education [1]. The Bergen ministerial conference of the Bologna Process in May 2005 discussed reforms pertaining to the degree structures, credit transfer, quality assurance and curricular development, which are transforming the European Higher Education Area. The European tertiary education systems are undergoing radical restructuring in line with the objectives defined by the Bologna Process. The design and structure of new curricula constitute significant processes of change and require cooperation and coordination [2]. In the midst of these developments lies an interesting opportunity for applying new technologies providing an interface for collaborative curriculum content management, including graphical representation of the available data.

These systems will facilitate institutional decision-making activities related to the creation of a well-balanced curriculum. The entire concept is oriented to the new paradigm according to conclusions of the Bologna process, which introduces outcome-based education. This performance-based approach at the cutting edge of curriculum development offers a powerful and appealing way of reforming and managing medical education. The emphasis is on the product – what sort of doctors are we producing – rather than on the educational process [3]. Outcome-based education has emerged as a priority for curriculum planners striving to align with the ever demanding societal needs. The CanMEDS initiative of The Royal College of Physicians and Surgeons of Canada has introduced the implementation of a national, needs-based, outcome-oriented, competency framework that sets out the knowledge, skills and abilities specialist physicians need for better patient outcomes. The framework is based on the seven roles that all physicians need to have, to be better doctors: Medical Expert, Communicator, Collaborator, Manager, Health Advocate, Scholar, and Professional. This multifaceted approach has enabled a large-scale curriculum change for outcome-based education [4, 5].

### 1.1 Outcome-Based Approach

Learning outcomes are arguably best viewed as a fundamental building block of the Bologna educational reforms and bring more transparency to higher education systems. They have the reputation of rather mundane and prosaic tools, yet it is this basic underpinning function that makes them so significant. It is important that there should be no confusion about their role, nature and significance, or the educational foundations of the Bologna process will be undermined. Learning outcomes have applications at three distinct levels: (i) the local level of the individual higher education institution (for course units/modules, programmes of study and qualifications); (ii) the national level (for qualifications frameworks and quality assurance regimes); and (iii) internationally (for wider recognition and transparency purposes). Learning outcomes and "outcome-based approaches" have a strong impact on curriculum design, teaching, learning and assessment, as well as quality assurance [2].

The use of learning outcomes implies a fundamental paradigm shift in curriculum design for many European institutions offering higher education. The adoption of learning outcomes means a shift of emphasis from the content (what staff teach) to outcome (what a learner is expected to know, understand and/or be able to do), or expressed in other words, a clear shift from a teacher-centred to a learner-centred perspective. These issues are closely related to the teaching and learning methods as well as student workload. Defining the learning outcomes (instead of teaching inputs) promotes the idea of the teacher as a facilitator of learning and recognizes that a great part of learning takes place outside the classroom.

It promotes the idea of learners who are actively involved in the planning and management of their own learning activities and who assume more responsibility for it [6]. S. Adam defined in [7] that a learning outcome is a written statement of what the successful student/learner is expected to be able to do at the end of the module/course unit, or qualification. The key aspect that each of the definitions has in common is the desire for more precision and consideration as to what exactly a learner acquires in terms of knowledge and/or skills when he or she successfully completes the period of learning. Modern information and telecommunications technologies offer the opportunity to revolutionize the way we provide education [8]. When using ICTs (Information and Communication Technologies) for education, the temptation is often to focus more on the technology and less on the learners and instructors, often to the detriment of the education quality [9]. The combination of appropriate computer technologies and methodical approach (e.g. outcome-based approach) can fundamentally improve the entire education process.

## 2 ICT Tools for Curriculum Management

A recommended curriculum could be identified through content analysis of curriculum documents and interviews with individuals responsible for teaching [10]. The question is how can one make the whole process of curriculum harmonization easier and more effective? The quantity of electronic education content in the today's information society is steadily increasing. Thus, academic institutions require new innovative web-based tools for user-friendly creation, effective organization and clear visualization of outcome-based curriculum data. The true power of visualization tools is that they can provide a broader view of the whole curriculum. Apart from the aforementioned functionalities, the key and obviously necessary task rests in appropriate data visualization for easier understanding and further analysis. The term visualization represents a set of technical means and methods allowing clear and unambiguous graphical presentation of

 ©2013 EuroMISE s.r.o.

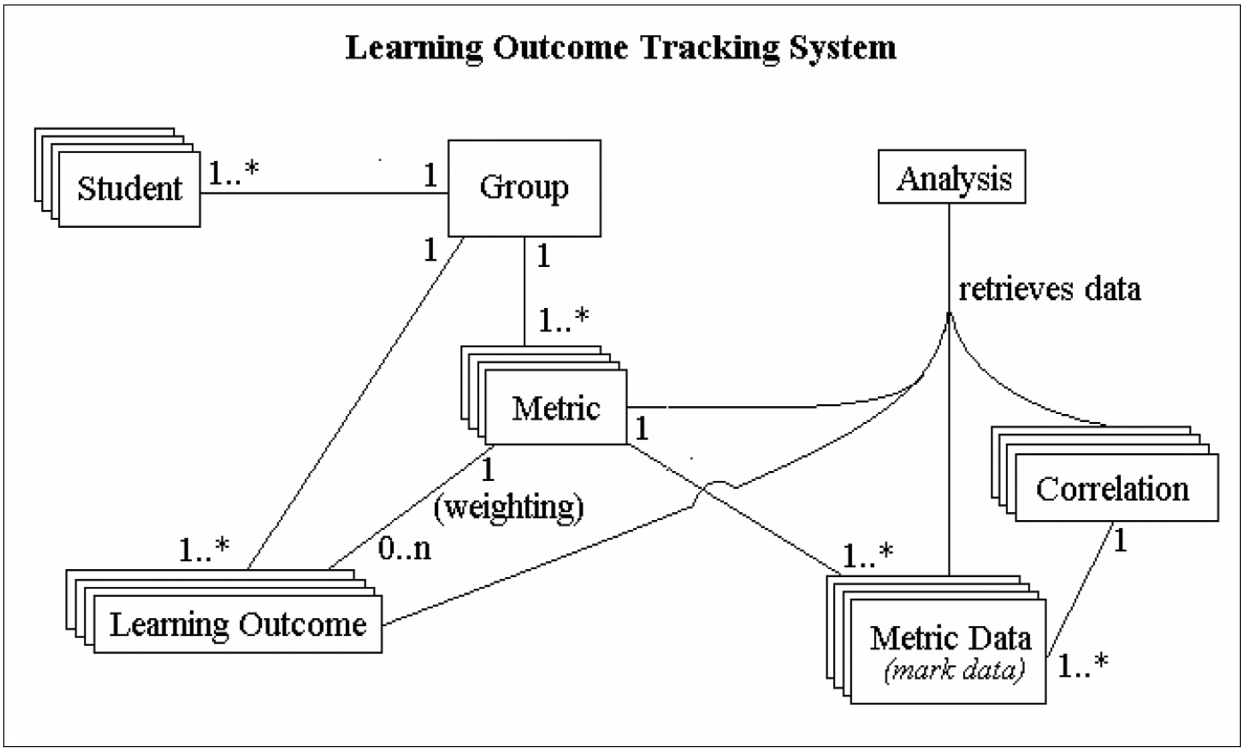


Figure 1: Relationship among the components for a single group instance in LOTS [13].

the data. The data stands in hidden relationships, its identification being the principal task of visual analysis that can be made with the integration of computer-based interactive tools and techniques based on cognitive, perceptual design principles [11].

There has been limited discussion on the original information technologies supporting outcome-based ideas. There are only a few available and published solutions, which have been developed for making educational outcome-based data accessible to the target group – students and teachers. In [12] A. Huang has presented an integrated outcome assessment application, for instance the design of a database to accumulate learner performance output by means of an outcome assessment application and storing it as learner profiles. These profiles could then serve as valuable input to produce customized learning content or to conduct overall performance evaluation. Y. Mong et al. [13] have described the web-based application LOTS (Learning Outcome Tracking System), which provides overall management of the learning outcomes and access for both the student and the teachers. In brief, LOTS consists of six components, namely group, metric, learning outcome, incident, correlation and analysis.

Fig. 1 illustrates the relationship among the components for a single group instance. The group includes student and group administration, where users can manage students and groups in the database. A group can be a course, a degree programme, or a subset of students. A great benefit is the integration of LOTS with an existing LMS (Learning Management System) environment. LOTS includes also several types of basic analysis with graphical output presentations (typically bar charts) for users to exploit the advantages of outcome-based education.

The “generic” electronic portfolio called ePortfolio [14] is an application, which is being used to support the evidencing of learning outcomes and to facilitate personal development planning. In modular courses, portfolios may provide focus on programme-level as well as module-specific learning outcomes. The whole process may help students become better at relating what they have learned to the requirements of teachers. The ePortfolio has been developed using robust platform-independent open source tools. In the design phase, some of the potential “value-added” features, which an IT approach can bring compared to paper-based portfolios, were considered, i.e. being highly customizable with multiple structures and views, easier cross-referencing, being searchable, integration with virtual learning environments, etc. A number of generic tools have been developed for supporting personal development planning, i.e. a reflective learning diary, outcomes/skills log, records of meetings with tutors, SWOT matrix and an Action Planning tool. The ePortfolio framework also allows sharing of specific content with supervisors, peers and others, with the facility for viewers to add formative comments.

Dynamic Learning Maps (DLM) has provided a novel way to navigate and engage with the curriculum and support learning. Curriculum drivers included the long-standing need for maps to aid the understanding of complex curricula in Medicine, and for promoting “cross-

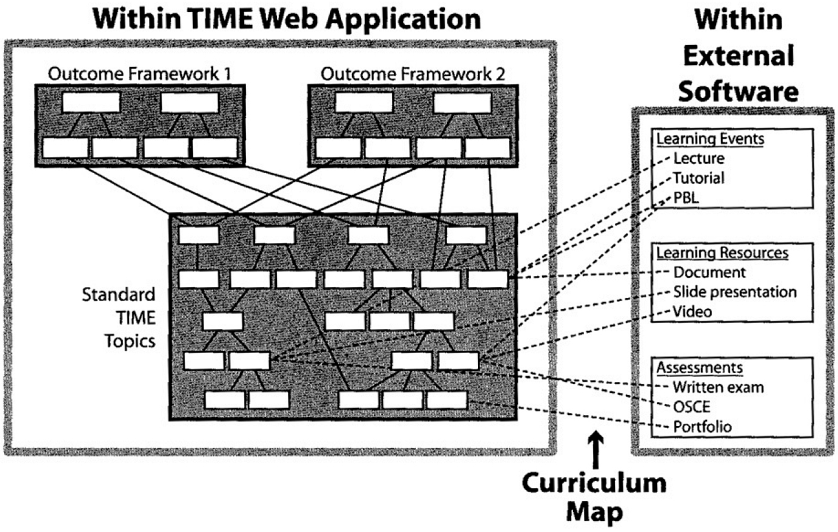


Figure 2: TIME includes Outcomes, Topics and Objectives [19].

modular" learning in modular programmes. DLM has been designed to be flexible and extensible and can be used for a range of different pedagogic/educational purposes. This web-based tool utilizes existing technologies and standards to provide access to information from curriculum databases, ePortfolios and other established sources. This approach provides a unique resource for learners, teachers and curriculum managers alike by providing a detailed and convenient way of accessing the formal and informal curriculum. DLM is an interactive tool, which can be aligned with an institution's Teaching and Learning Strategy, while at the same time it supports a wide range of requirements for specific programmes. There are two main views: "tree view" (hierarchical text) and "MindMap". For example, the maps can display and make links between granular curriculum information and overarching transferable skills frameworks and/or programme-level outcomes. This novel concept has been well received by learners, teachers and curriculum staff and has had considerable impact both at the Newcastle University and within the broader CETIS community (the Centre for Educational Technology and Interoperability Standards). DLMs took a rapid development/action research style approach, with emphasis on formative evaluation involving substantial input from students, teachers and other stakeholders. Open-source software was used for DLM development, in particular the Python-based rapid development framework Django, MySQL databases, jQuery and other Javascript libraries and the Freemind Flash browser [15, 16].

S. Kabicher et al have presented a sophisticated approach, the use of visual modelling within an interactive online environment (ActiveCC Web) for collaborative design, implementation and visualization of the curriculum structure and content. ActiveCC, an abbreviation for Active Curriculum for Computer Science, is an e-learning project at the Faculty of Computer Science, University of Vienna. The project aims to provide an insight into the structure and implementation of the new computer science curriculum to support the teaching staff in coordinating their course contents and to support the instructors at the faculty in obtaining an overview of the curriculum structure. It was technically based on the Cooperative Environment Web Services platform, a web service-oriented architecture for cooperation and learning, which offers a wiki module including the functionality of directly writing graph visualization code into the wiki. Using the graph visualization tool, the temporal arrangement of modules and dependency links among modules are visually modelled to show the modules' location and role within the curriculum. This approach offers an easily accessible and intuitively editable virtual space to facilitate collaboration on the curriculum content, provides a transparent view of the curriculum's structure and an insight into the module and course implementations, content and pedagogical methods [17, 18].

One of the options for describing content related to the curriculum is special taxonomy. T. G. Willett et al. have published a paper introducing TIME (Topics for Indexing Medical Education), a hierarchical taxonomy of topics relevant to medical education. The content and structure of the topics within TIME was developed in consultation with medical educators and librarians at several Canadian medical schools. As a web application, TIME can contain three types of items that are distinct but related to each other in hierarchical manner. It can contain Outcomes, which we operationally define as culminating demonstrations of learning at the end of medical training; Topics, which describe the content (subjects) of medical education; and Objectives, which we define as the more specific expectations associated with the courses or learning opportunities.

