EJBI 2010 ISSN 1801 - 5603

European Journal for Biomedical Informatics

Volume 7 (2011), Issue 1

An Official Journal of the European Federation for Medical Informatics

www.ejbi.eu

EuroMISE s.r.o.

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.

The generally accepted translations of the English version of the paper are to the following European languages:

List of European languages	ISO 639-1 code
Albanian	sg
Armenian	hy
Azerbaijani	az
Belarusian	be
Bosnian	bs
Bulgarian	bg
Catalan	ca
Croatian	hr
Czech	cs
Danish	da
Dutch	nl
English	en
Estonian	et
Finnish	fi
French	fr
Georgian	ka
German	de
Greek	el
Hungarian	hu
Icelandic	is
Irish	ga
Italian	it
Kazakh	kk
Latvian	lv
Lithuanian	lt
Luxembourgish	lb
Macedonian	mk
Maltese	mt
Norwegian	no
Polish	pl
Portuguese	pt
Romanian, Moldavian, Moldovan	ro
Romansh	rm
Russian	ru
Serbian	sr
Slovak	sk
Slovenian	sl
Spanish	es
Swedish	sv
Turkish	tr
Ukrainian	uk

Editors and Management

Editorial Board: National Members

Editorial Board: Representatives of Cooperating Journals

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
Tel: +420 226 228 904
Fax: +420 241 712 990

Instructions to Authors for the Preparation of Contributions

Abstract

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.

Keywords

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

Introduction

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.

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 the "For Authors" section of the EJBI Website: `www.ejbi.org`. 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.

Manuscript Preparation

You are strongly encouraged to use LaTeX 2_ε for the preparation of your manuscript. Only if you use LaTeX 2_ε can hyperlinks be generated in the online version of your manuscript. The LaTeX source of this instruction file for LaTeX users may be used as a template.

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. These instructions are explained for you in the included MS Word document. 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.

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

Lemmas, Propositions, and Theorems The numbers accorded to lemmas, propositions, and theorems, etc. appear in consecutive order, starting with Lemma 1, and not, for example, with Lemma 11.

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.

Basics of Graphic Composition Less is more! Avoid tables with columns of numbers. Summarise the main conclusion in a figure.

- 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.

Data Formats

- 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 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.

Figures and Tables in LaTeX For LaTeX users, we recommend using the *ejbi-figure* environment (Figure 1 shows an example). The lettering in figures should have a height of 2 mm (10-point type). Figures should be numbered and should have a caption which should always be positioned *under* the figures, in contrast to the caption belonging to a table, which should always appear *above* the table (see an example in Table 1). Short captions are centred by default between the margins and typeset automatically in a smaller font.

Table 1: Age, period, cohort modelling of coronary heart mortality, men, 30-74 yrs., Czech Republic, 1980-2004.

No.	Model	D	*df*	*p*-value
0	Interception	355388.0	44	<0.001
1	Age	15148.0	36	<0.001
2	Age-Drift	3255.5	35	<0.001
3a	Age-Age*Drift	2922.5	27	<0.001
3b	Age-Period	388.2	32	<0.001
3c	Age-Cohort	1872.6	24	<0.001
4	Age-Period-Cohort	28.7	21	0.121

Remark 1. 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 – which is the default if you use the *equation* environment, e.g.

$$\psi(u) = \int_o^T \left[\frac{1}{2} \left(\Lambda_o^{-1} u, u \right) + N^*(-u) \right] dt \, . \qquad (1)$$

Please punctuate a displayed equation in the same way as the ordinary text but with a small space before the end punctuation.

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]

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

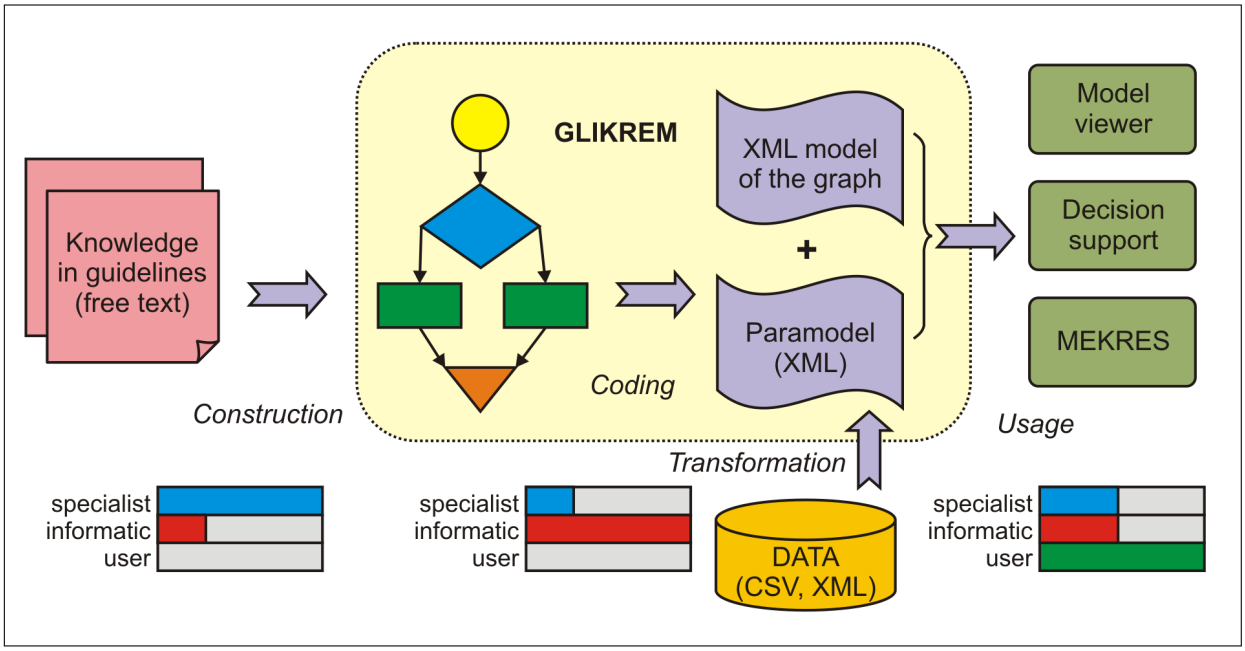


Figure 1: Construction, coding and use of GLIKREM.

Program Code

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

Citations

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. An example is given at the end of this information sheet.

For citations in the text please use square brackets and consecutive numbers: [1], [2, 3, 4]...

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).

Page Numbering and Running Heads

Please do not set running heads or page numbers.

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, in the year 2012 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.

EJBI Online

The online version of the full volume will be available at `www.ejbi.org`.

References

[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.

[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.

Contents

Semantic Interoperability in Biomedicine and Health Care II.

Štěpán Svačina, Jana Zvárová

The first issue of the European Journal for Biomedical Informatics in 2011 publishes peer-reviewed papers of students of the doctoral study at the 1st Faculty of Charles University in Prague. These papers were also presented as lectures given by Ph.D. students during the second workshop on the topic "Semantic interoperability in biomedicine and health care" held on November 24th, 2011 in Prague. The first workshop on the same topic was held on November 18th, 2010 in Prague and all papers of the first workshop were published in English and Czech languages in the European Journal for Biomedical Informatics Vol.6, Issue 1, 2010.

Semantic interoperability and the ability to obtain specific information by usage of technical means are essential conditions for the utilization of other telemedicine technologies and eHealth. The ability of systems to understand exchanged data (semantic interoperability) requires using the same terminology (i.e. classification systems and nomenclatures) and using the same language for communication and its recording (data standards). If information in biomedicine and health care is shared using a free text, a prerequisite for semantic interoperability is the access to its meaning. Existing standards (e.g. EN 13606) suppose the use of globally unique and uniquely defined terms that can be without much difficulty transferred to other classifications (e.g. by means of the Unified Medical Language System). Probably the best applicable general classification system for healthcare is SNOMED CT. It has arisen by a combination of American SNOMED (created by the Association of American Pathologists) and British Clinical Terms ("Read Codes"). In connection with this merger the International Health Terminology Standards Development Organization (IHTSDO), with the residence in Denmark, has been founded in 2007. IHTSDO is a not-for-profit association that develops and promotes use of SNOMED CT to support safe and effective health information exchange. SNOMED CT is a clinical terminology and is considered to be the most comprehensive, multilingual healthcare terminology in the world. SNOMED CT is now being used in a number of information systems for recording of clinical information within patient records. It is expected from modern information systems to work effectively with information and to exchange it mutually.

The task of the workshop supported by the project of the specific research at the 1st Faculty of Medicine of Charles University was also to present selected terms from papers of students and to make their description in English and classification by the SNOMED CT. Then the translations of these findings to the Czech language were presented at the workshop.

Statistical Analysis of Competing Risks: Overall Survival in a Group of Chronic Myeloid Leukemia Patients

Jana Fürstová[1], **Zdeněk Valenta**[1,2]

[1]3rd Medical Department, First Faculty of Medicine, Charles University in Prague, Czech Republic

[2]Department of Medical Informatics, Institute of Computer Science of the AS CR, Prague, Czech Republic

Abstract

Background: Survival analysis is a collection of statistical methods for inference on time-to-event data. If several causes of failure occur and the occurrence of one event precludes the occurrence of the other events, the situation is known as competing risks. Since the competing risks violate the fundamental assumption of independent censoring, specific methods for inference are needed.
Objectives: The aim of this paper is to recall the competing risks model and statistical methods for nonparametric analysis, and to illustrate the competing risks methods on a real data set of 118 Chronic Myeloid Leukemia (CML) patients from the Clinic of Haemato-oncology of the University Hospital in Olomouc.
Methods: The overall survival probability and risk factors of two types of failure (death due to CML and death from other causes) are assessed. Predicted probabilities of the two types of failure with stratification based on the risk factors (Sokal score, haematological response to treatment) are shown.
Results: Outcomes of the specific methods designed for the competing risks analysis are compared with the outcomes of the standard survival analysis methods. The effect of the Sokal score classification is found ambiguous. While the score should identify high- and low-risk CML patients, it seems to be predictive only for the failure due to other causes than CML.

Mgr. Jana Fürstová

Conclusions: The importance of careful censoring and the need of using proper methods of analyses of competing risks data is shown. The use of the Sokal score for classification of the CML patients should be considered more thoroughly.

Keywords

Competing risks, chronic myeloid leukemia (CML), overall survival, cause-specific hazard, cumulative incidence function

Correspondence to:

Mgr. Jana Fürstová
First Faculty of Medicine, Charles University in Prague,
Address: Kateřinská 32, Prague, Czech Republic
E–mail: jana.furstova@email.cz

EJBI 2011; 7(1):2–10
recieved: September 15, 2011
accepted: October 24, 2011
published: November 20, 2011

1 Introduction

Methods of survival analysis have become widely used in medical research in the past few decades. Standard survival data (also called time-to-event data) arise in studies where time from some origin to an end-point is measured. The end-point is defined by occurrence of a certain event of interest. The time until the specified event occurs can be characterized by several functions. The most widely used are the survival function, representing the probability of an individual surviving up to time t (i.e. the probability that the event has not occurred before t), and the hazard function, representing the rate of occurrence of the event at a given time. Under the assumption of in-

dependent censoring, these functions are estimated by the Kaplan-Meier estimator of the survival function and the Nelson-Aalen estimator of the hazard function (for more information, see e.g. [1] or [2]).

In some cases, several causes of failure are possible but the occurrence of one event precludes the occurrence of the other events (e.g. when failures are different causes of death, only the first one can be observed). This situation is known as competing risks. Often, only one event is chosen for analysis, the competing causes of failure are ignored and treated as right-censored observations, and classical survival methods are used for inference [3]. However, this approach leads to a bias in the Kaplan-Meier estimate [4]. The bias is caused by violating one of the fundamental assumptions underlying the Kaplan-Meier estimator – the assumption of independence of distribution of the time to the event and the censoring distribution. Furthermore, independence between distinct causes of failure cannot be verified on the basis of the observed competing risks data [5]. Specific methods are thus needed for the estimation of survival probabilities. The Cox proportional hazards model may be used for regression analysis, but the interpretation of the results becomes different [4].

This paper presents the competing risks model and statistical methods for nonparametric analysis. The methods are then illustrated on real Chronic Myeloid Leukemia (CML) data from the Clinic of Haemato-oncology of the University Hospital in Olomouc, Czech Republic. All statistical methods are implemented with the R software, using the *survival*, *cmprsk* and *mstate* packages [6].