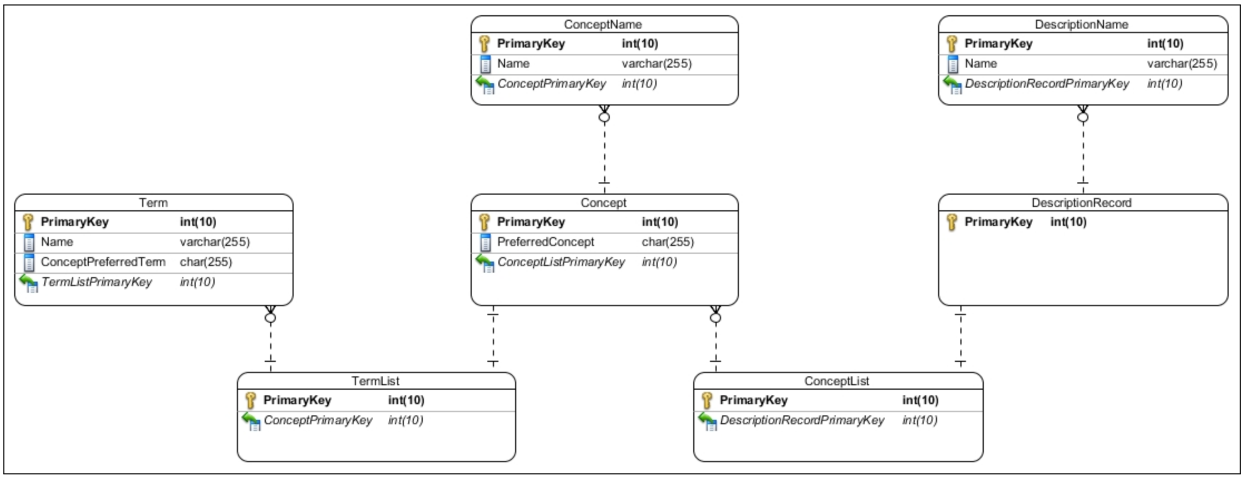


Figure 3: Conceptual data model covers all essential attributes for curriculum description using MeSH thesaurus.

The TIME is primarily intended to be a generic index for medical education. It is meant to be applicable to curriculum maps, as metadata for tagging reusable learning objects, as a classification for assessment objects, as a means of linking Objectives to Outcomes for a medical school, or potentially for any other indexing purpose in undergraduate medical education. The application itself does not enable mapping or indexing. This simply web-oriented tool allows the user to browse topics, create new outcomes and objectives, and link these items to the topics. The entire TIME content and structure can then be exported, via XML (eXtensible Markup Language), to external applications and used as an index for curriculum mapping [19]. Today, unfortunately, the TIME application is not available and cannot be viewed via the recommended web address and follow-up development was probably terminated.

## 3 Computer-Aided Management of Curriculum: Design

The idea of connection topics via the TIME web application is not the most suitable solution for complex and easy optimization of a curriculum, but it brings new inspiration for further development. Most importantly, we need an all-in-one web-oriented service, which covers all phases of the optimization process. This means an interface for data creation and editing, a user-friendly tool for transparent browsing supporting fast and accurate search, and a technology for graphical visualization of the curriculum relations. As stated above, there does not appear today to be any solution for how to parametrically describe, effectively manage and visualize clearly and in-depth the curriculum and all related education metadata in one system. This is the primary motivation for developing a new original technology for optimizing a curriculum, which supports an outcome-based approach. A similar solution based on the aforementioned parametric description of curriculum, including all appropriate metadata details (i.e. attributes of learning outcomes, units and objects), has not been widely seen.

Existing solutions that have been published are focused on the curriculum only from a certain perspective, offering the agenda together with selected functionalities and making the effort to provide them to students and teachers of the respective institution in a transparent format. However, we have yet to see a complex instrument that would cover all elements associated with global curriculum optimization, including a detailed parametric description down to the level of the learning units, and that would be linked to the learning objects. The main benefit of the development is a new concept resting on the following functionalities – a concept that has not yet been implemented in similar solutions.

- A proposal of an original medical curriculum optimization methodology across clinical and theoretical fields based on elaborate curriculum parameterization by using the outcome-based paradigm [3, 7, 13].
- A proposal of a formal database metadata arrangement describing the general medical curriculum independent of subsequent implementation.
- Integration of the Medical Subject Headings (MeSH) thesaurus [20] for standardized work with key words. As a specific field taught in the Czech language shall be optimized, a dictionary was selected that is standardized in English and offers Czech translation with updates on annual basis.
- A proposal of an elaborate access right system using a central verification mechanism through the Shibboleth technology [21] to allow potential future extension of access to the user level among academic community members.

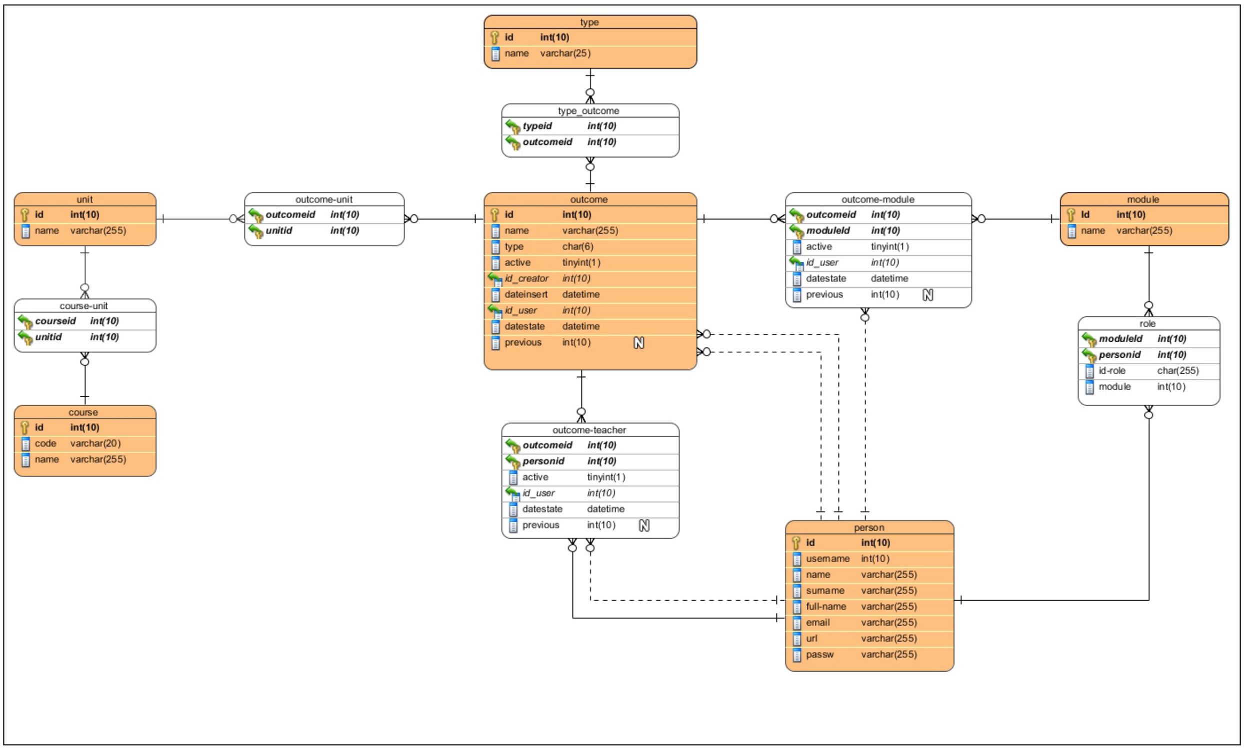


Figure 4: Entity relation data model represents key attributes of the proposed platform.

# 4 Results

The development of a new curriculum optimization technology and general methodology description are among the key attributes of our effort. The outcome-based approach opens the possibility of redefining the curriculum structure and properties in the form of a parametric description. The learning outcome is a summary of the requirements for graduates from the general medicine field, i.e. the list of knowledge and skills which the student should have upon completion of his/her studies, and signals the information essential for eventual work as a doctor. The organization of this data and its correlation will be provided in the curriculum data model that can be implemented without any restrictions within any database technology. The new methodology describes the different phases of the optimization process by using the platform to be developed. Its objective is to make all work of the user more efficient regarding the creation, editing and control mechanisms in the form of deep content inspection.

The architecture of the designed platform comprises two sections: FrontOffice and BackOffice. FrontOffice represents the interface to be used for presenting the content to the end user. The content of this section will be freely available only to the target group of users, i.e. students and teachers. BackOffice is the second platform component to be used by administrators and editors and is intended as the interface for adding and editing static content. Another feature is the adoption of the Czech version of the MeSH thesaurus where the objective is to standardize key words relating to the metadata stored in the database. The actual key words are defined and structured in many forms and there is a growing need for their unification with respect to the international framework. The bio-medicine dictionary MeSH in the English language has been published since 1960 by the US National Library of Medicine. The Czech translation of this thesaurus is being prepared by the Czech National Medical Library, which issues annual updates. The dictionary contains 26 142 entries with over 54 000 links [20]. Following a contractual agreement with the Czech National Medical Library, MeSH will also be used for the purposes of the emerging curriculum optimization platform. The main requirements for standardized dictionary integration are regular updates of the Czech mutation, which MeSH fulfils as the only solution available. No other language mutations are foreseen at the moment, but a possible change should not bring too many complications in terms of the proposed structure.

Besides the key words defined in an internationally recognized and standardized format, the learning unit description also features the so-called essential terms for which no dictionary exists. The database of essential terms contained in the field that the students come across during the six years of their study programme and which the graduates should be familiar with, will be fed by teachers and guarantors of studies. A unique dictionary of essential terms in medical studies will emerge and offer another view of the sophisticated and vast curriculum. To ensure at least partial standardization (quasi-standardization), users will be assisted by autocompletion as they enter words. The autocompletion will display expressions corresponding with the text already typed – existing words or corresponding items already defined by past authors. After the final clean-up, which cannot be fully automated, one of the outcomes (and hence the benefit) is the still non-existent dictionary of essential expressions from the general medicine field of study.

The draft entity relation model (see Fig. 4) is the basis for possible implementation in practice. It describes, in depth, the key entities and defines the relationships between them so that all selected links can later be displayed in a transparent manner. The entire optimization process is split into several dependent phases allowing efficient mapping of the curriculum of the respective field or specialization of study. A characteristic tool is recommended for each phase to give the end user – the teacher – the possibility of intuitively creating/browsing certain content. All tools have been developed based on real requirements and are the practical manifestation of how the method can be implemented when correct IT is applied. Unlike the platform being developed, the application of this method is fully independent of the information and communication technologies and of the focus (the selection of the field of study that will be optimized).

## 5 Discussion

Since the platform will serve users as a tool to oversee their curricula, attention will be focused in the future also on more efficient and fast search agenda. A language morphology analyzer is planned to be employed in the overall framework of the designed platform. For the Czech language, services of the original Majka analyzer [22] will be used. The objective is to integrate the script that will process different input chains into root forms and store this information in a database along with a link to the related metadata. If the user enters a search expression, then all words containing the root of the search expression will be displayed thanks to this functionality. As the current full-text search does not provide any reliable method of working with word forms for the Czech language, this processing method is an interesting and desirable solution. The root forms of words will also play an important role in the implementation of analytical and visualization tools. This domain will consist of data mining, data pre-processing (lemmatization, lexical categorization, creation of stop words list, generation of term frequency vector), data analysis (clustering, similarity measures), and visualization (graphs and networks). Additionally, selected methods of natural language processing and machine learning will be used for effective identification of information rich data relations and for simpler and easier way of understanding curriculum structure.

## 6 Conclusions

This approach is aimed to be primarily used for a global curriculum audit and it should determine the learning imperfections and potential overlaps across the chosen field of study. Building a powerful and robust system for the management, visualization and analysis of the curriculum could extend the basic learning outcome concept. What will such an approach mean for the student? It will provide clear information about what knowledge shall be obtained over the six years, what topics will be on the agenda, what fields will be covered repeatedly and what subjects are linked to the studies. For teacher's parametric description will mean an easier way of how to clearly describe their lessons and how to browse the curriculum data of all available courses at the medical faculty according to the pre-defined parameters. For the faculty management, this overview will represent a practical view of the teaching. It will provide clear and comprehensible data about who teaches what and in what scope, if an overlap is desirable, what is taught in clinical and theoretical fields and if the overall teaching pattern is correct or whether restructuring is necessary. The platform will also bring objective information used to define the new structure of the curriculum.

### Acknowledgements

Martin Komenda and Daniel Schwarz were supported from the grant project "MEFANET clinical reasoning" reg. n.: CZ.1.07/2.2.00/28.0038, which is funded by the European Social Fund and the state budget of the Czech Republic.

## References

[1] R. M. Harden and I. R. Hart, "An international virtual medical school (IVIMEDS): the future for medical education?," Med. Teach., vol. 24, no. 3, pp. 261–267, May 2002.

[2] R. Keeling, "The Bologna Process and the Lisbon Research Agenda: the European Commission's expanding role in higher education discourse," Eur. J. Educ., vol. 41, no. 2, pp. 203–223, 2006.

[3] R. M. Harden, J. R. Crosby, and M. H. Davis, "AMEE Guide No. 14: Outcome-based education: Part 1?An introduction to outcome-based education," Med. Teach., vol. 21, no. 1, pp. 7–14, 1999.

[4] J. R. Frank and D. Danoff, "The CanMEDS initiative: implementing an outcomes-based framework of physician competencies," Med. Teach., vol. 29, no. 7, pp. 642–647, 2007.

[5] J. R. Frank and M. D. Bernard Langer OC, "Collaboration, communication, management, and advocacy: teaching surgeons new skills through the CanMEDS Project," World J. Surg., vol. 27, no. 8, pp. 972–978, 2003.

[6] S. Adam, "Using learning outcomes - A consideration of the nature, role, application and implications for European education of employing 'learning outcomes' at the local, national and international levels," University of Westminster, Edinburgh, 2004.

[7] S. Adam, "An introduction to learning outcomes," Consid. Nat., 2006.

[8] B. Barnes and C. Friedman, "Using technology in continuing professional development," in The continuing professional

development of physicians, vol. 2003, Chicago: The continuing professional development of physicians, pp. 205–220.

[9] J. Sargeant, "Medical education for rural areas: Opportunities and challenges for information and communications technologies," J. Postgrad. Med., vol. 51, no. 4, pp. 301–7, Dec. 2005.

[10] D. T. Stern, "Practicing What We Preach? An Analysis of the Curriculum of Values in Medical Education," Am. J. Med., vol. 104, no. 6, pp. 569–575, erven 1998.

[11] D. Randjelovic and B. Popovic, "Visual analytics tools and theirs application in social networks analysis," in Telecommunications Forum (?FOR), 2011 19th, 2011, pp. 1340 –1343.

[12] A. H. Huang, "A framework of educational technology applications and development opportunities.," J. Comput. Inf. Syst., vol. 41, no. 3, pp. 54–58, 2001.

[13] Y. Mong, M. Chan, and F. K. H. Chan, "Web-based outcome-based teaching and learning - An experience report," in Advances in Web Based Learning - Icwl 2007, vol. 4823, H. Leung, F. Li, R. Lau, and Q. Li, Eds. Berlin: Springer-Verlag Berlin, 2008, pp. 475–483.

[14] S. Cotterill, "Design, implementation and evaluation of a 'generic' ePortfolio: the Newcastle experience," in Proc. ePortfolio 2004.

[15] S. Ball, "Dynamic Learning Maps –Final Project Report." 2009.

[16] P. Horner, S. Cotterill, J. Peterson, and G. Skelly, "Dynamic Learning Maps," in Learning Forum London 2010-Internet of Subjects Forum ePortfolio-Key Competencies-Identity London, 5-7 July 2010, Savoy Place Published by EIfEL ISBN 2-9524576-9-7., p. 98.

[17] S. Kabicher and M. Derntl, "Visual Modelling for Design and Implementation of Modular Curricula," Z. Für Hochschulentwicklung, 2008.

[18] S. Kabicher, M. Derntl, and R. Motschnig-Pitrik, "ActiveCC - A Collaborative Framework for Supporting the Implementation of Active Curricula," J. Educ. Multimed. Hypermedia, vol. 18, no. 4, pp. 429–451, 2009.

[19] T. G. Willett, K. C. Marshall, M. Broudo, and M. Clarke, "TIME as a generic index for outcome-based medical education," Med. Teach., vol. 29, no. 7, pp. 655–659, Jan. 2007.

[20] "Fact SheetMedical Subject Headings (MeSH®)." [Online]. Available: http://www.nlm.nih.gov/pubs/factsheets/mesh.html. [Accessed: 12-Jul-2012].

[21] M. Needleman, "The Shibboleth Authentication/Authorization System," Ser. Rev., vol. 30, no. 3, pp. 252–253, 2004.

[22] P. Šmerk, "Fast Morphological Analysis of Czech," RASLAN 2009 Recent Adv. Slavon. Nat. Lang. Process., p. 13, 2009.

 ©2013 EuroMISE s.r.o.

# Simple Use Case Evaluation Method Determining an EHR Integration Platform Design

**Daniel Krsička**[1], **Milan Šárek**[2]

[1] First Faculty of Medicine, Charles University in Prague, Czech Republic

[2] CESNET z.s.p.o., Prague, Czech Republic

## Abstract

Integration platform is a basic technical tool for realizing an interoperable Electronic Health Record (EHR). Our goal is to interrelate the knowledge about interoperability, the functions required for an EHR system and the formalized best practises for an integration platform. An evaluation method has been developed, testing dependencies between EHR use cases and logic implemented in the integration platform has been tested on the HL7 EHR System Functional Model. A dependency has been identified and is discussed in this article.

## Keywords

Interoperability Levels, Electronic Health Record, Healthcare Information System, Integration Platform, Integration Pattern, HL7 EHR System Functional Model

**Correspondence to:**

**Daniel Krsička**
First Faculty of Medicine, Charles University in Prague
Address: Kateřinská 32, Prague 2, Czech Republic
E–mail: dkrsicka@gmail.com

**EJBI 2013; 9(3):17–21**
recieved: August 15, 2013
accepted: October 28, 2013
published: November 20, 2013

## 1 Introduction

Massive penetration of Healthcare Information Systems (HIS) and eHealth resources in general the significance of Electronic Health Record (EHR) interoperability as an ability of two or more subjects to achieve a common goal or mutually support each other to achieve the individual goals respectively (synergic effect). Theoretical value can be expresses using the Metcalf's Law as the number of all possible connections among subscribers (HIS in our case). It can be asymptotically approximated by the quadratic polynomial of n2. Nevertheless the value of integrated HISs as a whole is not growing quadratic [1]. The HIS integration is not first the establishment of connections between HIS components. It is necessary to pinpoint and follow many protocols enabling an information interchange for particular HIS components and layers. That implies the definition of interoperability level.

Well known authors define several levels of interoperability and its maturity [2, 3]. The comparison between levels defined is in Table 1.

Our motivation is based on lessons learned about the technological interoperability insufficiency as a means of massive dissemination of interoperable EHR including all needed attributes. This statement is supported by professional publications focusing mainly on EHR system content and semantics. We have published the technological interoperability view inadequacy in [5] and [6]. We have demonstrated that the higher interoperability levels cannot be assured by and based on accepted and broadly used classification into technical layers according to ISO/OSI model in ISO/IEC:7498 [7]. The process and partly the semantic interoperability do not have any technical equivalent in ISO/OSI model, so these interoperability levels cannot be procured by technical resources only.

To reach the highest interoperability level is not necessary and should not be an automatic goal for each HIS, because not all the EHR system has to implement all the possible functions.

### 1.1 Hypothesis

Let us suppose that there is a mapping, assigning for each EHR use case an interoperability level required for its realization in a comprehensive EHR system including an integration platform. Evaluating a set of EHR use cases we will get a view on interoperability levels needed and we can use this approach for analysis and design of EHR integration platforms. The benefit is a software analysis simplification and EHR integration platform design op-

©2013 EuroMISE s.r.o. 

Table 1: Comparison of interoperability levels defined by other authors.

| Levels after Bloebel | Levels after Gibbons |
|---|---|
| Process / Service | Process |
| Semantic | Semantic |
| Syntactic | Technical |
| Structural | Technical |
| Technological | Technical |

timization. Creating an integration solution among 2 or more EHR systems, the method mentioned below should support analysis and design acceleration, implementation shortening, support of early prototype creation and anticipated decrease of the number of change request, so reducing the total solution costs.

# 2 Methods

We have developed a simple method evaluating every EHR use case in various dimensions. We have evaluated all the use cases for the EHR system defined in HL7 EHR System Functional Model [11]. This model serves as a input set of testing data for the method presented here due to we do not have any input data originating from the real EHR integration platform implementation. It is very hard to find some technical solution and gain access to the business analysis outcomes due to almost all the solutions are commercial in the Czech environment.

## 2.1 Interoperability Levels and Integration Platform Logical Parts

We have also defined a mapping between interoperability levels and various integration patterns. Using this mapping we are theoretically able to define which integration patterns (EHR integration platform logical functions) are needed for particular EHR use cases set. The description of individual integration platform layers is out of the scope of this article, so we introduce the list of patterns in each layer only. More information about specific patterns can be found in [8], providing a consistent vocabulary and visual notation to describe large-scale integration solutions across many implementation technologies. Considering the purpose and the added value of each pattern, they can be divided into following groups according to the interoperability level or rather the logical level of the technological solution (integration platform):

- Access Layer Integration Patterns: Channel Adapter, Competing Consumers, Correlation Identifier, Durable Subscriber, Event-driven Consumer, Idempotent Receiver, Message, Message Channel, Message Endpoint, Message Expiration, Messaging Gateway, Polling Consumer, Selective Consumer, Service Activator, Transactional Client,
- Transport Layer Integration Patterns: Channel Purger, Composed Message Channel, Document Message, File Transfer, Guaranteed Delivery, Message Bus, Message Dispatcher, Message Filter, Message Translator, Messaging, Messaging Bridge, Point-to-Point Channel, Publish / Subscribe Channel, Recipient List, Remote Invocation, Request / Reply, Return Address, Shared Database,
- Transformation and Routing Layer Integration Patterns: Aggregator, Content Filter, Dynamic Router, Format Indicator, Message Sequence, Message Router, Resequencer, Splitter

Table 2: EHR use case evaluation criterion: Space, answering questions: "Where the information communication takes place? How distant the points of presence are?"

| Score | Description |
|---|---|
| 0 | Integration in a work team only |
| 1 | Integration in one organization, mostly in one location |
| 2 | Integration among 2 or more organizations and/or locations |

Table 3: EHR use case evaluation criterion: Time, answering questions: "When the communication takes place? How fast and often it runs?"

| Score | Description |
|---|---|
| 0 | In real time / mostly “on-line” |
| 1 | On daily bases (once or more times a day – hours) |
| 2 | One or more times in a month (days / weeks |

 ©2013 EuroMISE s.r.o.

Table 4: EHR use case evaluation criterion: Subject, answering questions: "Who is communicating? What are the subject's skills?"

| Score | Description |
|---|---|
| 0 | Actors with practically the same knowledge and / or education (physicians) |
| 1 | Actors with a similar knowledge (physician and nurse) |
| 2 | Actors with completely different knowledge (physician and patient) |

Table 5: EHR use case evaluation criterion: Object, answering questions: "What is communicated? Why runs the communication? For what purpose?"

| Score | Description |
|---|---|
| 0 | Information with common syntax (sharing data) |
| 1 | Information with common semantics (sharing information) |
| 2 | Information for a deterministic action (sharing knowledge and skills) |

- Semantic Layer Integration Patterns: Canonical Data, Command Message, Content Enricher, Content-based Router, Claim Check, Datatype Channel, Envelope Wrapper, Event Message, Invalid Message, Messaging Mapper, Normalizer, Test Message
- Business Processes Layer Integration Patterns: Control Bus, Dead Letter Channel, Detour, Message Broker, Message History, Message Store, Pipes and Filters, Process Manager, Routing Slip, Smart Proxy, Scatter / Gather, Wire Tap

### 2.2 EHR Use Cases Classification

The core of presented method is a classification of each EHR use case from 4 different points of view. Each view focuses on different concept. Due to limited space, we cannot describe the method details, including its continual evolution. So we present only an overview. Inspired also by the HL7v3 Reference Information Model [9] and the law of 5W (Who, What, Where, When and Why) [10] we have proposed following classification criterions (see Tables 2, 3, 4 and 5).

Deploying the interoperability levels defined in [2], each EHR use case can get 0 to 8 points in total (4 criterions, 0 - 2 points in each criterion). The most important factor is the sum of score, determining the interoperability level needed for the use case. In case of sum equal 2 we propose a consideration of particular criterion values. If there is leastwise one score of 2 in one criterion, the target level should be syntactic interoperability. A score evaluation overview is in Table 6.

### 2.3 Experiments – Model EHR Use Cases and Interoperability

We have applied the aforementioned method on 64 functions required in HL7 EHR-S Functional Model [11]. For definition of use cases, we have assumed that these functions have to be realized in a common hospital. This way we can anticipate actors and other factor needed for the instantiation of a function (use case definition). An example of the use cases evaluation is in Table 7.

## 3 Results

Aggregating all of the 64 experiments we can summarize that the interoperability level needed for implementing all the functions in the HL7 EHR System Functional Model [11]. It means that we have used each function from this model, transform it into the EHR Information System Use Case and have applied the method presented here. This application results into a cumulative sets of values indicating the most intensive interoperability level needed to solve in the integration platform solution design. Results in graphics can be in Figure 1, the vertical axis represents the number of incidences, the horizontal axis depicts the score attained during the method application for each Use Case.

Table 6: Total score determining the target level of interoperability.

| Score attained (sum) | Target interoperability level |
|---|---|
| 0 – 2 | Technical / Structural |
| 2 – 3 | Syntactic |
| 4 – 5 | Semantic |
| 6 – 8 | Process / Service |

©2013 EuroMISE s.r.o. 

Table 7: A sample of evaluation of use cases derived from HL7 EHR-S Functional Model, section Direct Care (DC).

| Function ID used | Space | Time | Subject | Object | Sum |
|---|---|---|---|---|---|
| DC.1.1.1 | 1 | 0 | 2 | 0 | 3 |
| DC.1.1.2 | 1 | 1 | 2 | 1 | 5 |
| DC.1.1.3.1 | 2 | 1 | 1 | 1 | 5 |
| DC.1.1.3.2 | 2 | 1 | 2 | 0 | 5 |
| DC.1.1.3.3 | 2 | 2 | 2 | 0 | 6 |

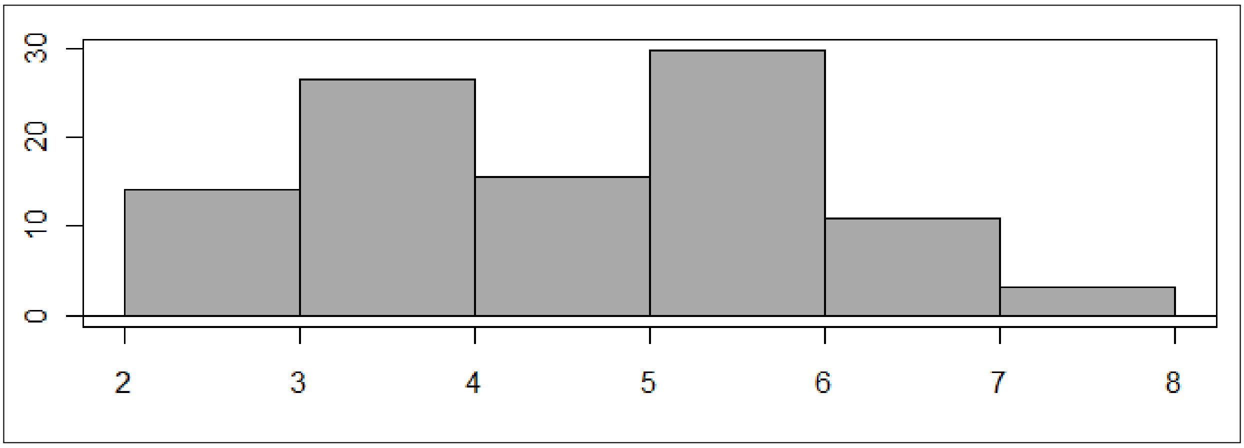


Figure 1: Histogram with results of HL7 EHR-S Functional Model use cases.