2 Methods

Competing risks are used to model a situation where subjects under investigation are exposed to several causes of failure. If failures represent different causes of death, only the first event to occur is observed. In other settings, second and subsequent failures may be observable, but not of interest. The violation of the assumption of independent censoring, leading to a biased Kaplan-Meier estimator, is an important issue in competing risks models. If the competing event time distributions were independent of the distribution of time to the event of interest, this would imply that at each time the risk of this event is the same for subjects that have not yet failed and are still under follow-up as for subjects that have experienced a competing event by that time [4]. However, a subject that is censored due to failure from a competing risk will certainly not experience the event of interest. Since subjects that will never fail (by the failure of interest) are treated as if they could fail (they are censored), the standard Kaplan-Meier estimator overestimates the probability of failure and underestimates the corresponding survival probability [4], [7].

The competing risks data are represented by the failure time T, the failure cause D and a vector of covariates $\mathbf{Z}$ (T is assumed to be a continuous and positive random variable, D takes values in the finite set $\{1,\ldots,m\}$). Former approach to competing risks used multivariate failure time models. In such models each subject was assumed to have a potential failure time for each type of event. The earliest event was actually observed and the others were latent. This approach focused on the joint distribution of the times $T_1,\ldots,T_m$ of the m different failure types, described by the joint survival function

$$S(t_1,\ldots,t_m) = P(T_1 > t_1,\ldots,T_m > t_m).$$

The marginal hazard function

$$h_j(t) = \lim_{\Delta t\to 0^+} \frac{P(t \le T_j < t+\Delta t | T_j \ge t)}{\Delta t}$$

is defined by the marginal survival

$$S_j(t) = P(T_j > t) = S(0,\ldots,0,t,0,\ldots,0).$$

However, without additional assumptions, neither the joint survival function is identifiable from the observed data, nor are the marginal distributions [2], [8], [5]. This "latent failure time" approach has thus little practical use.

A recent concept in competing risks models is *the cause-specific hazard function* and *the cumulative incidence function.* These two functions completely specify the joint distribution of (T, D), the failure time and the failure cause [9]. The cause-specific hazard function for the $j-$th cause is defined by

$$\lambda_j(t) = \lim_{\Delta t\to 0^+} \frac{P(t \le T < t+\Delta t, D = j | T \ge t)}{\Delta t},$$

for $j = 1,\ldots,m$. It represents the hazard of failing from cause j in the presence of the competing events. The cumulative cause-specific hazard is then defined by

$$\Lambda_j(t) = \int_0^t \lambda_j(u)du.$$

A function $S_j(t) = \exp(-\Lambda_j(t))$ should not be interpreted as a marginal survival function unless the competing event time distributions and the censoring distribution are independent (in case of independent censoring, the marginal distribution models the situation when competing events do not occur) [9]. The total hazard $\lambda(t)$ and the overall survival function $S(t)$ are defined in terms of the cause-specific hazards:

$$\lambda(t) = \lim_{\Delta t\to 0^+} \frac{P(t \le T < t+\Delta t | T \ge t)}{\Delta t} = \sum_{j=1}^{m} \lambda_j(t),$$

$$\begin{aligned} S(t) &= P(T > t) = \exp\left(-\int_0^t \lambda(u)du\right) = \\ &= \exp\left(-\sum_{j=1}^{m}\int_0^t \lambda_j(u)du\right) = \end{aligned}$$

$$= \exp\left(-\sum_{j=1}^{m}\Lambda_j(t)\right).$$

This overall survival function does have an interpretation: It is the probability of not having failed from any cause at time t [3].

The cumulative incidence function of cause j, $I_j(t)$, is defined by

$$I_j(t) = P(T \le t, D = j), \qquad j = 1, \ldots, m,$$

and represents the probability of a subject failing due to cause j in the presence of all the competing risks. It can be expressed in terms of the cause-specific hazard and the overall survival function as

$$I_j(t) = \int_0^t \lambda_j(u)S(u)du, \qquad j = 1, \ldots, m. \qquad (1)$$

This function is sometimes called "crude cumulative incidence function" or "subdistribution function". It is not a proper distribution function because the cumulative probability to fail from cause j remains less than unity, as $I_j(\infty) = P(D = j)$ [1]. The standard Kaplan-Meier estimator of the probability of failing due to cause j before or at time t satisfies

$$1 - S_j(t) = \int_0^t \lambda_j(u)S_j(u)du, \qquad (2)$$

which is similar to the expression of cumulative incidence function $I_j(t)$. Equations (1) and (2) differ by replacing $S(t)$ by $S_j(t)$. Since

$$S(t) \le S_j(t),$$

then

$$I_j(t) \le 1 - S_j(t),$$

with equality at t if there is no competition, i.e. if

$$\sum_{k=1,k\neq j}^{m}\Lambda_k(t) = 0.$$

This shows the bias of the Kaplan-Meier estimator if it is used to estimate $I_j(t)$ [4].

The cumulative incidence function can be estimated using the Kaplan-Meier methodology restricted to specific failures for each cause: Let $0 < t_1 < t_2 < \cdots < t_n$ be the ordered distinct times at which failures of any cause occur. Let d_{jk} denote the number of patients failing from cause j at t_k, and let $d_k = \sum_{j=1}^{m} d_{jk}$ denote the total number of failures (from any cause) at t_k. Let n_k be the number of patients at risk (i.e. patients still in follow-up who have not failed from any cause) at time t_k. Then the cumulative incidence function of cause j at time t is estimated by

$$\hat{I}_j(t) = \sum_{k:t_k\le t}\hat{\lambda}_j(t_k)\hat{S}(t_{k-1}),$$

where the discretized version of the cause-specific hazard $\lambda_j(t_k) = P(T = t_k, D = j|T > t_{k-1})$ is estimated by

$$\hat{\lambda}_j(t_k) = \frac{d_{jk}}{n_k}$$

and

$$\hat{S}(t) = \prod_{k:t_k\le t}\left(1 - \sum_{j=1}^{m}\hat{\lambda}_j(t_k)\right).$$

More detailed derivation of this estimator of $I_j(t)$ can be found in [1] and [4].

In addition to estimating the cumulative incidence functions of the events, it is often of interest to compare the cause-specific cumulative incidence functions among different groups of patients. In standard survival analysis this is done using the nonparametric tests comparing curves generated with the Kaplan-Meier method (e.g. the log-rank test, the Gehan-Wilcoxon test, etc.). In the presence of competing risks, however, these tests are inappropriate. Instead, Gray [10] proposed a class of generalized linear rank statistics for testing equality of the cumulative incidence functions. The tests are based on comparing weighted averages of the hazards of the cumulative incidence function for the failure type of interest.

Consider now a regression model for the competing risks. As in any other regression analysis, it is used to identify potential prognostic factors for a particular failure in the presence of competing risks, or to assess a prognostic factor of interest after adjusting for other potential risk factors in the model. First, consider a regression model for the cause-specific hazard functions. Since the cause-specific hazard functions are identifiable, regression on these functions is possible and a competing risks analogue of the Cox proportional hazards model becomes a logical choice [2]. It models the cause-specific hazard of cause j for a subject with a covariate vector $\mathbf{Z}$ by

$$\lambda_j(t, \mathbf{Z}) = \lambda_{0j}(t)\exp(\beta_j^T\mathbf{Z}),$$

where $\lambda_{0j}(t)$ is the baseline cause-specific hazard of cause j and β_j is a vector of the regression coefficients related to cause j. Both the baseline hazards and the regression coefficients are permitted to vary arbitrarily over the j failure types. Again, let $t_{j1} < t_{j2} < \cdots < t_{jk_j}$ denote the k_j times of type j failures, $j = 1, \ldots, m$, and let $\mathbf{Z}_{ji}$ be the covariates for the individual that fails at t_{ji}. Partial likelihood is constructed with conditioning at each failure time: (1) on the previous history of failures and censoring, (2) that at time t_{ji}, a single type j failure occurs [4]. The partial likelihood function then reads [2]:

$$L(\beta_1, \ldots, \beta_m) = \prod_{j=1}^{m}\prod_{i=1}^{k_j}\frac{\exp\left(\beta_j^T\mathbf{Z}_{ji}(t_{ji})\right)}{\sum_{\gamma\in R(t_{ji})}\exp\left(\beta_j^T\mathbf{Z}_{\gamma}(t_{ji})\right)},$$

where $R(t_{ji})$ is the risk set at time t_{ji}. Estimation and comparison of the regression coefficients β_j can be con-

structed by applying asymptotic likelihood techniques individually to the m factors.

Unfortunately, the cause-specific hazard function does not have a direct interpretation in terms of survival probabilities for the particular failure type. Moreover, the effect of a covariate on the cause-specific hazard function may be very different from the effect of the covariate on the corresponding cumulative incidence function [10]. Therefore, Fine and Gray [11] proposed a method for direct regression modeling on the cumulative incidence functions for the competing risks data. The Fine and Gray model is a semiparametric proportional hazards model using the partial likelihood principle and weighting techniques. It uses a $\log(-\log)$ transformation such that it is reasonable to assume a constant difference between the cumulative incidence functions independent of the time point t. For events of type j, the model reads

$$g_j\left(I_j(t,\mathbf{Z})\right) = h_{0j}(t) + \beta_j^T\mathbf{Z}, \qquad j = 1,\ldots,m,$$

where g_j is some known increasing function, $h_{0j}(t)$ is an invertible and monotone increasing function, $\mathbf{Z}$ is a covariate vector and β_j is a vector of regression coefficients related to cause j. The procedure is based on the transformation

$$g = \log\left(-\log(1-u)\right)$$

corresponding to the proportional hazards model, and it utilizes the subdistribution hazards (hazards related to the cumulative incidence functions) constructed by Gray in [10]. After the transformation, the model reads

$$I_j(t,\mathbf{Z}) = 1 - \exp\left(-\exp(\beta_j^T\mathbf{Z})h_{0j}(t)\right),$$

which allows to assess the effects of the covariates on the cumulative incidence function directly. The partial likelihood constructed by Fine and Gray differs from the traditional cause-specific hazard analysis: in the Fine-Gray model, the risk set for type j events is constructed so that subjects having already experienced events other than type j are always at future risk of a type j event, while in the traditional model the occurrence of an event other than type j removes an individual from future risk sets [11]. A comprehensive discussion may be found in [11] and [12].

3 Data

For illustration of the competing-risks techniques, data from the Clinic of Haemato-oncology of the University Hospital in Olomouc are used. The data contain 118 patients suffering from Chronic Myeloid Leukemia (CML). CML is a cancer of the white blood cells. It is a form of leukemia characterized by the increased and unregulated growth of predominantly myeloid cells in the bone marrow and the accumulation of these cells in the blood. The median age at time of the diagnosis of the disease is 53 years [in 1999], but all age groups, including children, are affected [13]. The natural history of CML is progression from a benign chronic phase to a blast crisis within three to five years [14]. Blast crisis is the final phase in the evolution of CML, and behaves like an acute leukemia, with rapid progression and short survival. The blast crisis is often preceded by an accelerated phase, which signals that the disease is progressing and transformation to blast crisis is imminent. Drug treatment can usually stop this progression if started early [13], [14], [15]. In the Czech Republic, there are about 200 newly diagnosed CML patients per year [16].

All 118 patients in the data set were treated in the Olomouc University Hospital in the years 1989–2010. The last admissible date of diagnosis for the analysis was in 2006 in order to have sufficient follow-up time for all the patients. There is one limitation of the data concerning its consistency: the treatment protocol was changed in 2001 because a new drug – Glivec – had been approved for treatment of the chronic phase of CML. Until 2001, patients were treated by Interferon.

For first-line treatment, Interferon was used for all patients in the Olomouc data set (even those diagnosed after 2001) and most of the patients surviving after 2001 were then treated by Glivec. Out of the 118 patients, 67 are males (57%). The age of the patients at the date of diagnosis ranges from 18 to 71, with the mean of 48 years and median of 50 years. At the date of diagnosis, the Sokal score [17] is evaluated for patients with CML. It identifies low- and high-risk patients according to their age, spleen size and blood cell count.

The high risk group (Sokal score 3) contains 21% of the Olomouc patients ($n = 25$), the low risk group (Sokal score 1) covers 39% ($n = 46$). All other patients were classified with the Sokal score 2. Complete blood count was recorded at the date of diagnosis and haematological response to the treatment was assessed. Overall, 73 patients (62%) achieved complete haematological response (CHR) to the Interferon treatment. The CHR is assessed by improvement of all parameters of the blood cell count of a patient. Median time of CHR is 3 months after the Interferon treatment. Although other types of failure could be considered as well (e.g. progression-free survival, after-treatment survival, etc.), the focus of this paper is the overall survival with initial point being the date of diagnosis of CML and terminal point being death.

The events of interest (competing risks) are two types of failure: death due to CML (includes accelerating disease, progressive disease and blast crisis), and death from other causes (different types of cancer, graft-versus-host disease after stem cell transplantation, suicide, other). By January 2010, 39 patients (33%) have died, 23 patients died due to CML (20%) and 16 due to other causes (14%). Seventy nine patients (67%) did not experience any of these events and were censored in January 2010. All the competing risks estimations are made in terms of the overall survival, i.e. time from the diagnosis of CML to death is considered.

4 Results and Discussion

Figure 1 shows the estimates of the probabilities of "CML-related death" and "death from other causes" for all patients. The CML curves are represented as survival curves, while the other event curves are represented as probability distribution functions (one minus survival) for greater clarity. Estimates based on the Kaplan-Meier method are grey, whereas the estimates of the cumulative incidence functions are black.

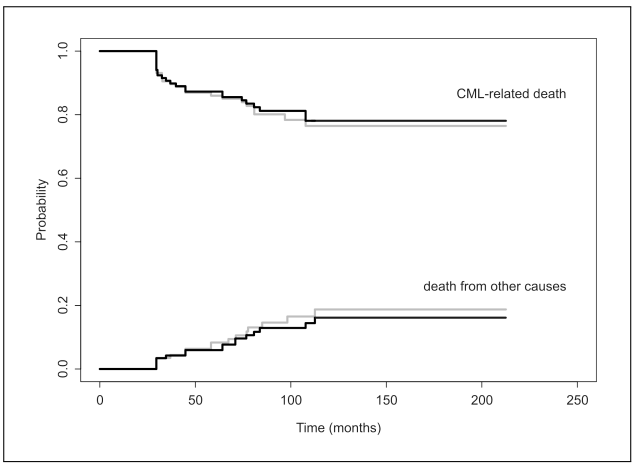


Figure 1: Estimates of probabilities of CML-related death (represented as survival curves) and death from other causes (represented as probability distribution functions), based on Kaplan-Meier (grey) and on cumulative incidence functions (black).

For this data, the two estimates are relatively close to each other, however, the difference between the curves is obvious. The estimates of probability of failure based on Kaplan-Meier after 10 years (120 months) of follow-up are $P = 0.24$ for CML-related event resp. $P = 0.19$ for other type of event, while cumulative incidence estimates are $P = 0.22$ and $P = 0.16$ for CML and other type of event, respectively. This illustrates the formerly mentioned claim that the Kaplan-Meier estimator overestimates the probability of failure and underestimates the corresponding survival probability.

Table 1: Basic characteristics of the continuous covariate variables: age, leukocyte count and haemoglobin level at the date of diagnosis.

	Mean	Median	Min	Max
Age (years)	48	50	18	71
Leu ($\times 10^9/l$)	131	86	2	777
Hgb (g/l)	125	126	70	161

Figure 2 shows the estimated cumulative incidence curves again, displayed in a different way – they are stacked. The bottom curve represents the estimate of the cumulative incidence function of CML ($\hat{I}_{CML}(t)$), the top curve represents the sum of estimates of the cumulative incidence functions of CML and other types of death ($\hat{I}_{CML}(t) + \hat{I}_{other}(t)$). This representation allows an easy comparison of the respective probabilities at any time t.

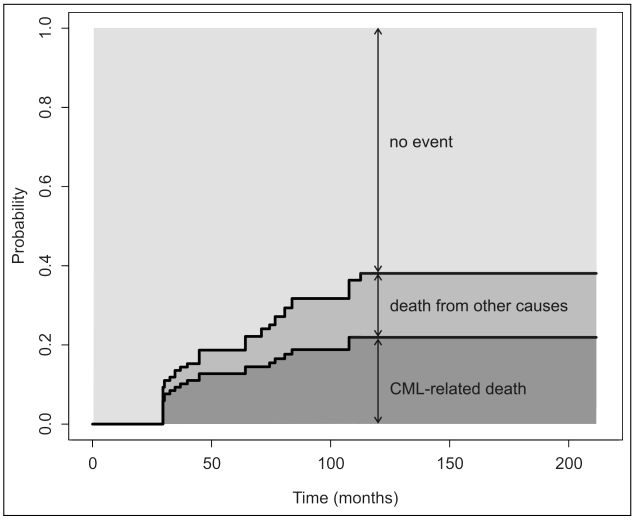


Figure 2: Cumulative incidence curves of CML-related death and death from other causes. Differences between the curves represent probabilities of the particular events.

For the regression analysis on cause-specific hazards, several covariates are used. Basic characteristics of the covariates are shown in Tables 1 and 2. Sex, Sokal score and complete haematological response to treatment (CHR) are categorical variables, whereas age at diagnosis, leukocyte count (Leu) and haemoglobin level (Hgb) at diagnosis are continuous. For purposes of the analyses, in order to make interpretation of results easier, these continuous variables were converted into dichotomous. The cut-off levels were set (by the medical staff) to 45 years of age, $50 \times 10^9/l$ of leukocytes and $110 g/l$ of haemoglobin.

Table 2: Basic characteristics of the categorical covariate variables. One value is missing in the Sokal score and the complete haematological response to treatment (CHR) variable.

		N	%
Sex	male	67	57
	female	51	43
Sokal score	1	46	39
	2	46	39
	3	25	21
CHR	yes	73	62
	no	44	37

Table 3 reports the results of the univariate Cox regression analysis with single covariates sex, age, Leu, Hgb, Sokal score and CHR. It is evident that the blood count has strong effect on the rate of occurrence of CML-related death. The leukocyte level above 50 negatively affects overall survival of the CML patients (hazard ratio (HR)

Table 3: Relative risk estimation for the CML-related death and death from other causes with single covariates, based on the Cox regression model on the cause-specific hazard functions.

	CML		other	
	$\exp(\hat{\beta}_{CML})$	$p-$value	$\exp(\hat{\beta}_{other})$	$p-$value
Sex (male)	1.30	0.55	0.52	0.20
Age (≥ 45)	1.40	0.46	1.43	0.51
Leu (≥ 50)	2.52	0.09	2.31	0.19
Hgb (≥ 110)	0.42	0.04	0.40	0.08
Sokal score	1.43	0.19	2.74	0.004
CHR (yes)	0.33	0.01	0.81	0.70

Table 4: Relative risk estimation for the CML-related death and death from other causes for the Sokal score represented as a pair of dummy variables. Based on the Cox regression model on the cause-specific hazard functions.

	CML		other	
	$\exp(\hat{\beta}_{CML})$	$p-$value	$\exp(\hat{\beta}_{other})$	$p-$value
Sokal score 2 versus 1	1.59	0.35	4.10	0.08
Sokal score 3 versus 1	2.05	0.20	8.92	0.007
Sokal score 3 versus 2	1.29		2.17	

= 2.52, $p = 0.09$), while the effect of haemoglobin level above 110 is protective (HR = 0.42, $p = 0.04$). Patients who achieve complete haematological response to treatment, are in a lower risk of death due to CML (HR = 0.33, $p = 0.01$). There is no evidence of any dependence of CML-related death rates on sex, age or the Sokal score.

On the other hand, the strongest effect on the rate of occurrence of other causes of death is achieved by the Sokal score. The hazard ratio for each extra point in the Sokal score is 2.74 ($p = 0.004$). Thus, an individual having Sokal score 3 has 7.54−times higher risk of death due to other causes compared to the individual having Sokal score 1 (the estimated coefficient $\hat{\beta}_{other} = 1.01$). In case of the Sokal score, it is not important whether the variable is coded as a single covariate (with three categories) or as a pair of dummy variables when modeling. The results are similar (see Table 4).

The effect of haemoglobin level above 110 is the same for other causes of death as for the CML-related death: haemoglobin level above 110 lowers the risk (HR = 0.40, $p = 0.08$). There seems to be no effect of sex, age, leukocyte count and the achievement of complete haematological response to treatment on the risk of death from other causes than CML. However, the results for the sex covariate are interesting. Although the effects are not statistically significant ($p = 0.55$ and $p = 0.20$ for CML and other type of death, respectively), they are opposite for the two types of failure.

In case of CML-related death, males may be in higher risk than females (HR = 1.30), while in case of other types of death, the hazard ratio for males relative to females is 0.52. Sex is the only covariate with such opposite effects on the two types of failure. In the multivariate Cox regression model, no combinations of the above mentioned six covariates prove to have statistically significant effects on the risk of failure due to any of the competing risks.

Table 5: Contingency table with counts of patients according to the Sokal score classification and the cause of death.

	cause		
Sokal score	CML	other	alive
1	7	2	37
2	10	7	29
3	6	7	12

Based on the results of the Cox regression, predicted cumulative incidence curves can be obtained. Figures 3 and 4 show the predicted occurrence of CML-related death and death from other causes for the groups of patients with and without complete haematological response to treatment and for the Sokal score classification. For the CML-related death, the CHR achievement has a strong protective effect: The predicted probabilities of failure due to CML after ten years (120 months) are $P = 0.15$ and $P = 0.38$ for the "CHR yes" and the "CHR no" groups, respectively. On the other hand, there seems to be no relationship between the CHR outcome and failure due to other causes than CML, which is to be expected. For both CHR groups, the predicted probability of death from other causes after ten years from the diagnosis is relatively low ($P = 0.15$). The CHR achievement after the Interferon treatment thus may be used as a reliable predictor of lower

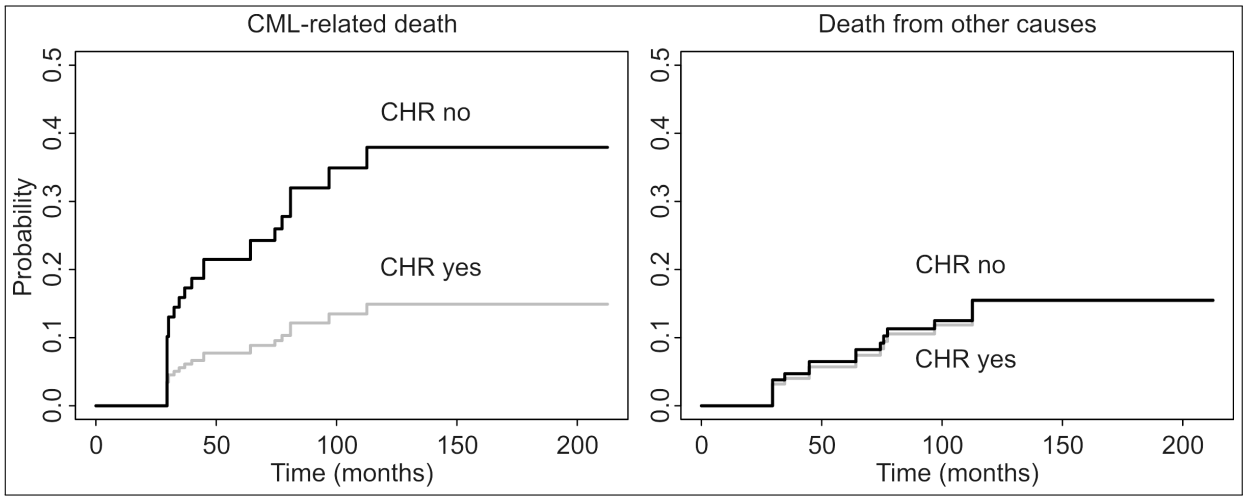


Figure 3: Predicted cumulative incidence functions for CML-related death (left) and death from other causes (right), for patients with and without complete haematological response to treatment, based on the proportional hazards model for the cause-specific hazards.

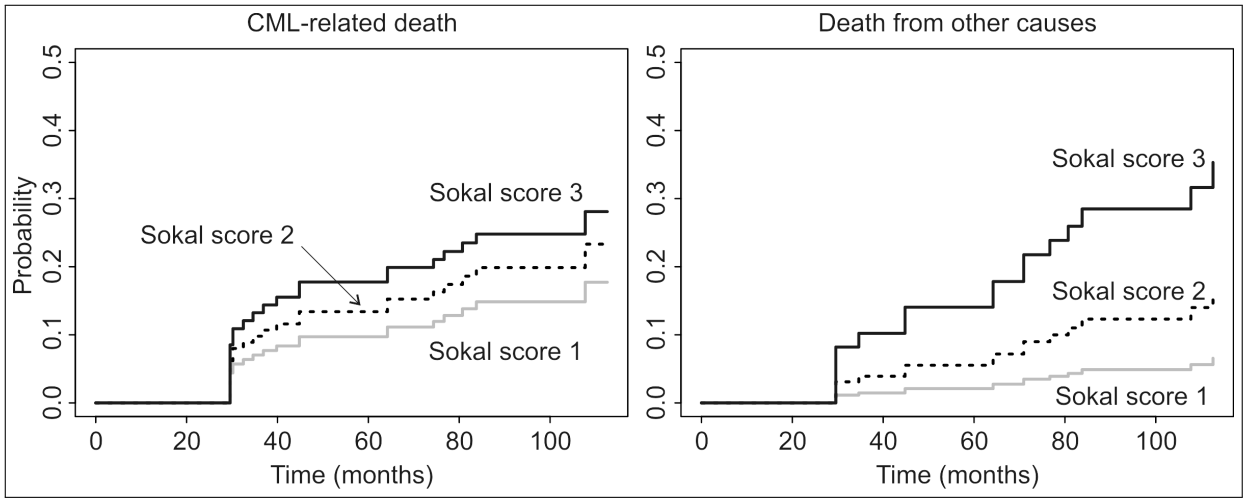


Figure 4: Predicted cumulative incidence functions for CML-related death (left) and death from other causes (right), for the Sokal score classification, based on the proportional hazards model for the cause-specific hazards.

risk of death due to CML. The effect of the Sokal score classification is ambiguous. While the score should identify high- and low-risk CML patients, it seems to be predictive only for the failure due to other causes than CML. The predicted probabilities of death from other causes after ten years are $P = 0.35$ and $P = 0.07$ for the Sokal score 3 group and the Sokal score 1 group, respectively. The predicted probabilities of death from CML after ten years are much closer one to another for all the groups – $P = 0.28$ for Sokal score 3 and $P = 0.18$ for Sokal score 1. For a better insight in the connection of Sokal classification to the different causes of death, contingency table with counts of patients is included (see Table 5). Other predicted cumulative incidence curves are not presented here, as they can easily be obtained from the results of the Cox regression (see Table 3).