It is clear that the majority of use cases evaluation scores does not exceed the value of 5. It means that the target interoperability for all the use cases derived from the HL7 EHR System Functional Model is at most the semantic interoperability. Of course the condition of semantic interoperability is the usage and implementation of all the lower interoperability levels. The integration patterns mentioned in 2.1 corresponding to these levels should be used.

On the other hands any investment into technologies and platforms supporting integration patterns related to the process interoperability level should be considered in detail.

# 4 Discussion

The presented method has been applied to 64 HL7 EHR-S Functional Model uses cases derived from [11]. The understanding of these rules is quite simple, so the use cases can be evaluated also by a person without a specialized training in computer science and software engineering (physician, manager ...). It offers the possibility to bridge the interdisciplinary gap among different actors. This way, a mapping between different Generic Component Model domains [12] is enabled in the integration platform development process.

The method implies the possibility of a structured view to the often heterogeneous set of (business) requirements. It has to be tested, whether the method can really simplify the analysis project phase and enable the development of an early integration platform prototype. The benefit of early prototyping is the possibility to test soon after the requirement specification, to decrease the number of change requests, to speed up the project, and to lower the total costs.

# 5 Conclusion

It seems that the method presented here should accelerate an EHR integration platform analysis and design and save time and costs in this way. The second, but not least, benefit lies in the possibility of gap bridging between various roles interested in above mentioned EHR software analysis and design. But there is one very important condition. The method must be tested on the real EHR integration use cases and only then we can compare the existing EHR integration designed in the traditional way with the design emerging from our method, assess its reliability and continually work on its optimization.

## Acknowledgements

The paper has been supported by the SVV-2013-266 517 project of Charles University in Prague.

# References

[1] Benson, T. Principles of health interoperability HL7 and SNOMED. New York: Springer, 2012, ISBN 978-144-7128-007.

[2] Bloebel, B. Architectural Approach to eHealth for Enabling

 ©2013 EuroMISE s.r.o.

Paradigm Changes in Health. Methods of Information in Medicine. 2010, 49, s.123-134.

[3] Gibbons, P. Coming to Terms: White Paper on Interoperability. In: HL7 [online]. 2007 [cit. 2012-08-14]. Via: http://www.hl7.org/documentcenter/public/wg/ehr/ComingtoTerms2007-03-22.zip

[4] ISO/EN 13606 Health informatics – Electronic health record communication. Geneva, Switzerland: International Organization for Standardization, 2008-2010.

[5] Krsicka, D. a M. Sarek. Automatizace vyuziti blokovych reseni pro vyvoj architektur IS. In: MEDSOFT 2012. Praha: Dum techniky CSVTS, 2012, s. 168-179. ISSN 1803-8115.

[6] Krsicka, D. a M. Sarek. Integracni vzory a jejich automaticke vyhodnocovani. In: Medsoft 2011. Praha: Creative Connections, 2011, s. 146-149. ISSN 1803-8115.

[7] ISO/IEC 7498-1:1994. Information technology - Open Systems Interconnection: Basic Reference Model: The Basic Model. Geneva, Switzerland: International Organization for Standardization, 1997.

[8] Hohpe, G. Enterprise integration patterns: designing, building, and deploying messaging solutions. Boston: Addison-Wesley, 2004, 683 s. ISBN 03-212-0068-3.

[9] Health Level Seven International:HL7 [online]. 2012 [cit. 2012-06-19]. Via: http://www.hl7.org

[10] Five Ws. In: Wikipedia: the free encyclopedia [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2012-08-15]. Via: http://en.wikipedia.org/wiki/Five_Ws

[11] HL7 EHR System Functional Model R1.1. HL7. HL7 [online]. 2012 [cit. 2012-10-29]. From http://www.hl7.org/implement/standards/product_brief.cfm?product_id=269

[12] Bloebel, B., Oemig, F. What Is Needed to Finally Achieve Semantic Interoperability? In: Doessel, O., Schlegel, W. C. (Edrs.) IFMBE Proceedings 25/XII. 2012, p. 411-415

# Overview and Multi-Criteria Analysis of Glucometers for Telemonitoring of Patient with Diabetes Mellitus

**Martina Oulická**[1], **Jan Mužík**[2], **Miroslav Mužný**[2], **Jan Brož**[3,4], **Eirik Årsand**[5,6]

[1] The First Faculty of Medicine, Charles University in Prague

[2] Spin-off Application Centre, The First Faculty of Medicine, Charles University in Prague

[3] Department of Internal Medicine, The Second Faculty of Medicine, Charles University in Prague

[4] Faculty Hospital Motol, Prague, Czech Republic

[5] Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway

[6] Department of Computer Science, University of Tromsø, Tromsø, Norway

## Abstract

**Background:** The telemonitoring is more and more used for compensation of diabetes in the last decade. Doctors are able to get an accurate and reliable data in real time using the telemonitoring. A remote monitoring affects the attitudes and behavior of patients and potentially improves their state of health. Conclusions of many studies show additional clinical implications of telemonitoring. But it has not been possible to generalize those conclusions yet.
**Objectives:** The aim of this study was to create an overview of current glucometers available on the market. And select those which would meet the required parameters for using in the telemonitoring with automatic data sharing.
**Methods:** The research is focused on researches from technical and grey literature and on websites of producers and medical device dealers. The questioning will be carried out in the Czech and English language. Multi decision making method helps to select a suitable glucometer.

**Conclusions:** Fifty five glucometers from nineteen producers have been found in the researches and by market survey. The summary table with all important parameters can be seen in the preview. Conclusions of the Multi decision making analysis showed using of Diamond Mini from ForaCare Inc. producer which is the most suitable for the project of the telemonitoring with automatic data sharing. It is necessary to consider the safety of sending data, data sharing and personal data protection before this glucometer will be used in our project.

## Keywords

Diabetes mellitus, telemonitoring, self management, glucometer, smart phone

**Correspondence to:**

**Martina Oulická**
General University Hospital in Prague
Address: U Nemocnice 2, 128 08 Prague 2, Czech Republic
E–mail: m.oulicka@gmail.com

**EJBI 2013; 9(3):22–27**
recieved: July 31, 2013
accepted: October 28, 2013
published: November 20, 2013

## 1 Introduction

The World Health Organization states in its report [1] the 1.7% prevalence of diabetes mellitus disease for the year 2000 and predicts prevalence of 4.4% for the year 2030. The data are worldwide and refer to all age groups. The number of patients is expected to increase from 171 million in the year 2000 to 366 million in 2030. The WHO predicts rise in direct costs for treatment of diabetes from 2.5% of the annual national budget to 15% in connection with increasing number of diabetic patients. Diabetic patients often suffer also from other diseases which results in decreasing quality of life together with rising costs of provided health care.

The telemonitoring is more and more used for compensation of diabetes in the last decade[2]. Doctors are able to get an accurate and reliable data in real time using the telemonitoring. A remote monitoring affects the attitudes and behavior of patients and potentially improves their state of health [3]. The telemonitoring has an incentive

 ©2013 EuroMISE s.r.o.

and educational effect for the patients [4]. Conclusions of studies focused on the impact of telemonitoring show additional clinical implications of telemonitoring. But it has not been possible to generalize those conclusions yet [5].

The telemonitoring is based on communication between sender and recipient in real time. This allows an immediate reaction of the doctor to the patient‘s impulse. The patient uses an interactive device which transmits data via the Internet to the doctor for checking patient‘s biological parameters. Then the doctor evaluates this data and makes decisions for the following steps [6]. Telemonitoring and gathering the data must not bother patients in terms of physical activity, time etc. Ideally this should be completely automatic [4, 7, 8].

Smart phones with installed applications which contain tools allowing the collection, evaluation and sending data to a doctor can be used for automatic operations [7, 9]. There are currently more than 137 mobile applications available for self-control of the diabetes [10]. The most important parameters are the level of glucose in blood, the dose of insulin, the physical activity and the diet [5]. The patient has to fill in manually most of these parameters in his diabetic diary.

Our goal is to create a telemonitoring system which would measure the patient‘s data and send them to a server as much as possible automatically. The goal is the patient should do the same procedures in the treatment of diabetes such as being without the use of telemonitoring. It should be possible to infer the influence of the telemonitoring to the compensations of diabetes on the basis of the feedback from patients and evaluation of their health status. Hence it is necessary to select such a technical device which allows automatic data sharing without additional burden for the patient

The aim of this study was to create an overview of current glucometers available on the market. And select those which would meet the required parameters and system of telemonitoring described above. Desired parameters are: detection of glucose in blood from 1.1 mmol/l to 33.3 mmol/l (it corresponds with 18 mg/dl to 594 mg/dl) and possibility to share data with a smart phone.

## 2 Methods

The survey was focused on researches from technical and grey literature and on websites of manufacturers and medical device dealers. Search keywords were: a glucometer, the measurement of blood glucose, ketones measurement, measurement + diabetes, glucose monitoring, regulation of glucose, glucose measurement, overview of glucometers, evaluation of glucometers, tests of glucometers. The questioning was carried out in the Czech and English language. The search was particularly focused on glucometers currently offered for sale. A summary table has been created and filled in with found devices and their parameters.

Entered parameters were: designation of the glucometer from the manufacturer; manufacturer; measuring range; measuring time; amount of the blood sample; blood draw location, glucometer memory size; weight; battery life; the ability to share data with a computer and connection specification; the ability to share data with smart phone and again connection specifications. These parameters were obtained from glucometer manuals. The availability on the Czech market, the indicative price and a link to the source are parameters stated in the comments.

The final product will be selected on the basis of a Multi decision making method of selected glucometers. This method is a simple and fast tool. Its conclusions should be objective decisions if comparable items are chosen. The criteria weight calculation formula (1) [11] has been used for assessment in this method. A represents the criteria weight, f is a number of preferences and n is a number of criteria.

$$A = \frac{2 \times f}{n \times (n-1)} \tag{1}$$

## 3 Results

### 3.1 Glucometers Overview

Fifty five glucometers from 19 manufacturers have been found in the researches and by market survey. Their list with all important parameters can be seen in the preview in the table in attachment. All glucometers detect level of glucose in the blood in the range from 1.1 to 33.3 mmol/l. Producers state that 42 devices are able to analyze the blood sample within 5 seconds, 11 devices should analyze the sample up to 10 seconds, two blood glucometers have not got this parameter specified. The memory size to store the measured data ranged from 99 to 4000 records. In some cases it is limited by memory of a smart phone to which the glucometer is connected. 20% of the glucometers state the minimum quantity of the blood sample of 0.3 $\mu$l, 22% states the quantity of 0.5 $\mu$l, 25% of 0.6 $\mu$l. One device required 2 $\mu$l of the sample and another one has no information about the sample quantity in its manual. The weight of the two glucometers was less than 20 grams; the weight of almost 70 % glucometers with battery was between 20 and 60 grams. 35 glucometers which represent more than 63 % of products allow blood draw from alternatives locations. Ten devices allow sampling from a thigh or a calf. Twenty models are available on the Czech market. Forty five devices have possibility of connection to a computer. Thirteen products via a USB cable, two by bluetooth. Another two glucometers is possible to connect via a special connector (e.g. used in iPhones). There are only three models which offer sharing data with smart phones – one via bluetooth and two via a special connector.

Only three glucometers have met the primary parameters – Diamond Mini a iDiamond form ForaCare Inc. and iBGStar glucometer from the producer Sanofi Aven-

Table 1: Evaluation of parameters of selected glucometers.

| Evaluated criteria | Diamond Mini | iDiamond | iBGStar |
|---|---|---|---|
| C1 | 10 | 10 | 5 |
| C2 | 10 | 0 | 0 |
| C3 | 10 | 0 | 5 |
| C4 | 10 | 5 | 5 |

tis. These were subsequently assessed by the multi decision making.

## 3.2 Selecting of Glucometer by Using the Multi Decision Making

Evaluating parameters for Multi decision making are:

- The availability of glucometer on the Czech market (marked C1).
- Possibility to connect devices with a smart phone via bluetooth (marked C2).
- An alternative sampling point of blood (marked C3).
- Equipped with a USB connector (marked C4).

Each parameter was rated from 1 to 10. Where number 10 is a maximum value. More satisfied parameters more assigned points. The Table 1 shows evaluation of parameters.

Criterion C2 – the possibility to connect devices with a smart phone, gives 3 preferences. The others received one preference. The criteria order was given. The criteria weights were calculated using the formula (1) [11]. Importance of each criterion was taken into consideration in the evaluation of parameters of selected glucometers (Table 2).

Conclusions of the Multi decision making analysis showed using of Diamond Mini from ForaCare Inc. producer is the most suitable for our project. This device gained more than double points in comparison with the competitors (Table 3).

# 4 Conclusions

The aim of the study was to carry out an overview of the currently available glucometers on the market. It is possible that not all glucometers were included in this thesis. Reason for leaving out some products is due to orientation on the Czech or English speaking market. Another reason could be difference in search expression. 55 devices from 19 producers were involved in the overview. Only three products have met requirements for using in the telemonitoring with automatic data sharing. Conclusions of the Multi decision making analysis showed the most universal and usable device which is Diamond Mini from ForaCare Inc. producer.

Before this glucometer will be used in this project, it is necessary to consider the safety of sending data, data sharing and personal data protection.

### Acknowledgements

The paper has been supported by the SVV-2013-266 517 project of Charles University in Prague.