To compare the results of the regression on the cause-specific hazards (shown in Table 3) and the regression on the cumulative incidence functions, the Fine and Gray model has been fitted to the data. The results of the Fine and Gray regression are reported in Table 6.

Both the regression models produce similar results for the CML data, thus the main difference between the two models is the interpretation of the results. Cause-specific hazards obtained from the Cox regression model may be translated into cumulative incidence curves, but the proportionality is lost by this process and the covariate effects on the cumulative incidence curves can no longer be expressed by a simple number [4]. Therefore, to determine the effect of a covariate on the cumulative incidence of an event of interest, the Fine and Gray approach using the proportionality of the subdistribution hazards is a better choice.

Table 7 reports the results of the Gray test of the cumulative incidence functions compared with the results of the log-rank test of the Kaplan-Meier estimates of the

Table 6: Relative risk estimation for the CML-related death and death from other causes with single covariates, based on the Fine and Gray regression model of the cumulative incidence functions.

	CML		other	
	$\exp(\hat{\beta}_{CML})$	$p-$value	$\exp(\hat{\beta}_{other})$	$p-$value
Sex (male)	1.42	0.41	0.51	0.17
Age (≥ 45)	1.32	0.55	1.39	0.52
Leu (≥ 50)	2.36	0.12	2.14	0.23
Hgb (≥ 110)	0.46	0.06	0.49	0.16
Sokal score	1.31	0.32	2.58	0.004
CHR (yes)	0.35	0.01	1.02	0.98

survival functions. The tests are computed for all the stratification groups used in the regression models.

The Gray test results are shown for both the competing events, the CML-related death and the death from other causes. For the log-rank test, two schemes of censoring are used: (1) the two types of death are considered separately, i.e. when focusing on the CML-related death, patients experiencing death from other causes are censored as well as patients experiencing no event (and vice versa when focusing on the death from other causes), (2) only event is considered – death from any cause, the differences in types of death are ignored, and censored are only those patients who have not died by January 2010. The censoring scheme (2) completely ignores not only the competing risks methods, but also the possibility of different causes of events.

Table 7: P-values resulting from the Gray test of the cumulative incidence functions (for competing risks) and the log-rank test of the Kaplan-Meier estimates of the survival functions (no competing risks). *Any = Death from any cause (censoring scheme (2)).

	Gray test		Log-rank test		
	CML	other	CML	other	any*
Sex	0.42	0.17	0.46	0.07	0.59
Age	0.53	0.52	0.76	0.33	0.40
Leu	0.11	0.22	0.16	0.14	0.04
Hgb	0.06	0.14	0.08	0.11	0.02
Sokal score	0.59	0.02	0.66	0.008	0.03
CHR	0.01	0.93	0.01	0.60	0.02

Unfortunately, this approach might be quite often in clinical studies where the information about the cause the patients' death are not available. While the results of the Gray test and the log-rank test of the scheme (1) censoring are similar, the results for the scheme (2) differ substantially. The scheme (2) finds statistically significant differences in overall survival between groups of patients stratified by Leu, Hgb, Sokal score and CHR.

However, these results are misleading, as the differences between the groups are limited to the "overall death" only and ignore the influence of the different causes of events. The results in Table 7 show the importance of careful censoring and the need of using proper methods of analyses.

5 Conclusion

The competing risks model and statistical methods for nonparametric analysis are recalled in this paper. The bias in the standard Kaplan-Meier estimator and the need for specific methods for inference on competing risks data is explained. The data set of Chronic Myeloid Leukemia (CML) patients from the Clinic of Haemato-oncology of the University Hospital in Olomouc is analyzed. The overall survival probability and risk factors of two types of failure (death due to CML and death from other causes) are assessed. The interesting role of sex and the Sokal score classification on the overall survival of the CML patients is discussed. Predicted probabilities of the two types of failure with stratification based on the chosen risk factors are shown. Results of the specific methods designed for the competing risks analysis are compared with the results of the standard survival analysis methods. The effect of the Sokal score classification is found ambiguous. While the score should identify high- and low-risk CML patients, it seems to be predictive only for the failure due to other causes than CML. The use of the Sokal score should be considered more thoroughly.

Acknowledgment

The authors are grateful to Assoc. Prof. Edgar Faber, M.D. and the Clinic of Haemato-oncology of the University Hospital in Olomouc for providing the data and introducing the authors into the subject of CML. The paper was supported by the project SVV-2011-262514 of Charles University in Prague.

References

[1] J. P. Klein, M. L. Moeschberger: *Survival Analysis. Techniques for Censored and Truncated Data.* Springer, New York, 2003.

[2] J. D. Kalbfleisch, R. L. Prentice: *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, New York, 2002.

[3] N. Porta, G. Gómez, M. Luz Calle: *The Role of Survival Functions in Competing Risks.* available at *http://www.eio.upc.es/nporta*, cited on June 20, 2011.

[4] H. Putter, M. Fiocco, R. B. Geskus: *Tutorial in Biostatistics: Competing Risks and Multi-State Events.* Statistics in Medicine 26 (2006), 2389–2430.

[5] A. Tsiatis: *A Nonidentifiability Aspect of the Problem of Competing Risks.* Proceedings of the National Academy of Sciences USA 72 (1975), 20–22.

[6] *http://www.R-project.org*

[7] T. A. Gooley, W. Leisenring, J. Crowley, B. E. Storer: *Estimation of Failure Probabilities in the Presence of Competing Risks: New Representations of Old Estimators.* Statistics in Medicine 18 (1999), 695–706.

[8] D. R. Cox: *The Analysis of Exponentially Distributed Lifetimes with 2 Types of Failure.* Journal of the Royal Statistical Society, Series B 21 (1959), 411–421.

[9] J. Lawless: *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons, New York, 2003.

[10] R. J. Gray: *A Class of K−sample Tests for Comparing the Cumulative Incidence of a Competing Risk.* The Annals of Statistics 16 (1988), 1141–1154.

[11] J. P. Fine, R. J. Gray: *A Proportional Hazards Model for the Subdistribution of a Competing Risk.* Journal of the American Statistical Association 94 (1999), 496–509.

[12] J.-H. Jeong, J. P. Fine: *Parametric Regression on Cumulative Incidence Function.* Biostatistics 8 (2007), 184–196.

[13] C. L. Sawyers: *Chronic Myeloid Leukemia. Review Article.* The New England Journal of Medicine 340 (1999), 1330–1340.

[14] S. Faderl, M. Talpaz, Z. Estrov, H. M. Kantarjian: *Chronic Myelogenous Leukemia: Biology and Therapy.* Annals of Internal Medicine 131 (1999), 207–219.

[15] *http://en.wikipedia.org*, cited on July 13, 2011.

[16] *http://diagnoza-cml.cz*, cited on June 20, 2011.

[17] J. Sokal, M. Baccarani, D. Russo, S. Tura: *Staging and Prognosis in Chronic Myelogenous Leukemia.* Seminars in Hematology 25 (1988), 49–61.

[18] J. Fürstová: *Competing Risks of CML-related Death and Death from Other Causes.* Doktorandský den 2011, Matfyzpress, Praha, 2011.

[19] P. Hougaard: *Analysis of Multivariate Survival Data.* Springer, New York, 2000.

[20] H. T. Kim: *Cumulative incidence in Competing Risks Data and Competing Risks Regression Analysis.* American Association for Cancer Research 13 (2007), 559–565.

[21] R. L. Prentice, B. J. Williams, A. V. Peterson: *On the Regression Analysis of Multivariate Failure Time Data.* Biometrika 68 (1981), 373–379.

[22] H. Putter: *Tutorial in Biostatistics: Competing Risks and Multi-State Models. Analysis Using the mstate Package.* available at *http://cran.r-project.org/ web/ packages/ mstate/ vignettes/ Tutorial.pdf*, cited on June 1, 2011.

[23] L. Scrucca, A. Santucci, F. Aversa: *Competing Risk Analysis Using R: An Easy Guide for Clinicians.* Bone Marrow Transplantation 40 (2007), 381–387.

[24] T. M. Therneau, P. M. Grambsch: *Modeling Survival Data. Extending the Cox Model.* Springer, New York, 2000.

[25] P. R. Williamson, R. Kolamunnage-Dona, C. T. Smith: *The Influence of Competing-Risks Setting on the Choice of Hypothesis Test for Treatment Effect.* Biostatistics 8 (2007), 689–694.

Electronic System for Data Record and Automatic Diagnosis Assessment in the Temporomandibular Joint Disorders

Radek Hippmann [1,2,4], **Miroslav Nagy** [2,3], **Taťjana Dostálová** [1,2], **Jana Zvárová** [2,3,4], **Michaela Seydlová** [1,2]

[1] Department of Paediatric Stomatology, Second Faculty of Medicine of Charles University, Prague, Czech Republic

[2] Centre of Biomedical Informatics, Czech Republic

[3] Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic

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

Abstract

Background: The research goal of the Dental segment of the Centre of Biomedical Informatics is focused on the electronic health record (EHR) development for dentistry.
Objectives: At the beginning there has been constructed an electronic dental cross "DentCross", which was representing patients dental data in the graphical form. It has been completed with the system of the automatic speech recognition (ASR) and voice synthesis module (TTS).
Methods: The main goal of this work was to reach the high entirety of the system and its automatization. For this reason it has been completed with the special record medium for the temporomandibular disorders (TMD).
Results: Concerning the experience with the old version the knowledge database (KB) for TMD has been structured differently. A classification diagnostic schema by the American Academy of Orofacial Pain (AAOP) has been used. The KB has been created in the MUDR KB Editor application. On this basis a relational database has been constructed and a user interface for data collection based on MUDR and MUDRLite EHR systems was developed.
Conclusions: The main advantage of this system is determination of probable diagnosis of the disease (AAOP) by the system ("custom" component). It is based on the characteristic data, which have been recorded in the electronic form after the investigation. For the creation of the component of its alone MS Visual Studio.NET 2003 development tool has been used. The whole component is programmed in C♯ language.

MUDr. Radek Hippmann

Keywords

Temporomandibular joint, temporomandibular joint disorders, DentCross, electronic health record, AAOP classification

Correspondence to:

MUDr. Radek Hippmann
Department of Paediatric Stomatology,
Second Faculty of Medicine of Charles University, Prague
Address: FN Motol, V Úvalu 84, Prague 5, 150 06
E–mail: r.hip@seznam.cz

EJBI 2011; 7(1):11–16
recieved: October 2, 2011
accepted: November 1, 2011
published: November 20, 2011

1 Introduction

Nowadays, the level and general development of medical care for patients brings not only much higher demands on knowledge and skills of providers (medical personnel, especially doctors), but also on appropriate systems for storing and handling the increasing amount of data. The main task of these systems for electronic health record (EHR) is another quality of care. Earlier data recording in the classical paper form is no longer for its complexity and quantity of data so convenient and development of IT technologies enable us to leave this type of data storage. Today's EHR systems thus serve not only for easier data manipulation, faster communication and even

greater data protection against abuse, but also allow more comprehensive, more effective and safer medical care.

All these parameters and needs apply in different areas of medicine. Dentistry is no exception and there is an effort to develop an EHR for this area, with all the specifics that it brings. It is about a graphic design application for dentistry, and shot the whole issue of orofacial system, including contingent multifactorial diseases such as temporomandibular joint disorders in particular (TMD). Electronic health records just for temporomandibular joint disease are not yet available and its development is due to the complexity of the problem very complicated.

The temporomandibular joint (TMJ) itself is among the most complex joints throughout the body. It is the compound joint and is closely linked to the chewing muscles and periodontal complex. The American Academy of Orofacial Pain, defined in 2008 the stomatognathic system as a "functional and anatomical relationship between the teeth, jaws, muscles of mastication and TMJ" [1]. Due to the complexity of the system of conditional multifactorial pathology, and many system functions (speech, mastication) [2] is the correct diagnosis, screening and choosing right therapeutic regimen a very demanding process that requires especially good access and organization of all information.

There are several internationally recognized classifications, but few of them are used extensively and are also usable internationally. Probably the most used is the AAOP (American Academy of Orofacial Pain) classification [1], which is used more for clinical use and classification of the RDC / TMD (Research Diagnostic Criteria for Temporomandibular Disorders) [3], which is more suitable for scientific purposes.

The actual study and treatment of TMD are very important and desirable because it is a relatively neglected area. TMD being deemed to be the most frequent cause of pain in the orofacial area of non-dental origin [4]. Among the most common clinical signs of TMD belong muscle or joint pain, joint sound phenomena, limited mouth opening or deviation of jaw when opening mouth [5]. Furthermore, it can be associated with symptoms as headache or earache, neuralgia, vertigo, and toothache.

EHR development in the field of dentistry and TMD are engaged in a long time and the experience with its use and modern trends, so we decided to update and rearrange the electronic form of TMD. The aim is to achieve better control and transparency for the end user, the complexity of the knowledge base for the TMD in the final stage of preparation and applications for automatic diagnosis from the specified anamnestic data.

2 Methods

2.1 MUDR EHR, MUDRLite, DentCross

For purposes of our research the MUDR KB Editor modeling tool, which has been developed in the EuroMISE Centre, has been used to create a model for MUDR EHR. In this application the knowledge base for the field of dentistry [6, 7] has been already constructed and now the new version of TMJ knowledge base was added. MUDR EHR is a predecessor of the MUDRLite application, which is serving rather for the smaller environment needs (e.g. dental outpatient department).

Based on this two applications the DentCross system has been developed (for easier applicability by the final user and better graphical image), which represents user interface and TMD form. This form has been transformed after clinical experiences and completed with the system for automatic diagnosis selection. Voice control by the Automatic speech recognition (ASR) system and module for the voice synthesis Text-to-speech (TTS) are also included (same as the older version of DentCross) [7].

2.2 Classification TMD (AAOP, RDC/TMD)

As mentioned above, the best scheme for the TMD classification seems to be from our point of view the classification performed by the American Academy of Orofacial Pain [1], which is suitable for the clinical use and is concerned with the whole problematic of temporomandibular joint (TMJ) and surrounding diseases.

It is more extensive than RDC/TMD [3, 7], which is dealing more with the temporomandibular joint and is omitting many facts. Both of them have been involved in the knowledge database for TMJ, but for the user interface just the AAOP classification has been used.

AAOP

1. TMJ
   (a) *Congenital diseases:*
      I. *Aplasia* – connected with hemifacial mikrosomia, face asymmetry, congenital, insufficient growth of the one side condylus, movement to the affected side, restricted mouth opening.
      II. *Hypoplasia* – growth limitation, other bones could be affected, less serious than I., restricted mouth opening, asymmetry during mouth opening.
      III. *Hyperplasia* – excessive skull growth, face asymmetry, asymmetry during mouth opening.
      IV. *Dysplasia* – childhood and adolescence, slowly excessive growth and fibrotic tissue changes, face asymmetry and mouth opening.
      V. *Tumors* – complex affection, medical imagining with pathological findings, pain, restricted mouth opening e.g.

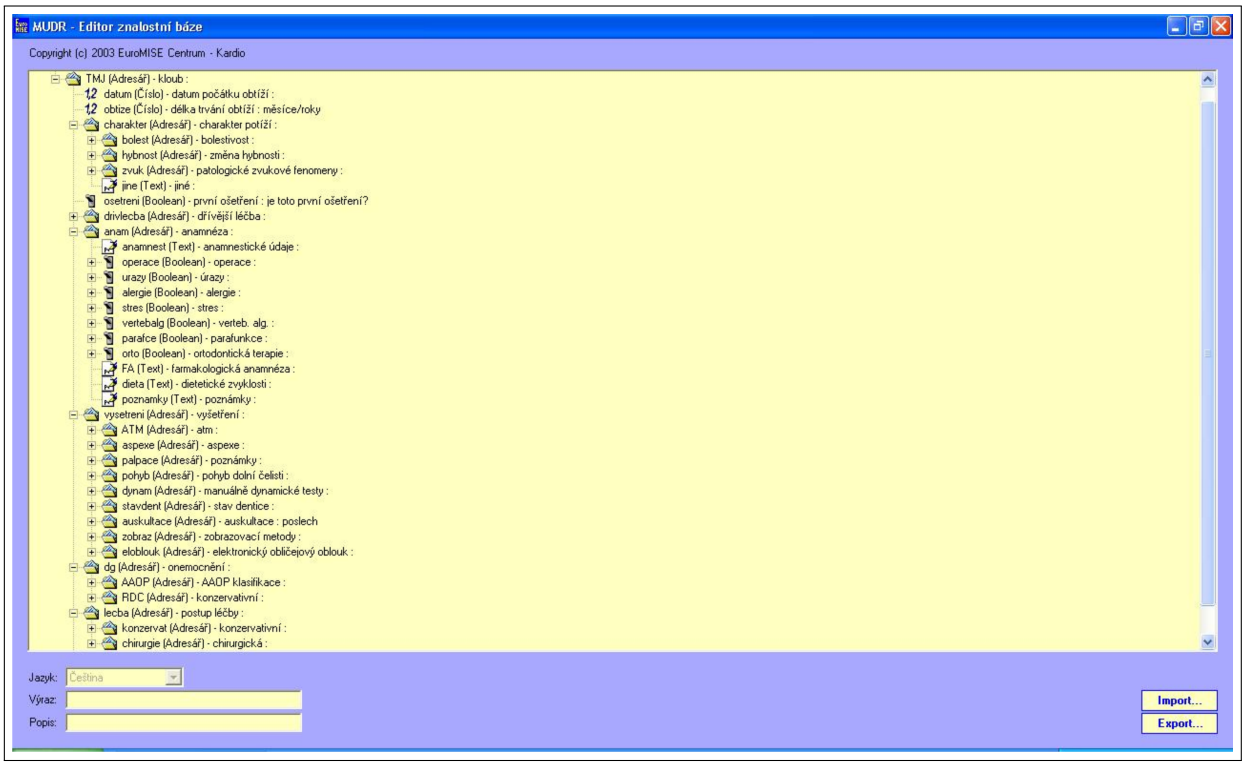


Figure 1: Knowledge base for TMD.

(b) *Joint disc diseases:*

I. *Dislocation with reduction* – joint sounds during jaw opening and closing, asymmetric opening (S-shape deviation towards affected side), restricted mouth opening.

II. *Dislocation without reduction* – restricted and asymmetric mouth opening (movement towards affected side), without joint sounds.

(c) *Dislocation TMJ (luxation)* – acute state, pain, patient is not able to close his mouth.

(d) *Inflammatory diseases* – arthritis (capsulitis, synovitis, polyarthritis included) – restricted mouth opening, pain, joint sounds, painful palpation in the TMJ region.

(e) *Non-inflammtory diseases* – arthosis – primary (etiology not known) and secondary (consequence of the trauma or systemic disease). Chronic affection. Restricted mouth opening, without pain, rigidity feeling, sound phenomena (crepitation, creaking), asymmetric opening (movement towards affected side).

(f) *Ankylosis* – restricted to impossible mouth opening, protrusion and lateropulsion limitation, face asymmetry, chronic disease.

(g) *TMJ fracture* – trauma in anamnesis, pain (spontaneous as well as during palpation), positive X-ray findings, possible restricted mouth opening, disrupted articulation, asymmetric opening, acute origin.

2. Masticatory muscles affections:

(a) *Local myalgia (muscle pains)* – painful TMJ movement and muscle palpation, could be connected with stress, disrupted articulation and muscle injury.

(b) *Myofascial dysfunctional syndrome* – rest pains, trigger points present, could be connected with vertigo, steaming, parafunctions, stress etc.

(c) *Central provoked myalgia* – chronic, long lasting muscle pain similar to myositis, but without inflammatory sings. Could be combined with neurological inflammation symptoms. Important is presence of the persistent pain.

(d) *Myospasmus (muscle contraction)* – acute involuntary tonic muscle contraction, restricted mouth opening, without severe pain.

(e) *Myositis (muscle inflammation)* – rest and movement muscle pain, red skin in region, swelling possible.

(f) *Myofibrotic contraction* – chronic, pain not present. Caused by fibrotic degeneration of tendons, ligaments and rarely of muscle fibers. Manifested as muscle shortening.

(g) *Muscle tumors* – benign and malignant character, wide scale of symptoms, positive histological findings and medical imagining.

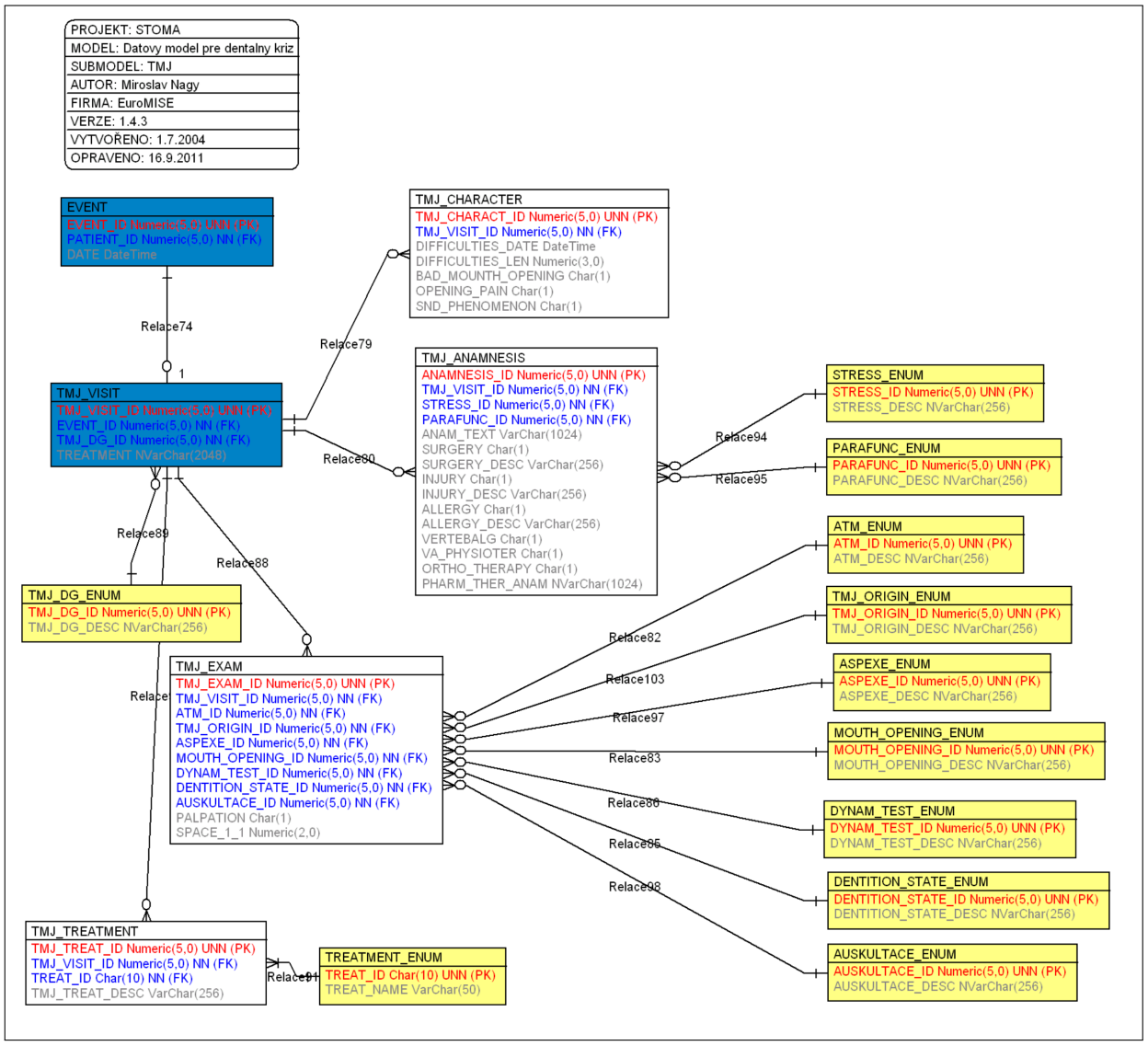


Figure 2: Relational data model for TMD.