Table 2: Criteria preferences.

| Table of preferences | C1 | C2 | C3 | C4 | Number of preferences (f) | Hierarchy of criteria | Scale |
|---|---|---|---|---|---|---|---|
| C1 | x | 0 | 1 | 0 | 1 | 2 | 0.167 |
| C2 | 1 | x | 1 | 1 | 3 | 1 | 0.500 |
| C3 | 0 | 0 | x | 1 | 1 | 2 | 0.167 |
| C4 | 1 | 0 | 0 | x | 1 | 2 | 0.167 |

Table 3: Weight calculation for each criterion of selected glucometers.

| Evaluation the criteria | Scale | Diamond Mini | iDiamond | iBGStar |
|---|---|---|---|---|
| C1 | 0.167 | 1.67 | 1.70 | 0.83 |
| C2 | 0.500 | 5.00 | 0.00 | 0.00 |
| C3 | 0.167 | 1.67 | 0.00 | 0.83 |
| C4 | 0.167 | 1.67 | 0.83 | 0.83 |
| Sum | | **10.00** | 2.50 | 2.50 |


©2013 EuroMISE s.r.o.

## References


[1] World Health Organization. Global status report on noncommunicable diseases 2010. World Health Organization. 2010.

[2] Stone R.A. The Diabetes Telemonitoring Study Extension: an exploratory randomized comparison of alternative interventions to maintain glycemic control after withdrawal of diabetes home telemonitoring. Journal of the American Medical Informatics Association 2012; (66): 973–979. http://www.ncbi.nlm.nih.gov/pubmed/22610495 (accessed 5. December 2012).

[3] Paré G, Jaana M, Sicctte C. Systematic Review of Home Telemonitoring for Chronic Diseases: The Evidence Base. Journal of the American Medical Informatics Association 2007; 14(3):269-277.

[4] Glasgow R.E. D net diabetes self management program: long-term implematation, outcomes, and generatization results. Preventive medicine 2003; (36): 410-419. http://www.ncbi.nlm.nih.gov/pubmed/12649049 (accessed 10. December 2012).

[5] Jaana, M. Home telemonitoring of patients with diabetes: a systematic assessment of observed effects. Journal of Evaluation in Clinical Practice 2006; (13): 242–253. http://www.ncbi.nlm.nih.gov/pubmed/17378871 (accessed 10. December 2012).

[6] Středa L. eHealth a telemedicína: Klinická telemedicína, Zdravotnické noviny: týdeník pro pracovniky ve zdravotnictví. 2013 Aug 26.

[7] Kumar V.S. The DAILY (Daily Automated Intensive Log for Youth) Trial: A Wireless, Portable System to Improve Adherence and Glycemic Control in Youth with Diabetes. Diabetes Technology Therapeutics 2004; (6): 445-453. http://www.ncbi.nlm.nih.gov/pubmed/15320998 (accessed 5. December 2012).

[8] Cava D.A., Everett J, Plougmann S, et al. Use of the Internet to optimize self-management of type 1 diabetes: preliminary experience with DiasNet.. Journal of Telemedicine and Telecare 2003; (9): 50-52. http://www.ncbi.nlm.nih.gov/pubmed/12952722 (accessed 5. December 2012).

[9] Årsand, E. The few touch digital diabetes diary : userinvolved design of mobile self-help tools for peoplewith diabetes. Tromso: University of Tromso, Faculty of Science; 2009.

[10] Chomutare T., Fernandez- Luque L., ?rsand E., et al.. Features of Mobile Diabetes Applications: Review of the Literature and Analysis of Current Applications Compared Against Evidence-Based Guidelines. Journal of Medical Internet Research 2011; (): e65.

[11] Doubravová, H. Vícekriteriální analýza variant a její aplikace v praxi, 1.st ed.; Jihočeská univerzita v Českých Budějovicích, 2009.

[12] Glasgow R.E. Media for Diabetes Self-Management: Issues in Maximizing Public Health Impact. Med Decis Making 2010; (30): 745-758. http://www.ncbi.nlm.nih.gov/pubmed/21183760 (accessed 10. December 2012).

[13] Seto E. Cost Comparison Between Telemonitoring and Usual Care of Heart Failure: A Systematic Review. Telemedicine and e-Health 2008; (14): 679-686. http://online.liebertpub.com/doi/abs/10.1089/tmj.2007.0114 (accessed 10. December 2012).World Health Organization. Global status report on noncommunicable diseases 2010. World Health Organization. 2010.

Appendix 1: Glucometers review based on the market research.

| No. | Type of glucometer | Producer | Range of measurements (mmol/l) | Measurement time (s) | Memory range | The size of a blood sample (μl) | Weight (g) with battery | Possibility of connection with computer | Specification connection with computer | Connection with smarth phones | Specification connection with smarth phones | Posibility alternative place of blood sampling | Availability on the Czech market | Place of blood sampling |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | One Touch UltraEasy | LifeScan | 1.1 - 33.3 | 5 | 500 | 1 | 40 | yes | USB cable | no | - | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 2 | One Touch Vita | LifeScan | 1.1 - 33.3 | 5 | 500 | 1 | 58 | yes | USB cable | no | - | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 3 | One Touch Ultra | LifeScan | 1.1 - 33.3 | 5 | 150 | 1 | 42,5 | yes | USB cable | no | - | yes | yes | finger, forearm, upper arm, hand |
| 4 | One Touch UltraSmart | LifeScan | 1.1 - 33.3 | 5 | 3000 | 1 | 79 | yes | USB cable | no | - | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 5 | One Touch Verio Pro | LifeScan | 1.1 - 33.3 | 5 | 750 | 0,4 | 85 | yes | USB cable | no | - | yes | yes | finger, forearm, hand |
| 6 | Diamond Prima | ForaCare Inc. | 1.1 - 33.3 | 5 | 450 | 0,5 | 52 | yes | USB cable | no | - | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 7 | Diamond Mini | ForaCare Inc. | 1.1 - 33.3 | 5 | 450 | 0,5 | 27 | yes | USB cable, bluetooth | yes | bluetooth | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 8 | iDiamond | ForaCare Inc. | 0.5 - 33.3 | 5 | iphone memory | 0,5 | 34 | yes | iphone connector | yes | iphone connector | not specified | yes | not specified |
| 9 | FreeStyle Lite | Abbot Diabetes Care Inc. | 1.1 - 27.8 | 5 | 400 | 0,3 | 31 | yes | data cable | no | - | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 10 | FreeStyle Freedom Lite | Abbot Diabetes Care Inc. | 1.1 - 27.8 | 4 | 400 | 0,3 | 45 | yes | data cable | no | - | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 11 | FreeStyle Optimum | Abbot Diabetes Care Inc. | 1.1 - 27.8 | 5 | 450 | 0,6 | 42 | yes | not specified | no | - | yes | yes | finger, forearm, upper arm, hand, thigh, calf |
| 12 | FreeStyle InsuLinx | Abbot Diabetes Care Inc. | 1.1 - 27.8 | not specified | 495 | 0,3 | 66 | yes | USB cable | no | - | no | no | finger |
| 13 | FreeStyle Optimum Xceed | Abbot Diabetes Care Inc. | 1.1 - 27.8 | not specified | 450 | 0,3 | 42 | yes | not specified | no | - | yes | no | finger, forearm, upper arm, hand |
| 14 | FeeStyle Lite | Abbot Diabetes Care Inc. | 1.1 - 27.8 | 5 | 400 | 0,3 | 31 | yes | not specified | no | - | no | yes | finger |
| 15 | Precison Xtra Overview | Abbot Diabetes Care Inc. | 1.1 - 27.8 | 5 | 450 | 0,6 | 46 | yes | data cable | no | - | yes | no | finger, forearm, hand |
| 16 | Accu Chek Performa Nano | Roche | 0.6 - 33.3 | 5 | 500 | 0,6 | 40 | yes | infrared | no | - | yes | yes | finger, hand |
| 17 | Accu Chek Aviva | Roche | 0.6 - 33.3 | 5 | 500 | 0,6 | 60 | yes | infrared | no | - | yes | no | finger, hand |
| 18 | Accu Chek Aviva Combo | Roche | 0.6 - 33.3 | 5 | 1000 | 0,6 | 103 | yes | bluetooth | no | - | no | no | finger |
| 19 | Accu Chek Performa | Roche | 0.6 - 33.3 | 5 | 500 | 0,6 | 52 | yes | infrared, cable | no | - | yes | yes | finger, hand |
| 20 | Accu Chek Mobile system | Roche | 0.6 - 33.3 | 5 | 2000 | 0,3 | 129 | yes | infrared | no | - | no | no | finger |
| 21 | Accu Chek Advantage | Roche | test strips | 5 | 480 | not specified | 60 | yes | data cable | no | - | no | no | finger |
| 22 | Accu Chek Aviva Active | Roche | 0.55 - 33.3 | 5 | 200 | 1 | 57 | yes | infrared | no | - | yes | no | finger, forearm, hand |
| 23 | Contour Link | Bayer | 0.6 - 33.3 | 5 | 480 | 0,6 | 53 | yes | not specified | no | - | no | yes | finger |
| 24 | Contour next USB Meter | Bayer | 1.1 - 33.3 | 5 | 2000 | 0,6 | 34 | yes | USB cable | no | - | yes | no | finger, forearm, hand |
| 25 | Contour XT Meter | Bayer | 0.6 - 33.3 | 5 | 480 | 0,6 | 48 | no | - | no | - | no | no | finger |
| 26 | Contour | Bayer | 0.6 - 33.3 | 5 | 480 | 0,6 | 48 | no | - | no | - | yes | no | finger, forearm, hand |
| 27 | Contour TS | Bayer | 0.6 - 33.3 | 8 | 250 | 0,6 | 57 | yes | data cable | no | - | no | yes | finger |
| 28 | Breeze 2 | Bayer | 0.6 - 33.3 | 5 | 420 | 1 | not spedified | yes | data cable | no | - | no | no | finger |

| No. | Type of glucometer | Producer | Range of measurements (mmol/l) | Measurement time (s) | Memory range | The size of a blood sample (µl) | Weight (g) with battery | Possibility of connection with computer | Specification connection with computer | Connection with smarth phones | Specification connection with smarth phones | Posibility alternative place of blood sampling | Availability on the Czech market | Place of blood sampling |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | Didget | Bayer | 1.1 - 33.3 | 5 | 480 | 0,6 | 76 | yes | Nintendo integrated card | no | - | yes | no | finger, forearm |
| 30 | SD Codefree | Standard Diagnostics, Inc. | 0.55 - 33.3 | 5 | 500 | 0,9 | 48 | yes | data cable | no | - | yes | yes | finger, forearm, hand |
| 31 | SD-Check Gold | Standard Diagnostics, Inc. | 0.6 - 33.3 | 5 | 400 | 0,9 | 50 | no | - | no | - | no | yes | finger |
| 32 | SeNova | Chdiagnostics | 1.1 - 33.3 | 10 | 250 | 0,6 | 60 | yes | not specified | no | - | no | yes | finger |
| 33 | True2go | NIPRO Diagnostics | 1.1 - 33.3 | 4 | 99 | 0,5 | 17 | no | - | no | - | yes | no | finger, forearm |
| 34 | TRUEresult | NIPRO Diagnostics | 1.1 - 33.3 | 4 | 500 | 0,5 | 47 | no | - | no | - | no | no | finger |
| 35 | TRUEtrack | NIPRO Diagnostics | 1.1 - 33.3 | 10 | 365 | 1 | 47 | no | - | no | - | yes | no | finger, forearm |
| 36 | TRUEbalance | NIPRO Diagnostics | 1.1 - 33.3 | 10 | 365 | 1 | 47 | no | - | no | - | no | no | finger |
| 37 | TRUEread | NIPRO Diagnostics | 1.1 - 33.3 | 10 | 200 | 1 | 47 | no | - | no | - | yes | no | finger, forearm |
| 38 | Clever Chek | BBI Healthcare | 1.1 - 33.3 | 7 | 450 | 0,7 | 70 | yes | USB cable | no | - | yes | no | finger, forearm, upper arm, hand, thigh, calf |
| 39 | Omnitest 3 | B Braun | 0.55 - 33.3 | 3 | 365 | 0,3 | 54 | yes | data cable | no | - | no | no | finger |
| 40 | Omnitest plus | B Braun | 0.6 - 33.3 | 5 | 250 | 1 | 41 | yes | data cable | no | - | no | yes | finger |
| 41 | GlucoMen LX Plus | A. Menarini Diagnostics LTD | 1.1 - 33.3 | 4 | 4000 | 0,3 | 75 | yes | audiojack | no | - | yes | no | finger, forearm, hand |
| 42 | Glucocard MX | A. Menarini Diagnostics LTD | 0.6 - 33.3 | 5 | 500 | 0,3 | 46 | yes | not specified | no | - | yes | no | finger, hand |
| 43 | Glucocard G+ | A. Menarini Diagnostics LTD | 0.6 - 33.3 | 5,5 | 450 | 0,6 | 50 | yes | audiojack | no | - | yes | no | finger, forearm, hand |
| 44 | Glucocard X | A. Menarini Diagnostics LTD | 0.6 - 33.3 | 5 | 360 | 0,3 | 45 | yes | not specified | no | - | yes | no | finger, forearm, hand |
| 45 | Glucofix mio plus | A. Menarini Diagnostics LTD | 1.1 - 33.3 | 4 | 400 | 0,3 | 75 | yes | not specified | no | - | no | no | finger |
| 46 | GlucoMed GM | A. Menarini Diagnostics LTD | 0.6 - 33.3 | 7 | 250 | 0,5 | 39 | no | - | no | - | yes | no | finger, forearm, hand |
| 47 | GlucoRx Nexus | TaiDoc Technology Corporation | 1.1 - 33.3 | 5 | 1000 | 0,5 | 67 | yes | USB cable | no | - | no | no | finger |
| 48 | IME-DC | Arctic Medical | 1.1 - 33.3 | 10 | 100 | 2 | 57 | yes | USB cable | no | - | yes | no | finger, hand |
| 49 | One Touch Ultra 2 | LifeScan | 1.1 - 33.3 | 5 | 500 | 1 | 42 | yes | data cable | no | - | yes | no | finger, forearm, hand |
| 50 | Discreet all in One | Mendor | 1.1 - 33.3 | 5 | 250 | 0,5 | 28 | yes | data cable | no | - | no | no | finger |
| 51 | BGStar | Sanofi Aventis | 1.1 - 33.3 | 6 | 1865 | 0,5 | 48 | yes | not specified | no | - | no | no | finger |
| 52 | iBGStar | Sanofi Aventis | 1.1 - 33.3 | 6 | 300 | 0,5 | 9 | yes | not specified | yes | iphone connector | yes | no | finger, forearm, hand |
| 53 | CareSens N | Spirit Healthcare | 1.1 - 33.3 | 5 | 250 | 0,5 | 50 | yes | USB cable | no | - | yes | no | finger, forearm, hand, thigh |
| 54 | WaveSense Jazz | Aga Matrix, Inc | 1.1 - 33.3 | 4 | 1865 | 0,5 | 48 | no | - | no | - | yes | no | finger, forearm, hand |
| 55 | PURA | Ypsomed | 0.6 - 33.3 | 5 | 500 | 1 | not spedified | yes | not specified | no | - | yes | no | finger, hand |

# Risk Perception and Environmental Risks Management in Environment and Health Protection Context

**Vladimír Bencko**[1], **John M. Quinn**[1]

[1] Institute of Hygiene and Epidemiology, First Faculty of Medicine, Charles University, Prague, Czech Republic

## Abstract

**Risk Assessment:** The qualified expert assessment of potential ecological/environmental and health risks rising from the planned industrial, transport and waste treatment facilities and other construction activities becomes indispensable. Whereas the initial phase of risk assessment, its identification or potential human exposure are of pure scientific character, the actual risk assessment increasingly assumes the arbitrary aspects (e.g. safety coefficients), risk communication, its control and management by way of psychological aspects; collective decision making then becomes a hotly debated political issue.
**Risk Perception:** When evaluating perception of environmental risks, psychosocial and psychosomatic factors may be of fundamental importance. This is the case in particular where our knowledge of the true health consequences of exposure to given factor is incomplete or its action is within the range of values where we do not anticipate the measurable biological effect. This applies not only in the case of the indoor environment related complains but also e.g. to that of non-ionizing electromagnetic radiation and electro-ionic microclimate. A serious consequence found in the syndrome of mass hysteria is the fact that due to differently motivated disinformation, part of the population can suffer from some psychosomatic symptoms. Those imply objective suffering and deterioration quality of life for those affected.
**Challenges for Prevention:** The prevention of such conditions can either be systematic: early educational or popularisation campaigns, specific health education orientated to the development of industrial, transportation, or other types of constructions, and integration of the local civic activities in the program. The purpose of this should not be a cheap belittling of the risk but reasonable explaining of its acceptable rate, and also the likely advantage to benefit from the realization of the structures. Any later efforts to inform the public about the true state of affairs is usually accepted with distrust and disbelief, in belief this information had been well-paid by the government, industry and market forces, the military or some other institution trying to camouflage the actual condition.

## Keywords

risk perception, environmental risks assessment, psychosomatic aspects, psychosocial aspects of risk, health an illness, scientific and social models, public health

**Correspondence to:**

**Vladimír Bencko**
Institute of Hygiene and Epidemiology,
First Faculty of Medicine, Charles University
Address: Studničkova 7, 128 00 Prague
E–mail: vladimir.bencko@lf1.cuni.cz

**EJBI 2013; 9(3):28–31**
recieved: June 26, 2013, 2013
accepted: October 10, 2013, 2013
published: November 20, 2013

## 1 Introduction

### 1.1 Scientific and Social Models of Health and Illness

When contemplating the problem of a healthy environment in relation to a sick one, it is necessary to define the relationship of health and illness in general. Currently, health is conceived as a condition of physical, psychic, and socioeconomic wellbeing. Contrarily, illness involves an extensive set of different experiences or behaviours of the affected person. Different experience in the negative sense against the generally accepted standard is implying the deteriorated or endangered subjective condition or social function, feeling of undesirability, of being unwelcome and/or unexpected. The illness induces some activities which aim is an improvement of the condition [4].

Every society responds to such impaired function by charging a number of individuals or institutions with duties to evaluate and interpret the actual condition and provide the necessary measures. Public health key stakeholders (owing to public health being both an institution and scientific discipline) whose representatives include physi-

 ©2013 EuroMISE s.r.o.

cians are expected to react to these social requirements and provide best practices and outcomes. Medicine tries to build up a scientific model of illness, its diagnostics, treatment and prevention, yet this model often is neither identical nor congruent with the social one. There is a difference between illness and disease; the same as the difference of views concerning the therapeutic and preventive approach [6].

The priority of the scientific approach is the attempted objectivity and criticism in collecting data and interpreting it. On the contrary, the social model is mostly based on subjective and strongly emotional attitudes. Both, the expert and lay community are not immune against the harmful influence of myths. Science, however, is closer to truth, but not exceptionally, the science-based, as well as lay models, tend to misinterpret the situation, and provide altered or skewed approaches [7]. Using objective methods rooted in the scientific process, it becomes clear to be able to reflect upon failures, where the subjective approach often resists logical argumentation and organized methodology.

Nevertheless, even the scientific process operates with some traditional elements. Max Planck has lamented, "the new scientific truth would not win by convincing the opponents, but rather by letting the opponents die, and the new generation then adopts a new, and own truth." If rationally removing harmful effects and providing for a healthy living environment we have to consider both the scientific and social aspects, i.e. the views and needs of people living in particular environment.

## 2 Assessment of Ecological and Health Risk Factors and Settings

The assessment of potential ecological and health risk rising from the planned industrial transport and waste treatment facilities and other construction activities becomes indispensable [10]. Of course, the public health aspects of such waste management activities is no novelty as such cited projects have legally been controlled and approved by the district or regional public health authorities, within the scope of preventive supervision and best practices.

Whereas the initial phase of risk assessment, its identification or potential human exposure are of pure scientific character, the actual risk assessment increasingly assumes the arbitrary aspects (e.g. safety coefficients), risk communication, its control and management by way of psychological aspects; collective decision making then becomes a hotly debated political issue [5]. As illustrating examples we can use problems related to conflicting views concerning the health risk and associated effects of electromagnetic field and electronic microclimate [11].

The present approach to quantitative risk assessment artificially separating physiologically based pharmacokinetic (PBPK) model and biologically based dose response (BBDR) model needs to be substantially improved. The modelling procedure must go beyond the current organ-tissue based PBPK model as well as the hard-to-modify two-stage BBDR model. It is clear that a model must be flexible and capable of incorporating information about pharmacokinetics and cell signalling response, among other transparent metrics that help to elucidate the situation [5, 18].

A limitation of the present approach to risk assessment is low dose extrapolation of cancer incidence data from both animal (experimental) and human (epidemiology) studies that are most frequently based on models that assume linearity at low doses and low exposures [18]. There are situations in which this assumption could be considered unreasonable. However, because of the lack of data and no alternative methodology for risk extrapolation at present, the model of low - dose linearity continues being used despite existence of qualitative evidence evidencing the contrary. This is specifically relevant in the case of many non-genotoxic carcinogens modulating mitogenic stimulation or suppression of apoptosis - processes regulated by signalling through its impact on gene expression [12]. Dioxins (TCDD) can serve as example of non-genotoxic carcinogen, endocrine disrupter acting through the Ah receptor. It is a general consensus that to resolve this problem, we need to develop a methodology incorporating biological data on mechanisms operating at the cellular or molecular level.

## 3 Psychic Infection and Mass Hysteria

As every expert knows dealing with clients may sometimes bring about a number of both material and psychological problems. Besides, addressing a group of individuals, who, moreover, feel endangered is more complicated still, especially when these groups previously organized in harmony and through a certain hierarchy start to change into disintegrated ones where behaviour suggest the behaviour of masses or of the mob. The mass psychology may appear whenever a sufficient number of persons are gathering around one point of common interest.

The psychology of the group never makes a mere sum of the member's psychology but it has its own individual characteristics. The group as a whole shows better quality than the most inferior members, but the worse judgement and lower IQ compared to the best individuals of the group, and it is prone to getting influenced by emotion rather than so by reality. Another characteristic is behaviour of the group as a mob (aggressive, panicking, etc.) whose activities are more often worse than those of an individual [2, 3, 17, 7].

The basic characteristic of mass dynamics is the "psychic infection" due to increased suggestibility responsible for the sensation of symptoms and subsequent chain reactions. A person in the mob then is capable of acts they would otherwise never have committed as an individual on their own. The cases of mass psychoses are well known

©2013 EuroMISE s.r.o. 

from many literary descriptions of "mass hysteria" in real or supposed exposure to toxic substances, or in health problems and symptoms connected with the indoor environment (sick building syndrome) found in air conditioned structures [16, 15, 19, 14, 13, 8].

In such cases, it can be considered "objective", (i.e. the patient really suffers from them). They are reminded of such symptoms of acute distress but they are less intense and last for a longer period of time, (e.g. for many days, weeks, or months). The affected are aware of the overall stress and tension, fright, shyness, of sensations of oppressiveness and worries, when addressing other people, and vague stressing uncertainty for the future. All these symptoms are accompanied by chronic fatigue, headache, insomnia and other sub-acute vegetative disorders. As the syndrome is not fully debilitating, the patient feels chronically unwell in both his daily duties and his reaction towards other people. Often their capacity of cognition and making sense of daily activities becomes reduced as the result of chronic fatigue and impaired concentration.

The symptomatology fully corresponds to the term "somatization" introduced in the ICD-10 international classification. The point is that emotion - here a very strong one - finds its vegetative correlate occurring in the somatic sphere. An important role in further development plays the "interpretative model" of the patient being xenochtonous in our case (the cause of all trouble comes from outside) and the patient is aware of it (sick building, nearby radar station, TV tower, waste incineration plant etc.). This mass reaction can manifest by two syndrome levels: one prevails the state of anxiety and the other prevails motoric symptoms (e.g. the medieval processions of flagellants praying for aversion of a pest).

The symptoms may appear separately or combined, or occur in turn in the patient. Mass hysteria afflicts men less frequently than women, especially those living in poorer socioeconomic conditions. Mass hysteria is closely connected with the problems of "sick indoor environment" illness. Important here is the firm conviction of outside noxiea responsible for any kind of symptom, further tendency to hypochondria and stress and also hostile attitudes of the patient to anybody to blame for these conditions; in practice materialized by endless weary court trials. In a sense, also collective insistence on Unidentified flying Objects (UFOs) and other paranormal encounters belong to this category.

Yet, not all mass-occurring pathological symptoms are mass-hysteria-related. For example, the mass poisoning of school children in the school canteen in London can be mentioned, manifested by gastrointestinal troubles shortly after lunch. The complex microbiological, hygienic, and toxicological examination included a questionnaire for children, which showed a significant link between the symptoms and consumption of raw cucumber (relative risk 6.1). Microbiologically the cucumbers were safe but pesticide contaminated, as proved by toxicology tests. In the discussion, the authors warn against any overhasty diagnosis of mass hysteria.

Another example can be found with vaccination and concerns of a disproportionate amount of adverse side effects, including cognitive and development impairment. This new version of vaccine denialism has caused some eradicated disease such as measles, mumps and rubella to be once again found in the developed world due to significantly reduced herd immunity due to parents opting out of such otherwise required vaccines. Vaccine-preventable diseases have been a major cause of illness, death, and disability throughout human history. The advent of the modern vaccine era has changed this significantly. In more recent times, there has been much debate in the lay press regarding vaccine safety — namely what possible side-effects vaccines cause and whether these outweigh the risks of leaving a population without a vaccination program. Despite most of the hysteria, some key literatre relating vaccines and childhood development have been comeplteyl rebuked and and withdrawn from the literature. Present use vaccines provide disease coverage to populations, prevent illness and save lives.

Even when the concentration of toxicants fails to reach the risk values, other factors may be involved, e.g. ambient temperature, air humidity, etc., which have up to now not been included in our models but which are able to objectively influence the clinical course, morbidity and mortality rate [1]. There even may occur combination of the actual infection and mass hysteria. In some people evident hypersensitivity to some substances exists: their pathophysiological reaction then is capable of psychogenic effects on the environment.

Nevertheless, we presume the psychosocial aspects may be of basic importance in understanding the potential health risks.

Furthermore, we can expect such problems when our knowledge of actual health effects of human exposure is incomplete or the intensity of exposure oscillates in levels raising doubts as to possible biological effects [9, 5]. Very serious problems, mostly in psychologically unstable patients, are neuro-psychic and psychosomatic symptoms resisting to treatment. Despite the difficulty in objectification, they represent suffering that should not be underrated considering the quality of patient's life.

## 4 Challenges for Prevention

The prevention of such conditions can either be systematic: early educational or popularisation campaigns, specific health education orientated to the development of industrial, transportation, or other types of constructions, and integration of the local civic activities in the program. The purpose of this should not be a cheap belittling of the risk but reasonable explaining of its acceptable rate, and also the likely advantage to benefit from the realization of the structures. Any later efforts to inform the public about the true state of affairs is normally accepted with distrust and disbelief, in belief this information had been well-paid by the government, industry and market forces,


©2013 EuroMISE s.r.o.

the military or some other institution trying to camouflage the actual condition.

## 5 Conclusion

The cases of mass psychoses are well known from many literary descriptions of "mass hysteria" in real or supposed exposure to toxic substances, or in health problems and symptoms connected with the indoor environment (sick building syndrome) found in air conditioned structures [16, 15, 19, 14, 13].

It is therefore recommended to carry out a relevant, competent epidemiological pilot study on potential incidence of some health problems (tumours, congenital malformations, etc.) still before starting the structures, to compare - using a set of reliable data, when the building had already been approved for use - the incident phenomenon with the previous conditions. Such a study, of course, is no alibi. In cases of positive findings the study could serve as basis for rational measures to minimize the health risk due to the operation of the particular facility. The concept of health risk minimization must be included as a theme in all stages of the design and realization, covering all potential risks for the environment and human health. In medicine, the Hippocrates' statement still holds: Life is short, and Art is long; the occasion is fleeting, experience fallacious, and judgment difficult. The physician must not only be prepared to do what is right himself, but must also make the patient, the attendants, and externals to co-operate. If we honour this in therapy, we should do so in prevention of environment related health risks twice as much [6].

### Acknowledgements

The presentation was elaborated within research activities supported by project PRVOUK P28/LF1/6 and by the specific research project no. 266517 "Semantic Interoperability in Biomedicine and Health Care", Charles University in Prague.