2.3 Description of the Module for Diagnosis Assessment Support

The new version of the MUDRLite TMD form was simplified and in comparison to the former version the form was altered to match the examination process of the patient performed by the physician. Another major improvement of the TMD form is a custom component, defined by the MUDRLite system to enable the extension of the forms functionality in the sophisticated manner. In our case the form was enabled to propose a diagnosis based on data collected during the patient examination.

The new component, designed to support decisions of the physician in the form of diagnosis proposals, is based on the AAOP classification. For each diagnosis contained in this classification a logical formula was created, forming a set of implications. The assumptions of each implication are formed from parameters collected via the TMD form. If the conditions of a given implication are met then the relevant diagnosis is added to the proposed diagnoses list and presented to the physician.

The MS Visual Studio .NET 2003 was used to develop the decision support component and the programming language used for this was C♯. The component is connected to the MUDRLite system in the form of dll library.

3 Results

3.1 Knowledge Database for TMD

The whole knowledge base, which was created in the MUDR KB Editor application [6], has been completed and sorted according to the new findings and complies to a different logical system. It is now more oriented on the logical order during patient investigation (Fig. 1).

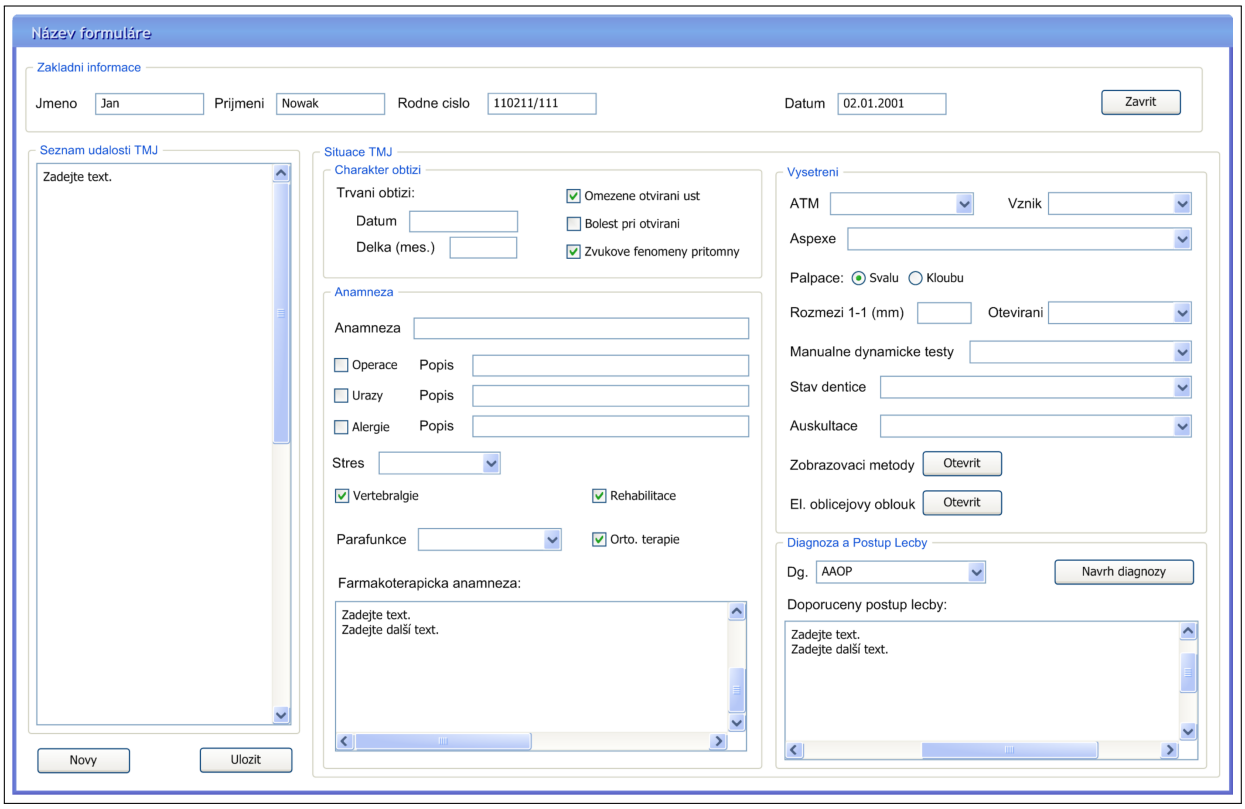


Figure 3: User interface window for TMD record in MUDRLite EHR.

3.2 Data Model

Before data conversion from the TMD knowledge database to the user application it was necessary to choose important modalities, which are essential for patient investigation. They have been transformed into a relational database model by using MUDRLite EHR (Fig. 2). This model represents relations among modalities and thanks to base on this database it is possible to create a form, which is integrated with the DentCross application.

3.3 Application with Automatic Diagnosis Statement for TMD

The form window for anamnestic and investigation data record is (as it was in the old version) a part of the DentCross application and could be opened from the main window. Better information structuring has been made with the respect to logical attitude by investigation and also restriction on important data has been made. Subsequently, separate windows have been added for imaging with description (e.g. RTG) and electronic face bow data (Fig. 3). As already mentioned, the diagnosis assessment is based on the modified AAOP classification and is executed by an automatic component, which is fixing this probable diagnosis in compliance with the patient investigation. This simplifies the therapy process for the physician, because of a very wide variety of TMD states and diagnoses.

Following therapy modalities are different. They are directly connected with the stated diagnosis and seriousness of presented symptoms. They could be conservative (soft diet, splint, non-steroid antiflogistics, dry heat, isometric exercise etc.), miniinvasive (arthrocentesis, arthroscopy), surgical (operation with joint discus reposition, joint substitution) or combined.

4 Discussion

The entire EHR application for dentistry, the DentCross, seeks to provide a comprehensive recording options for the whole stomatognatic system. The comprehensive record is further complemented by the information that a voice control (ASR) further adds and backward interpretation of the record (TTS) by a computer. This allows quick control and contactless applications and data recording [6, 7].

Further expansion is aimed at trying to automate the process of diagnosis and thus a therapeutic plan. Our system is now complemented by the automatic speech recognition and clinically tested at a specialist department for TMD in the Faculty Hospital in Motol.

It should enable the less experienced dentists to perform quality evaluation of the patient with a record of all important data, not only anamnestic data. Then, based on these data, the diagnosis is established and is followed either by a treatment or by referral sent to a specialized

department for a too complicated condition of the case, which should be followed by a comprehensive therapy.

This objective has many pitfalls and should be further complemented and tested in particular. It should be noted that all the time at least basic knowledge of this issue by the dentist will be needed. The role of human factor in final decision making and control of automated applications will still be significant.

Acknowledgements

This paper was supported by the project 1M06014 Ministry of Education, Youth and Sports CR, by the project SVV-2011-262514 of Charles University in Prague and the project AV0Z10300504 of the Academy of Sciences of the Czech Republic.

References

[1] The Americia Academy of Orofacial Pain, de Leeuw R. Orofacial pain: Guidelines for assessment, diagnosis and management, ed 4. Chicago, Quintessence Publishing, 2008.

[2] The Academy of Prosthodontics. Glossary of prosthodontic terms, ed 6. Chicago. Mosby, 1994.

[3] Dworkin S, LeResche L. Research diagnostic criteria for temporomandibular disorders: review, criteria, examinations and specifications, critique. J Craniomandib Disord Fac Oral Pain 1992; 6: 301-355.

[4] LeResche L. Epidemiology of temporomandibular disorders: implications for the investigation of etiologic factors. Crit Rev Oral Biol Med 1997; 8:291 305.

[5] Laskin DM. Etiology of the pain-dysfunction syndrome. J Am Dent Assoc 1969; 79:147-153.

[6] Hippmann R, Nagy M. Dostalova T, Zvarov Seydlova M, Feltlova E: Electronic health record for temporomandibular joint disorders support in therapeutic process. EJBI 2010.

[7] Hippmann R., Dostalova T., Zvarova J., Nagy M., Seydlova M., Hanzlicek P., Kriz P., Smidl L., Trmal.: Voice supported electronic health record for temporomandibular joint disorders. Methods of information in medicine, 2010, 49:168-172.

Traditional Measures of Diversity and Sensitivity of Power Entropies

Martin Horáček[1,2], **Jana Zvárová**[1,2]

[1] Center of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

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

Abstract

Objectives: We dealt with the traditional measures of diversity and their sample estimates. We also studied a way to compare sensitivity to changes of different measures of diversity.
Methods: We proposed a new estimator of measures of diversity. We compared our estimator with three established estimators in a simulation study. We introduced a function called sensitivity to changes of a measure of diversity H and we described its basic characteristics.
Results: The proposed estimator compares favorably to other well established estimators. The sensitivity to changes has a clear interpretation and is easy to compute.
Conclusions: The sensitivity of measure of diversity to changes could be used to compare behavior of different measures of diversity and to select one or few that are the most suitable for a given problem.

Mgr. Martin Horáček

Keywords

Diversity, entropy, diversity estimates, sensitivity

Correspondence to:

Mgr. Martin Horáček
Center of Biomedical Informatics,
Institute of Computer Science AS CR
Address: Pod Vodárenskou věží 2, 182 07 Prague
E–mail: horacek@euromise.cz

EJBI 2011; 7(1):17–21
recieved: September 29, 2011
accepted: October 24, 2011
published: November 20, 2011

1 Introduction

In this paper, we dealt with functions that aim to capture the diversity of a given population. The diversity may relate e.g. to the genetic diversity - diversity of alleles of a chosen gene, to the species diversity in a chosen location, but also to a language or economic diversity. We were namely interested in the situation when the amount of diversity of the population depends solely on the probabilities p_i that an individual randomly sampled from the population contains the i-th out of r possible different mutually exclusive features. When these functions satisfy some additional requirements (described in the next paragraph) that are natural for a function that captures the diversity of a population, they are called the traditional measures of diversity.

More formally, traditional measure of diversity is a real functions H defined on the domain $\Delta^r = \{p = (p_1, \ldots, p_r) : \sum_{i=1}^r p_i = 1,\ p_i \geq 0\ \forall\, i\}$ that is

- nonnegative,
- symmetric with respect to permutations,
- minimal when one $p_i = 1$ (only one feature appears in the population)
- maximal when all $p_i \equiv 1/r$ (features are uniformly distributed),
- when one greater p_i increases at the expense of one smaller p_j, the value of $H(p)$ should not rise.

It may be useful to note that if a function $H : \Delta^r \to R^+$ is Schur-concave and nonnegative, it satisfies all five requirements (see [3]).

There are several frequently used traditional measures of diversity. We present some of them in the next section. Most of them are included or closely related to the f-entropies - a family of generalized entropies - proposed by Zvárová [1]. This family of entropies, their characteristics and how they can be used as measures of diversity is further studied i.e. in Zvárová, Vajda [2] and Horáček [3].

2 Traditional Measures of Diversity and Their Estimates

In this section, we introduce some of the most common traditional diversity measures like Simpson's index, Shannon's entropy, Renyi's entropy of order α, Hill's index and others. We introduce the topic of sample estimates of traditional measures of diversity and we develop a new type of estimator. Parts of sections 2 and 3 were published in the proceedings of the 7th Summer School on Computational Biology [4].

2.1 Examples of Traditional Measures of Diversity

The most often mentioned and used diversity measures include the number of features (e.g. alleles or species)

$$H_0(p) = \sum_{i=1}^{r} I_{(0,1]}(p_i) - 1$$

(where I denotes the identity function), the Simpson's index

$$H_2(p) = 1 - \sum_{i=1}^{r} p_i^2$$

and the Shannon's entropy

$$H_1(p) = -\sum_{i=1}^{r} p_i \ln p_i.$$

These three indices are generalized by the family of power entropies

$$H_\alpha(p) = (\alpha - 1)^{-1} \left(1 - \sum_{i=1}^{r} p_i^\alpha\right), \quad \text{when } \alpha > 0, \alpha \neq 1,$$

defined as limits when $\alpha = 0$ (identical to number of features) and $\alpha = 1$ (Shannon's entropy). When $\alpha = 2$, we get the Simpson's index.

Another frequently mentioned and used indices include the γ-entropic function

$$H_{\mathrm{A},\gamma}(p) = (1-\gamma)^{-1} \left[1 - \left(\sum_{i=1}^{r} p_i^{1/\gamma}\right)^{\gamma}\right], \text{ if } \gamma > 0, \gamma \neq 1,$$

Hill's index

$$H_{\mathrm{H},\alpha}(p) = \left(\sum_{i=1}^{r} p_i^\alpha\right)^{\frac{1}{1-\alpha}}, \quad \text{when } \alpha > 0,\ \alpha \neq 1$$

and Rényi's entropy of order α

$$H_{\mathrm{R},\alpha}(p) = (1-\alpha)^{-1} \ln \left(\sum_{i=1}^{r} p_i^\alpha\right), \quad \text{when } \alpha > 0, \alpha \neq 1.$$

The introduced generalized parametric indices are handy in several ways. Namely they could be used to improve properties of some procedures based on the common Shannon's entropy. When the Shannon's entropy is replaced by a suitable parametric index, its variable parameter could be used to fine-tune the procedures. This is done e.g. in Andrade and Wang [5].