## References

[1] Aldous JC, Ellam GA, Murray V, et al. (1994) An outbreak of illness among schoolchildren in London: Toxic poisoning not mass hysteria. J. Epidemiol. Community Health, 48,(1):41-45.

[2] Ali-Gombe A, Guthrie E, McDermott N (1996) Mass hysteria: One syndrome or two? Brit. J. Psychiat., 186, (5):633-635.

[3] Bartholomew RE (1993) Redefining epidemic hysteria: An example from Sweden. Acta Psychiat. Scand., 88 (3):178- 182.

[4] Bencko V, Honzák R. (1997) Importance of psychosocial factors for the rational management of environmental health risks (in Czech) Prakt Lék, 77 (8): 379-381

[5] Bencko V.(2010) Use of epidemiological data for cancer risk assessment: Approaches, concepts and issues. The Open Epidemiology J. (3):147-151

[6] Bencko V.(2011) Psychosomatic and psychosocial aspects of risk perception. In: Bencko V. at al. Hygiene and Epidemiology. Karolinum Press, Charles University in Prague pp.265-269

[7] Burns, W. J., Slovic, P. (2012) 'Risk Perception and Behaviors: Anticipating and Responding to Crises.' Journal of Risk Analysis. Volume 32, issue. DO.10.1111/j.1539-6924.2012.01791.x

[8] Clarc S, Bencko V (1996) Environmental aspects of the reuse of former military lands, NATO//CCMS pilot study. Centr Eur J Publ Health (4) :266-267

[9] Cikrt, M., Bláha, K., Fuchs, A (1993) Chemical risk assessment and management in Czech and Slovak Republics. Centr, Eur. J.Publ. Hlth. (1) : 4-6.

[10] Drew, C.H., Grace, D., Silbernagel, S., Hemmings, E., Smith, A., Griffith, W.C., et al. (2003) 'Nuclear waste transportation: case studies of identifying stakeholder risk information needs.' Environmental Health Perspectives. 2003;111:263–272.

[11] Goldsmith, J R (1996) Epidemiological studies of radiofrequency radiation: Current status and areas of concern. The Sci of the Total Environ 180. Special issue No. 1. :3-8.

[12] Hernández, L.G., van Steeg, H., Luijten, M., van Benthem, J (2009) 'Mechanisms of non-genotoxic carcinogens and importance of a weight of evidence approach,' Mutation Research/Reviews in Mutation Research, Volume 682, Issues 2–3, September–December 2009, Pages 94-109

[13] Holcátová, I., Bencko, V (1997) Health aspects of formaldehyde in the indoor environment. Czech and Slovak experience. Centr. eur. J. publ. Hlth 5 (1):38-42.

[14] Chang, C. C., Ruhl, R. A., Halpern, G. M., et al. (1994) Building components contributors of the sick building syndrome. J. Asthma, 31 (2):127-137.

[15] Klein, D. F.(1993) False suffocation alarms, spontaneous panics, and related conditions: An integrative hypothesis. Arch. Gen. Psychiat., 50 (4):306-317.

[16] Klein, O., Bencko, V(1991) Ecology of Man and Healh (In Czech) Phare 91/24/ SER/3 Min Environ CZ Educatiuon Programme for Environmentally Sustainable Development, p. 189.

[17] Rothman, A L, Weintraub, M I (1995) The sick building syndrome and mass hysteria. Neurol. Clin., 13, (2):405-412.

[18] Ricci, P.F., Sammis, I.R. (2012) 'Regulatory-Science: Biphasic Cancer Models or the LNT – Not Just a Matter of Biology!' Dose – Response; Special Issue on the role of linear and nonlinear dose-response models in public decision-making, Volume 10, Number 2 / 2012 pp. 120 - 154

[19] Ruhl, R. A., Chang, C. C., Halpern, G. M., et al. (1993) The sick building syndrome. II. J. Asthma, 30 (4):297-308.

# Annotation Framework for the Physical Activity Evaluation in Real-Time

Miroslav Mužný[1], Jan Mužík[1]

[1] Spin-off Application Centre, First Medical Faculty, Charles University in Prague, Czech Republic

## Abstract

**Objectives:** Work aims to create a portable tool with a decision support, providing relevant methods for purposes of physical activity evaluation in real-time.
**Methods:** We have utilized accelerometer equipped ez430Chronos watch in conjunction with a preconfigured RaspberryPi-based setup. Wireless transmission of accelerometer data into a running web application instance, which served as a user frontend, is provided through the WebSocket protocol. Decision support is based on a Weka classifier.
**Results:** The proposed framework is ready to be used for the annotation and basic evaluation of physical activity data in a Wi-Fi covered areas. Minor issues are related to the occasional instability of data transmission, which has to be handled consequently.
**Conclusions:** We found the overall framework architecture robust enough to serve its purpose. Next steps in the development will lead to an expansion of outlined functionality.

## Keywords

Human Activity recognition, WebSockets, Activity counts, ez430Chronos, Annotation Framework

**Correspondence to:**

**Miroslav Mužný**
Spin-off Application Centre, First Medical Faculty,
Charles University in Prague
Address: Studnickova 7, Prague 2, 120 00, Czech Republic
E–mail: miroslav.muzny@lf1.cuni.cz

**EJBI 2013; 9(3):32–37**
recieved: July 31, 2013
accepted: November 2, 2013
published: November 20, 2013

## 1 Introduction

Human activity recognition (HAR) is an extensive area of research for a long time. In this respect, variety of different solutions relying on an accelerometer sensor have been proposed. It is evident that recently, many of these solutions started to utilize the pervasiveness of the internet to enrich the knowledge, when operating in a real-life environment. We show, that advances in technology allow to make a tool for acquisition and evaluation of physical activity data in real-time. The usage of the this tool is not bounded by borders of testing under laboratory conditions, but it is utilizable in real-life experiments. The project we are going to present here, takes an advantage of commonly available hardware and software instruments. Therefore, it is particularly easy to reproduce it.

As far we have been using ez430Chronos watch [1] for a purpose of collection of accelerometer data. The watch features wireless transmission of raw accelerometer data. However, the utilization of the watch itself for purpose of analysis of physical activity data in real-time is limited in many ways. One of the shortcomings is directly given by the reasonably small range of the Texas Instruments USB dongle, which is supplied with the watch. This fact contributes to the difficulty of recording accelerometer data in some environments. Another shortcoming is a lack of annotation tool. Therefore, we have decided to develop a portable solution for the collection and annotation of user's physical activity data. The JavaScript enhanced web application served as a user frontend for our more complex framework behind the scene.

Similar approach, introduced as a smart phone application, was proposed by Lara et. al. [2]. The relevant part of the complex real-time HAR system was able to fetch a set of recognizable activities from the remote Application server using a HTTP protocol. The recognition itself took a place on a smart phone and only aggregated feedback data were passed to the remote server.

Another similar approach was proposed by Riboni et. al. [3]. Their COSAR context-aware activity recognition tool based on ontological reasoning and statistical inferencing aimed to recognize more complex activities on a basis of Android device's integrated sensors. Similarly as in the previous case, merged sensor's data were used to form feature vectors, which were consequently evaluated by a remote reasoner.

 ©2013 EuroMISE s.r.o.

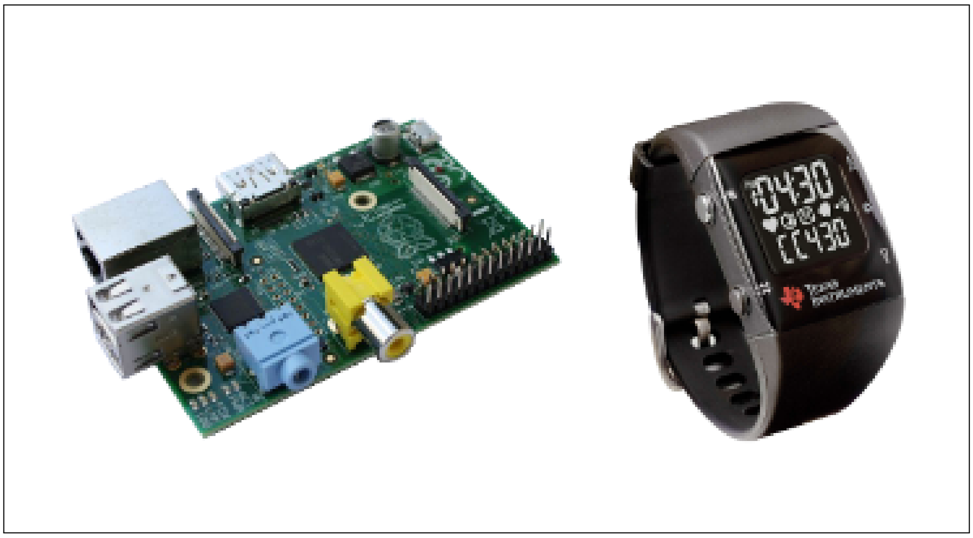


Figure 1: RaspberryPi-based portable setup.

Finally, recently published study about a multi-layer mHealth monitoring system introduced a platform featuring a real-time data transmission based on WebSockets [4]. The platform gathered data from various wirelessly connected sensors in a body area network (BAN) and sent them to a remote server. However, contrary to previously referred projects, this platform served a monitoring purpose and therefore lacked the decision support.

In this work, we deal with a design of annotation framework enhanced by a basic decision support and usable in a real-life environments. We are neither targetting to create a complex BAN, nor the specific smart phone application for an activity recognition. Our goal is an easy to setup, robust tool with a real-time data transmission based on a WebSockets technology, which can be used for a further analysis of physical activity data in real-time.

# 2 Methods

It is important to mention, that in this stage of development we are not focusing on particular aspects of activity recognition process. The robustness and overall stability were primary goals though the real-time nature of the whole system. However, to enable the full data flow through the system, we have configured the crucial elements contributing to the recognition of physical activity according to results in other HAR related works [2, 5].

Modern web browsers provide a wide range of features related to the real time data processing. The WebSocket protocol [6] is one of the technologies, which makes possible to conveniently develop such applications, while retaining odds of web based user interface.

Therefore, our early aim in this project was to ensure that the capabilities of the WebSockets protocol are enough to handle the real-time transmission of accelerometer data. Furthermore, we have experienced a strong demand on the convenient data annotation and visualization. The preservation of the portability of whole setup was our intention as well.

## 2.1 Instrumentation

The portable setup, intended to be worn at user's waist or in a backpack, consisted of ARM-based computer, RaspberryPi [7], along with a connected Texas Instruments USB dongle. Wireless internet connectivity ensured the low-powered USB Wi-Fi adapter [8]. All of the components were powered by an external rechargeable battery pack. That allowed us to keep the overall setup up and running at least for 3 hours. Figure 1 shows the RaspberryPi and the ez430Chronos watch, which served as a wirelessly connected source of raw accelerometer data.

## 2.2 Configuration

RaspberryPi is capable of running several different Linux distributions. We chose an ArchLinux distribution [9] because we have experienced a Wi-Fi connectivity issues when using other distributions. Upon the startup, the device connects to the nearby Wi-Fi network. At the same time, the reverse SSH tunnel with a preconfigured remote computer is automatically established. Tunnel itself solves multiple forthcoming issues. At first, it bypasses eventual NAT and Firewall restrictions of the network. Furthermore, the tunnel makes the annotation framework accessible from the public internet.

We have configured multiple wireless profiles in order to cover the most of the user's space of movement. In case when the device is suffering from a weak signal, the network is automatically switched to the one with a better coverage.

### 2.3 System Architecture

On a basis of our previous positive experience, we have managed to use the JRuby scripting language together with a Rails 4 technology, which already supports WebSockets protocol. Integration of the Weka machine learning framework [10] through its native Java bindings facilitated our development process in many aspects.

Figure 2 shows the system architecture from the viewpoint of basic functional blocks. The core Rails 4 application bundle is denoted by a dashed line. This bundle is actually deployed and runs on the RaspberryPi device.

Sensor-equipped Tester actor is a person who wears an ez430Chronos watch and the RapberryPi setup as well. Administrator actor represents a person who services the Web-based user interface. Administrator has an access to raw sensor data, some of preprocessed data and also to activity labels. Besides a passive participation, Administrator can initiate certain actions towards the framework. This feature is provided by WebSockets full-duplex connectivity [11]. Therefore, existence of Backward parameters setting paths on the Figure 2 should not be understood as a way of automatic adjustment of parameters in an autonomous system. They are dedicated to accomplish actions initiated from the Administrator's side and their specific purpose is further discussed.

The Data preprocessing component handles incoming data from the ez430Chronos watch. The watch's accelerometer sampling frequency is set to 30Hz. For now, we configured the component to extract median average deviation, root mean square and curtosis features from a half overlapping windows with 128 samples. These extracted features are consequently forwarded to a Classification component, which takes a further action.

In our proposed architecture scheme, the Data processing component also serves for the computation of features, which are not passed to the Classification component. These features are meant to be visualized in the Web-based user interface and that is why they are passed directly to the Communication component. We demonstrate this possibility on the example of activity counts, which are further examined in a recent study from the University of Portland [12]. Computation of activity counts commonly precedes the estimation of energy expenditure. We have integrated the estimation as well as the visualization of activity counts of 1 second epochs into our framework. While we are not taking benefit of the activity counts output yet, its estimation looks promising for the future calculation of energy expenditure.

The Classification component currently holds the implementation of C4.5 decision tree from the Weka machine learning framework. The C4.5 classifier was trained to distinguish, on a basis of time domain features from the Data preprocessing module, between several levels of walking and running activities.

We put the Weka activity classification model in a separate file, aside of the Classification component's processing logic. Besides the apparent design intention, there was another reason to make this step. Since the WebSockets protocol supports also transmission of binary data, the separate classification model can be easily replaced by another one from the Web-based user interface.