2.2 Sample Estimates of Traditional Measures of Diversity

Let $p = \{p_1, \ldots, p_r\} \in \Delta^r$ be a vector of unknown probabilities p_i that an individual randomly chosen from a population has a feature of type A_i out of r possible features. In this situation, the estimate of measure of diversity $H(p)$ is usually done on the basis of relative frequencies $\hat{p}_n = (X_1/n, \ldots, X_r/n) = (\hat{p}_i, \ldots, \hat{p}_r)$ of features observed in a sample of n individuals selected from the population randomly with replacement. In that case, the distribution of the vector $X = (X_1, \ldots, X_r)$ is multinomial $\mathrm{M}(n, p)$. Several estimators that use the observed relative frequencies were suggested in the past. Their qualities, namely their bias and variance, respectively their mean squared error, may vary depending on the chosen diversity index and on the population in which they are used.

The most commonly used estimator, often called the "plug-in" estimator, consists in simply replacing the unknown probabilities p_i with the observed relative frequencies $\hat{p}_i$. However, despite $\hat{p}_i$ is an unbiased estimate of p_i, the plug-in estimator is generally biased.

Sometimes, the bias could be easily corrected. For example, the mean value of the plug-in estimate of Simpson's index is

$$\begin{aligned}
\mathsf{E}H_2(\hat{p}_n) &= 1 - n^{-2} \sum_{i=1}^{r} \mathsf{E}X_i^2 \\
&= 1 - n^{-2} \sum_{i=1}^{r} \left[\mathsf{var}X_i + (\mathsf{E}X_i)^2\right] \\
&= 1 - n^{-2} \sum_{i=1}^{r} \left[np_i(1-p_i) + n^2 p_i^2\right] \\
&= \left(1 - n^{-1}\right) H_2(p).
\end{aligned}$$

Thus, the unbiased estimate of Simpson's index is

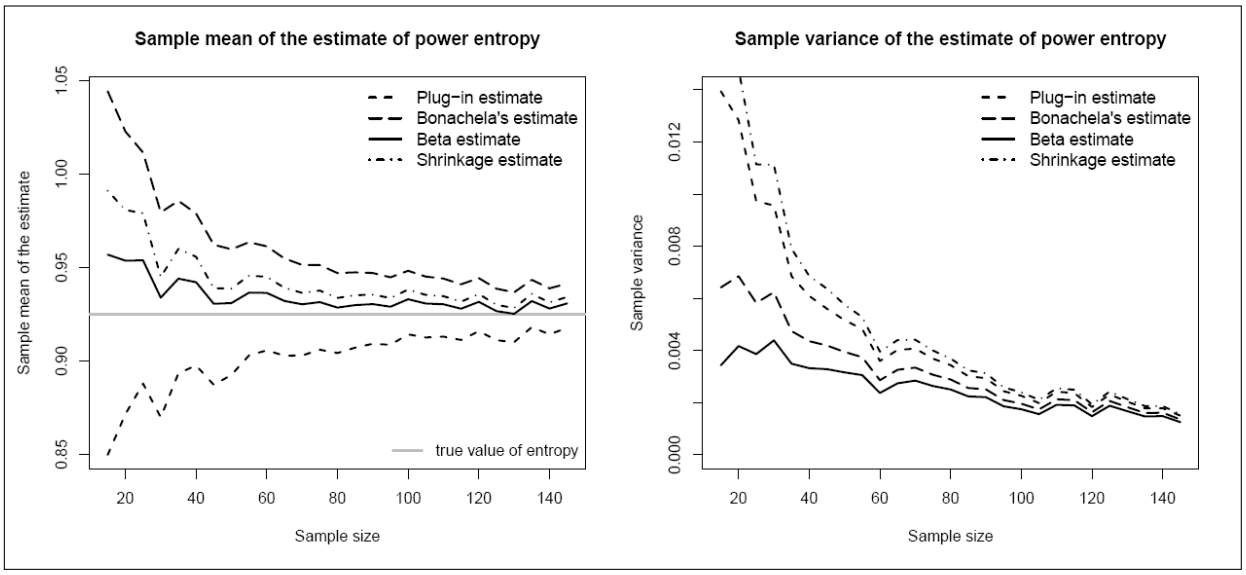


Figure 1: The sample mean and sample variance - estimates of power entropy $H_{3/2}$.

given by

$$\hat{H}_2(\hat{p}_n) = n(n-1)^{-1} H_2(\hat{p}_n).$$

However, it is often difficult or impossible to find a good correction of the plug-in estimate for other diversity measures. For example, it can be shown that Shannon's index does not have an unbiased estimate (Blyth [6]). Hence, several authors dealt with this problem and suggested more sophisticated estimators. We present an estimator proposed by Bonachela et al. [7] that is called the balanced estimator. We suggested a modification of this estimator that takes into account the likely distribution of values of p_i in the interval $[0, 1]$.

Let us assume that the diversity measure is in the form

$$H(p) = F\left(\sum_{i=1}^{r} h(p_i)\right),$$

where F and h are an arbitrary real continuous functions. This form includes all previously mentioned indices save the number of features.

Bonachela et al. proposed their estimator in the form

$$\hat{H}(X) = F\left(\sum_{i=1}^{r} \zeta(X_i)\right),$$

where the function ζ is chosen to minimize

$$\Phi^2_\zeta(p_i) = [\mathsf{E}(\zeta(X_i) - h(p_i))]^2 + \mathsf{var}(\zeta(X_i))$$

possibly weighed by a function $w(p_i)$ when we have some prior knowledge about the distribution of values $p_i \in [0, 1]$. This way, if we disregard the possible influence of the function F and the correlations, we are able to simultaneously reduce the variance and the square of bias of the estimate. The weighted average error is then given by

$$\bar{\Phi}^2_\zeta(p_i) \quad = \int_0^1 \Phi^2_\zeta(p_i) w(p_i) \mathrm{d}p_i. \tag{1}$$

The necessary condition for minimality of the error is a zero value of the derivatives

$$\frac{\delta}{\delta\zeta(k)} \bar{\Phi}^2_\zeta(p_i) = 0, \quad k \in \{0, \ldots, n\}.$$

Therefore, we chose such ζ that

$$\frac{\delta}{\delta\zeta(k)} \int_0^1 \Big[h^2(p_i) - 2h(p_i) \sum_{j=0}^{n} P(X_i = j)\zeta(j) + \\ + \sum_{j=0}^{n} P(X_i = j)\zeta^2(j)\Big] w(p_i)\mathrm{d}p_i = 0$$

which can be simplified to

$$\int_0^1 \Big[\zeta(k)P(X_i = k) - h(p_i)P(X_i = k)\Big] w(p_i)\mathrm{d}p_i = 0.$$

Since

$$P(X_i = k) = \binom{n}{k} p_i^k (1 - p_i)^{n-k},$$

we got

$$\zeta(k) \quad = \frac{\int_0^1 h(p_i) w(p_i) \binom{n}{k} p_i^k (1-p_i)^{n-k} \mathrm{d}p_i}{\int_0^1 w(p_i) \binom{n}{k} p_i^k (1-p_i)^{n-k} \mathrm{d}p_i}. \tag{2}$$

Bonachela et al. [7] derived the form of balanced estimator for Shannon's and power entropies with the weight function equal to 1 on the whole interval $[0, 1]$ of possible values of p_i.

However, if we have zero prior knowledge about the values of the components of vector p, it is natural not to prefer any point from Δ^r. Vector p can be thus viewed

as a realization of a random vector $Y = (Y_1, \ldots, Y_r)$ that has a uniform distribution on Δ^r. The weight function could then be chosen proportional to the expected value of the components p_i, i.e. as a marginal density of random variable Y_i.

This marginal density is proportional to

$$\begin{aligned} f(y_1) &\propto \int_0^{1-y_1} \ldots \int_0^{1-y_1-\ldots-y_{r-2}} \mathrm{d}y_{r-1} \ldots \mathrm{d}y_2 \\ &= \frac{(1-y_1)^{r-2}}{(r-2)!}, \end{aligned}$$

which is (outside a multiplicative constant) a density of Beta distribution $\mathsf{B}(1, r-1)$.

We found the ζ function that minimizes vaha with $w(p_i)$ chosen as $(1-p_i)^{r-2}$ and we derived the corresponding estimator. We called this estimator a β-estimator. We describe the derivation of the β-estimator for Shannon's entropy, i.e. when $h(p_i) = -p_i \ln p_i$ and $F(x) = x$. The symbols Γ, B and Ψ denote the Gamma, Beta and Digamma functions, respectively.

First, we replaced $h(p_i)$ and $w(p_i)$ with the appropriate forms and calculated the integral in the denominator of equation zeta1

$$\zeta(X_i) = \frac{\int_0^1 h(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i}{B(X_i+1, n-X_i+r-1)}.$$

The partial derivative of the Beta function satisfies

$$\frac{\delta}{\delta x} B(x,y) = B(x,y)[\Psi(x) - \Psi(x+y)],$$

and the numerator can be expressed as

$$\begin{aligned} &\int_0^1 h(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i \\ &= -\int_0^1 p_i \ln(p_i) p_i^{X_i} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i \\ &= -\lim_{\alpha \to 0} \int_0^1 \frac{p_i^\alpha - 1}{\alpha} p_i^{X_i+1} (1-p_i)^{n-X_i+r-2} \mathrm{d}p_i \\ &= \lim_{\alpha \to 0} \frac{1}{\alpha} [B(X_i+2, n-X_i+r-1) - \\ &\quad - B(X_i+\alpha+2, n-X_i+r-1)] \\ &= B(X_i+2, n-X_i+r-1)[\Psi(n+r+1) - \\ &\quad - \Psi(X_i+2)]. \end{aligned}$$

Therefore, the ζ function follows

$$\begin{aligned} \zeta(X_i) &= \frac{X_i+1}{n+r}[\Psi(n+r+1) - \Psi(X_i+2)] \\ &= \frac{X_i+1}{n+r} \sum_{k=X_i+2}^{n+r} \frac{1}{k} \end{aligned}$$

and the β-estimator of Shannon's entropy is

$$\hat{H}_1(X) = \sum_{i=1}^{r} \frac{X_i+1}{n+r} \sum_{k=X_i+2}^{n+r} \frac{1}{k}.$$

The β-estimator for power entropies, whose satisfy $F(x) = x$ and $h(p_i) = (\alpha-1)^{-1}(p_i - p_i^\alpha)$, could be derived in a similar manner. With the weight function chosen as $w(p_i) = (1-p_i)^{r-2}$, the β-estimator of power entropies satisfies

$$\hat{H}_\alpha(X) = (\alpha-1)^{-1} \left[1 - \sum_{i=1}^{r} \frac{B(n+r, \alpha)}{B(X_i+1, \alpha)}\right].$$

In Fig. 1 we can see a comparison of the β-estimator, Bonachela's original balanced estimator, the plug-in estimator and the James-Stein shrinkage estimator [8] in a population with 6 possible different features distributed as $p = (24/50, 11/50, 9/50, 3/50, 2/50, 1/50)$. The figures show the sample mean and sample variance computed out of 300 trials. The figures were done in R [9]. Another comparison of sample variance and absolute values of sample bias is presented in Table 1. This time, we compared estimators of $H_{1/2}$ and $H_{5/2}$ when $p_1 = (13, 9, 2, 2, 1)/27$, $p_2 = (12, 8, 5, 3, 3, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1)/46$ and sample size is $n = 50$.

Table 1: Absolute sample bias and sample variance of estimates.

		$H_{1/2}$		$H_{5/2}$	
		|bias|	var	|bias|	var
p_1	plug-in	0.0843	0.0339	0.0094	0.0010
	balanced	0.0059	0.0143	0.0169	0.0008
	beta	0.0052	0.0135	0.0009	0.0006
	shrink	0.0354	0.0211	0.0056	0.0010
p_2	plug-in	0.7476	0.1486	0.0075	0.0002
	balanced	0.2867	0.0281	0.1673	0.0002
	beta	0.1288	0.0218	0.0038	0.0001
	shrink	0.0651	0.0817	0.0052	0.0002

3 Sensitivity to Changes

The indices used to measure diversity differ more or less in their qualities. Their characteristic that is frequently of interest is the rate of the change in value of diversity measure connected to changes in frequencies of a given feature. Several authors dealt with this problem, namely Boyle et al. [10], who were interested mostly in the empirical results, and Izsak [11], who tried to construct a sensitivity measure on a theoretical background. On the suggestion of I. Vajda, we propose a sensitivity measure that is easier to compute and has a clearer interpretation, compared to the Izsak's sensitivity.