Event handling and Communication component deserve a special attention. As it acts as a mediator of nearly all communication, its flawless functionality is crucial for the whole system. We had to deal with multiple asynchronous inputs and therefore we have utilized a reactor pattern [13], which was the most likely pattern to meet design requirements. More specifically, we used the EventMachine library, which implements the reactor pattern in Ruby language. Figure 3 shows the Event handling and Communication component in a detailed view with all inputs to better illustrate the bindings to asynchronous events. As noticeable, events originating from the network are based on incoming WebSockets notifications in a JSON format. On the other side, events from the inside of application are fired by a Redis database channel. Redis database is an advanced key value store [14]. The publish/subscribe pattern was implemented in version 2.0. Therefore, it allows to specify a series of channels that a user can subscribe to. This was a basic presumption for an implementation of an event handler, which receives notifications from different parts of the accelerometer data processing pipeline.

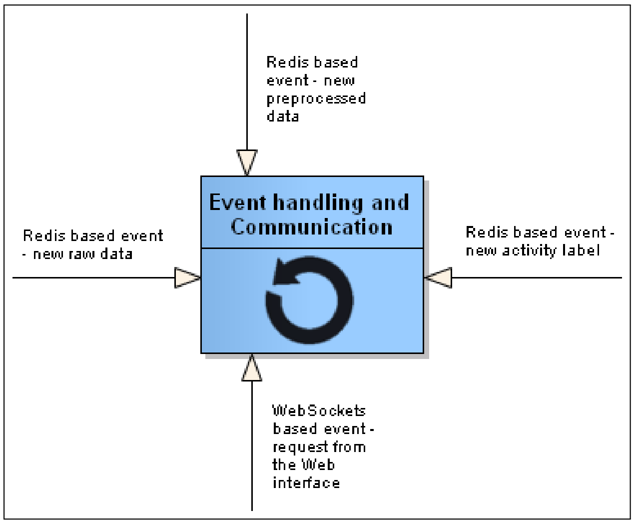


Figure 3: Event handler.

### 2.4 Data Visualization and Interpretation

This section covers the description of Visualization and Management component. Originally, we have been mostly interested in the real time visualization of the raw accelerometer data. For that purpose, we have used the FlotChart, JavaScript graph component [15]. The FlotChart component contains additional set of plugins, which makes easier to consecutively inspect and annotate captured data. From this perspective, our framework currently supports zooming, selection, continuous recording,

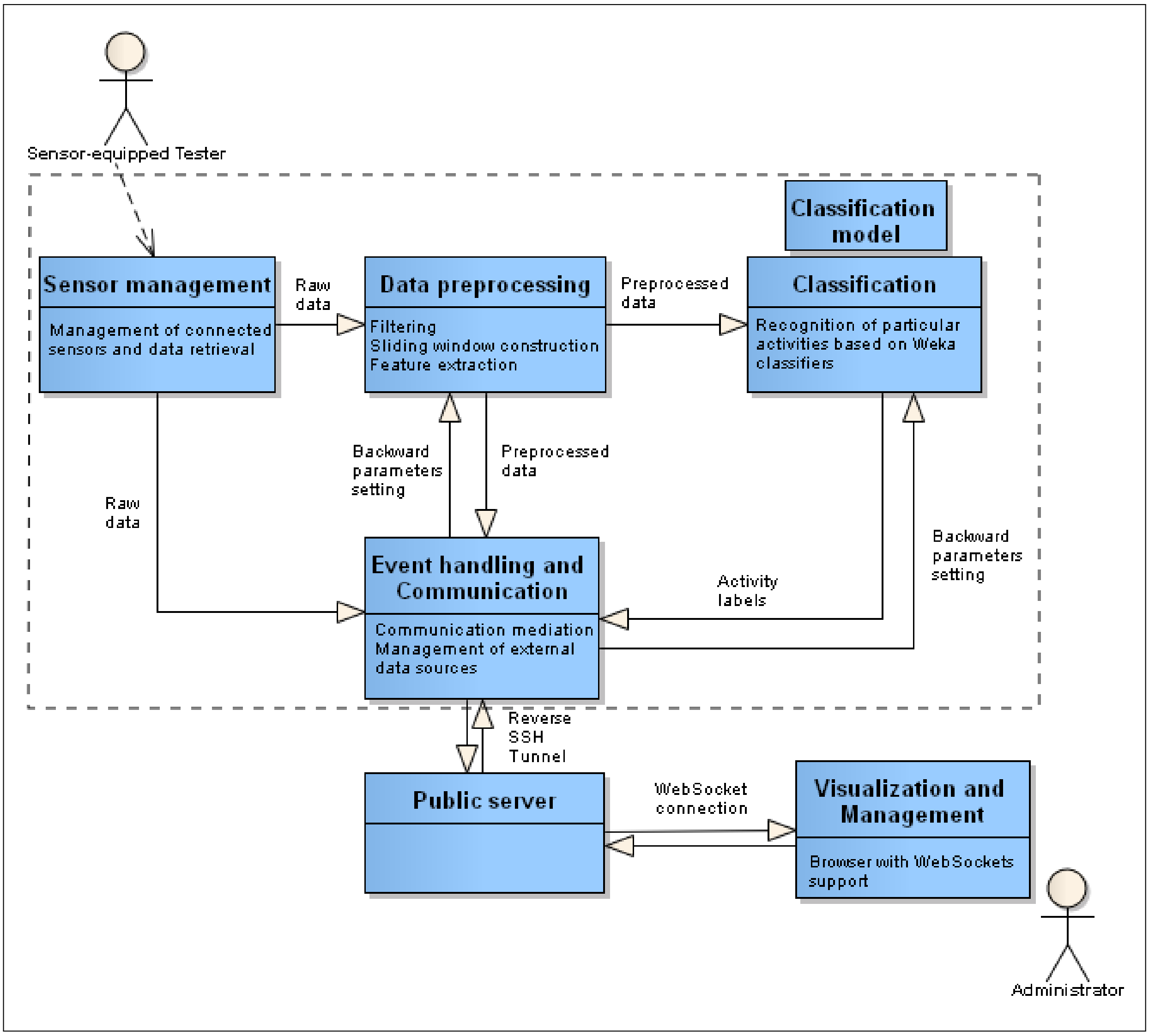


Figure 2: System architecture.

displaying of received annotation flags and insertion of own annotation flags. Besides that, the Administrator is allowed to select accelerometer axes, which has to be displayed. Among implemented functionality counts also an export of recorded data with annotation flags into a file in a CSV format.

The Web-based user interface also serves as a remote management interface. As stated earlier, the Administrator can remotely update the classification model and also adjust settings of Data preprocessing component. At the moment, the only possible adjustment is the change of the time epoch length for the computation of activity counts.

Figure 4 shows the plotting part of the Web-based user interface. The red, blue and yellow series represent incoming data from the 3-axis accelerometer. The green serie in the bottom of figure represents estimated activity counts. On the Figure 4 are also visible annotation flags as the output of the Classification component.

# 3 Results

We have prepared a basic architecture ready to serve for purposes of real-time physical activity evaluation. Fully automatic startup simplified the operation steps on the plugging of the power cable into a battery pack. Relatively small and lightweight package can be worn on a belt around the waist. Also, 3 hours of operation looks promising for most of planned testing scenarios. However, the prototype is not suitable for a daily activity tracking purposes in a comparison with other activity trackers on the market.

Occasionally, we have been dealing with a non-fluent data transmission. As its consequence, the drawing of the chart got stuck until fully re-buffered the data packets. Therefore, we have analyzed the average delay between receiving the pair of incoming packets for the last second. On the basis of its value, we have eventually stopped redrawing the raw accelerometer data, so the transmission

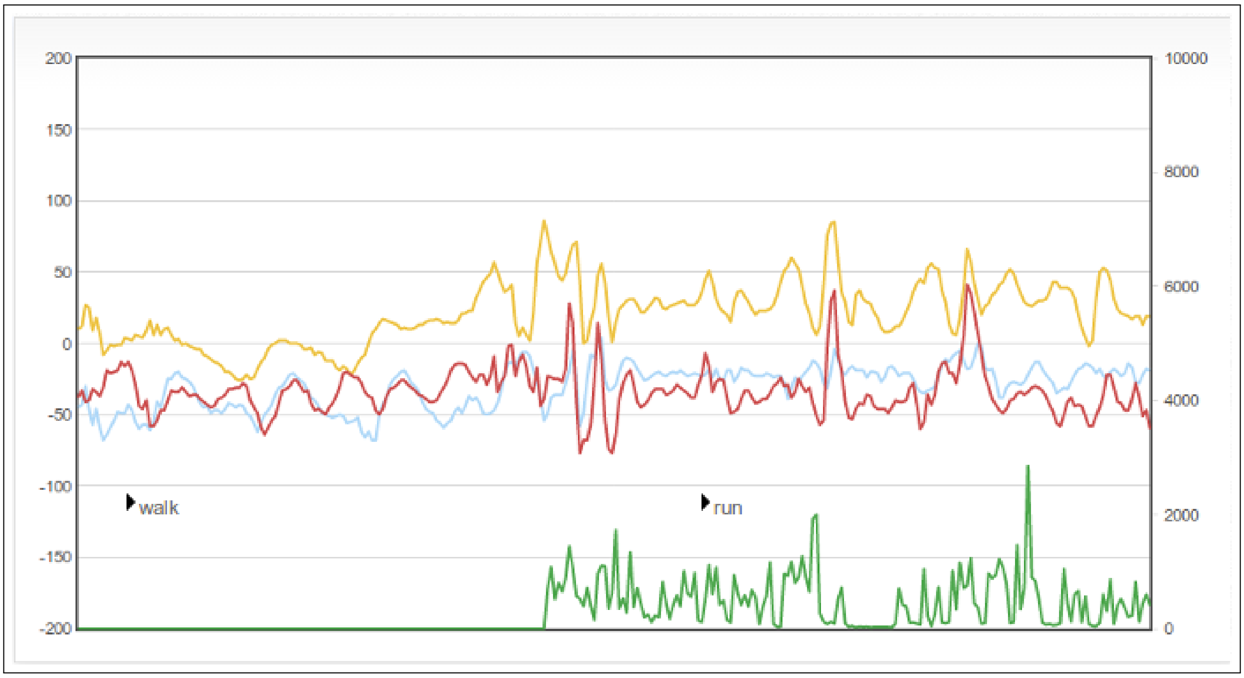


Figure 4: Plotting part of the Web-based user interface.

was cut down to an activity counts data together with output of the Classification component. This feedback-initiated fallback mode is supported by an automatic reconnection process implemented in the WebSockets protocol.

Besides the data transmission related issues, the framework has several minor flaws. For the proper over the internet functionality the framework requires preconfigured computer with a public IP address. With a slight modification, the framework can however be used completely without a network connection. This modification resides in locally storing all data for the future offline analysis. The framework is also dependent on a linux environment for many bindings to native libraries. Therefore, it is not possible to implement it as a smart phone application yet.

## 4 Discussion

We are about to extend the existing framework functionality. Besides its current purpose, the solution could serve as a more complex framework for the collection of data from various sources. That involves also low-powered Bluetooth equipped sensors, which spread out recently. Next steps in the development process will also lead to the expansion of features of the Web-based interface, as most of them currently serve as testimonials of the system architecture.

Results of the framework related research will not be limited to a single device used in our experiments but will affect a wide group of devices. That includes particularly smart phones, which are available to a public audience. Taking advantage of a daily used device with a sufficient computation power could open a range of possibilities regarding the real-time processing of activity data.

After all, we think that our framework holds a potential for a future usage and that we made a step towards achieving a platform, which can be utilizable for the physical activity evaluation in a wide scale.

### Acknowledgements

The paper has been supported by the SVV-2013-266 517 project of Charles University in Prague.

## References

[1] Ez430 - Chronos. [Internet]. [cited 2013 Oct 31]; Available from: http://processors.wiki.ti.com/index.php/EZ430-Chronos

[2] Lara Oscar D, Labrador Miguel A. A mobile platform for real-time human activity recognition. In Consumer Communications and Networking Conference (CCNC), IEEE, 2012, 667-671.

[3] Riboni D, Bettini C. COSAR: hybrid reasoning for context-aware activity recognition. In Personal and Ubiquitous Computing, 2011, 15(3), 271-289.

[4] Zhang W, Stoll R, Stoll N, Thurow K. An mHealth Monitoring System for Telemedicine Based on WebSocket Wireless Communication. Journal of Networks, 2013, 8(4), 955-962.

[5] Chernbumroong S, Atkins A S, Yu H. Activity classification using a single wrist-worn accelerometer. In Software, Knowledge Information, Industrial Management and Applications (SKIMA), 2011, 5th International Conference on (pp. 1-6). IEEE.

[6] The WebSocket Protocol. [Internet]. 2011 [cited 2013 Oct 31]; Available from: http://tools.ietf.org/html/rfc6455

 ©2013 EuroMISE s.r.o.

[7] Raspberry Pi. RASPBERRY PI FOUNDATION. [Internet]. [cited 2013 Oct 31]; Available from: http://www.raspberrypi.org/

[8] Edimax EW-7811Un. EDIMAX. [Internet]. [cited 2013 Oct 31]; Available from: http://www.edimax.com/en/produce _detail.php?pd_id=347pl1_id=1

[9] Arch Linux Arm. [Internet]. 2009 [cited 2013 Oct 31]; Available from: http://archlinuxarm.org/platforms/armv6/raspberry-pi

[10] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

[11] Furukawa Y. Web-based Control Application Using Websocket, ICALEPCS2011, 673-675.

[12] Bernstein D, Coleman E, Lambert A, Peach D, Shindler J, Callender H, Niederhausen N. Comparison of Physical Activity Assessment Tools. [Internet]. 2012 [cited 2013 Oct 31]; Available from: http://reu-ret.hosted.willamette.edu/reuret2012/2012results/UP_REU _Final_Presentation.pdf

[13] Schmidt D. C. Using design patterns to develop reusable object-oriented communication software. In Communications of the ACM, 1995, 38(10), 65-74.

[14] Sanfilippo S, Noordhuis P. Redis. [Internet]. [cited 2013 Oct 31]; Available from: http://redis. io.

[15] Flot: Attractive JavaScript plotting for JQuery. [Internet]. [cited 2013 Oct 31]; Available from: http://www.flotcharts.org/