Define the sensitivity of diversity measure H to changes in the j-th feature as

$$S_H(p|j) = \lim_{\epsilon \to 0} \frac{H(p_{j,\epsilon}) - H(p)}{\epsilon H(p)}$$

where

$$p_{j,\epsilon} = \frac{(p_1, \ldots, p_{j-1}, p_j + \epsilon p_j, p_{j+1}, \ldots, p_r)}{1 + \epsilon p_j}.$$

This way, the sensitivity of the measure $H(p)$ to changes in p_j is defined as (a limit form of)

$$\frac{\text{relative change of } H}{\text{relative change of } p_j}$$

and reflects the ratio between relative changes of $H(p)$ and p_j when given p_j alters by a small margin.

3.1 Sensitivity of Power Entropies

The derivation of the formula for sensitivity of power entropies was done in Horáček [3]. If all $p_i > 0$, the sensitivity of power entropies satisfies

$$S_{H_\alpha}(p|j) = \alpha \frac{\sum_{i=1}^{r} p_i^{\alpha-1}(p_i - \delta_{ij})}{1 - \sum_{i=1}^{r} p_i^{\alpha}},$$

if $\alpha \neq 1$ and

$$S_{H_1}(p|j) = \frac{\sum_{i=1}^{r} (p_i - \delta_{ij}) \ln p_i}{\sum_{i=1}^{r} p_i \ln p_i}.$$

A comparison of the sensitivity in a population with $p = (24/50, 11/50, 9/50, 3/50, 2/50, 1/50)$ is shown in Fig. 2.

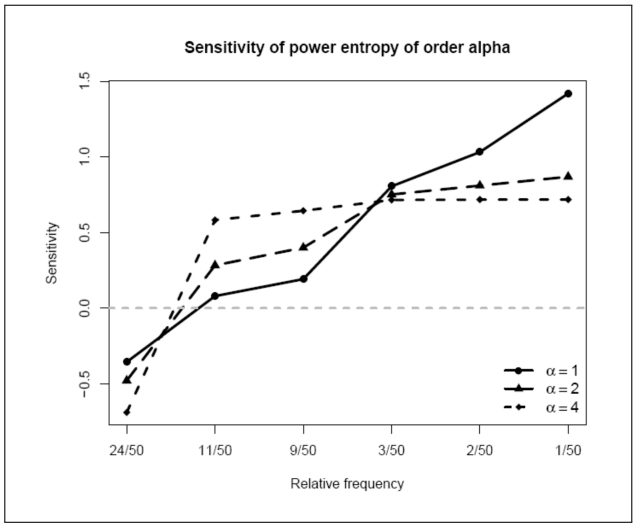


Figure 2: Comparison of the sensitivity to changes - power entropies.

We can see that with decreasing α, the power entropies are more sensitive to the fluctuations in the features that are rare in the population. If we look for example at the sensitivity of Shannon's entropy, say a 10 % increase in $p_5 = 1/25$ would result in about 10 % increase in $H_1(p)$, while a 10 % increase in $p_1 = 24/50$ would result in about 3 % decrease of $H(p)$ and a small change in $p_2 = 11/50$ wouldn't likely change the $H(p)$ much at all.

Acknowledgements

This work was supported by the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic and by the project SVV-2011-262514 of Charles University in Prague.

References

[1] Zvárová J.: On Measures of Statistical Dependence. Časopis pro pěstování matematiky 1974; 99: 15–29

[2] Zvárová J., Vajda I.: On Genetic Information, Diversity and Distance. Methods of Inform. in Medicine 2006; 2: 173–179

[3] Horáček, M.: Measures of biodiversity and their applications. Master thesis, Charles university, Prague, supervisor J. Zvárová 2009

[4] Horáček, M., Zvárová J.: Traditional Measures of Diversity, Their Estimates and Sensitivity to Changes. Proceedings of the 7th Summer School on Computational Biology. 2011; 73–81.

[5] Andrade, M. de, Wang, X.: Entropy Based Genetic Association Tests and Gene-Gene Interaction Tests. Statistical Applications in Genetics and Molecular Biology 2011; 10: Iss. 1, Article 38

[6] Blyth, C. R.: Note on estimating information. Annals of Math. Stat. 1959; 30: 71–79

[7] Bonachela, J. A., Hinrichsen, H., Muñoz, M. A.: Entropy estimates of small data sets. J. of Phys. A: Math. and Theor. 2008; 41: 1–9

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

[9] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org* 2011

[10] Boyle, T. P., Smillie, G. M., Anderson, J. C., and Beeson, D. R.: A sensitivity analysis of nine diversity and seven similarity indices. Research Journal Water Pollution Control Federation 1990; 62: 749–762

[11] Izsak, J.: Sensitivity Profiles of Diversity Indices. Biom. J. 1996; 38: 921–930

Quality Assessment of Fetal Nuchal Translucency Measurements in the First Trimester of Pregnancy

Martin Hynek[1,2,4], **David Stejskal**[1], **Jana Zvárová**[3,4]

[1]Gennet, Center for Fetal Medicine, Prague, Czech Republic

[2]Department of Gynecology and Obstetrics, Thomayer University Hospital, Prague, Czech Republic

[3]Center of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

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

Abstract

Objectives: To evaluate and compare the performance of various quality control methods for nuchal translucency (NT) measurements.
Methods: Fetal NT measurements performed over a one-year period in a single center were used for the study. The retrospective quality review methods proposed by the Fetal Medicine Foundation (FMF) and the Woman & Infants Hospital of Rhode Island (WIHRI) were assessed in the whole dataset and in sonographer-specific distributions. Further prospective statistic process control (SPC) methods were applied (Shewhart $\overline{x}$ and s charts, exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) charts).
Results: Three thousand five hundred and seventy eight NT measurements obtained by seven sonographers were eligible for designed analysis. In the assessment of the sonographer-specific NT distributions three of them did not meet due to the underestimation the FMF and one the WIHRI criteria. Using SPC methods, three sonographers presented unsatisfactory performance with underestimation, three sonographers overall satisfactory performance with transient periods of over- and underestimation and one sonographer showed perfect performance.

MUDr. Martin Hynek

Conclusions: Assessed SPC methods showed close agreement with the retrospective ones, but with the advantage that they can be applied prospectively allowing the prompt action in case of malperformace. The EWMA and CUSUM methods were regarded as the most suitable.

Keywords

Nuchal translucency, control chart, statistical process control, cumulative sum, exponentially weighted moving average, prenatal screening

Correspondence to:

MUDr. Martin Hynek
Gennet, Center for Fetal Medicine
Address: Kostelní 9, 170 00 Praha 7
E–mail: martin.hynek@gmail.com

EJBI 2011; 7(1):22–32
recieved: September 15, 2011
accepted: October 24, 2011
published: November 20, 2011

1 Introduction

Nuchal translucency (NT) is the sonographic appearance of a collection of fluid under the skin behind the fetal neck and can be identified and measured in the first trimester of pregnancy [1, 2]. It has been shown to be the single most effective marker of trisomy 21 and all other major chromosomal abnormalities [3, 4]. NT thickness increases with gestational age (GA) and GA is determined by means of crown-rump length (CRL) measurements. In screening for chromosomal aneuploidies patient-specific risk is derived by multiplying the *a priori* maternal age and gestational-related risk by a likelihood ratio, determined from the deviation of the fetal NT measurement

from the normal median for given CRL. One of the usual approach to quantifying the NT deviation from the normal median, similar to the approach used for laboratory values, is to divide NT measurement by the normal median to produce a multiple of median (MoM) value [5, 6]. In MoM method it is assumed that the distributions of the $\log_{10}$ transformed MoM values in trisomy 21 and unaffected pregnancies are Gaussian and the ratio of the heights of the distributions at a particular MoM, which is the likelihood ratio for trisomy 21, is used to modify the *a priori* maternal age-related risk to produce a patient-specific risk [5, 7]. Using NT, the detection rate for trisomy 21 for a fixed false-positive rate of 3% in screening by a combination of maternal age and fetal NT reaches about 70%, and in screening by maternal age, fetal NT and biochemical markers free β-human chorionic gonadotropin (fβ-hCG) and pregnancy-associated plasma protein-A (PAPP-A) is increased to about 85% [4, 8].

However, NT measurements are displaying the higher variability than biochemical markers due to the lack of automatisation and significant dependance on the operator [9]. Moreover, it has been reported that even minor deviations in NT measurements cause changes in screening efficacy [10]. Thus, to minimize variability and maintain the satisfactory screening performance, it is of a high importance to have clear international guidelines and ongoing quality review programmes established [2, 11].

International technical guidelines specifying the standardized measurement conditions are provided by the Fetal Medicine Foundation (FMF) which have also set up a training programme with a process of accreditation and ongoing quality control [2]. Basically, quality assurance can be qualitative and quantitative [11, 12, 13]. The former one includes image-scoring system, when individual images are reviewed by an expert, and will not be part of this study. The latter involves the comparison of NT measurements to reference values or their distribution assessment.

The first NT quality review was established by the FMF and was based on annual determination of the proportions above and below certain centiles [2]. In 2008, Palomaki et al. [14] from the Women & Infants Hospital of Rhode Island (WIHRI) proposed to use for the quality review the same three epidemiological parameters that have been proven useful in monitoring biochemical markers, ie. median NT MoM, logarithmic standard deviation of NT MoM (SD $\log_{10}$ (NT MoM)) and percent increase in NT thickness per gestational week.

To look to the issue of quality control more globally, there is a wide range of methods for statistical process control (SPC), originally developed in industry to monitor the quality of manufactured products. Dated back to the year 1926, Walter Shewhart, commissioned by Bell Laboratories to improve the quality of telephones manufactured, developed a simple graphical method [15, 16] - the first of subsequently growing range of SPC charts. Since then, these methods have proven very useful and beneficial in industry [15]. Typical Shewhart control chart is the chart with center line (CL) representing the average or target value of the quality characteristic and two control lines, upper control limit (UPL) and lower control limit (LCL), representing the interval within which the quality characteric value should fall with a great probability if the process is 'in control'. In case the values fall outside the control limits, the process is regarded as 'out of control' leading to subsequent investigation of possible causes and corrective actions. The quality characteristic measured by a sample statistic is typically the average $\overline{x}$ ($\overline{x}$ chart), standard deviation s (s chart), range R (R chart) and others [15, 16]. However, these charts are relatively insensitive to small shifts in the process approximately on the order of about 1.5σ or less [15]. In this case suitable alternatives represent special types of control charts, such as the cumulative sum control chart (CUSUM) and exponentially weighted moving average chart (EWMA), which are able to detect even small shifts quickly due to the fact that they do not use only information in the last plotted point as Shewhart charts but all the 'historical' information contained in all previous ones.

In the 1970s the first implementation of SPC methods for analysing medical data was reported [17]. Their power to detect the suboptimal clinical performance has been confirmed in various settings - interventional procedures, general, cardiovascular and thoracic surgery, anaesthesia or ortopedics, namely monitoring the success and complication rate of the procedures, mortality as well as infection rates [18, 19, 20]. The SPC methods have the main advantage of being prospective and therefore allow the early detection of deviation from target performace, with prompt feedback and correction. The most commonly SPC method used in medicine is CUSUM [17]. And it was this very method which was first proposed by Biau et al. [17] for the NT quality review, using deviations in millimeters from the expected NT median. Subsequently, Sabria et al. [9] designed a CUSUM models applying NT deviations in MoMs.

The aim of this study is to evaluate the performance of previously applied methods in NT quality control (FMF, WIHRI, CUSUM) with the extension to Sherhart and EWMA charts, using a real dataset of NT measurements from our center. As the screening policy in our center is based on NT deviations in MoMs, all the assessment will regard MoM-based approches only.

2 Patients and Methods

Patients

The fetal NT measurements performed during a one-year period between July 2010 and June 2011 in the Center for Fetal Medicine Gennet in Prague were used for designed study. From our local database, for each NT measurement we retrieved fetal CRL, date of ultrasound scan and sonographer's name. Only fetuses from singleton pregnancies, without known chromosomal or struc-

tural anomalies and with a CRL between 45 and 84 mm were included in the study. The analysis was restricted to NT measurements between 0.1 and 4.0 mm. Since our center follows the FMF guidelines, as a reference NT median a formula decribed by Nicolaides et al. [21] was used: $\log_{10}\mathrm{NT} = -0.3599 + 0.0127\mathrm{CRL} - 0.000058\mathrm{CRL}^2$, SD $\log_{10}$ (NT MoM) = 0.12. Each NT measurement was converted into NT MoMs and $\log_{10}$ NT MoMs. Assumption of normality of $\log_{10}$ NT MoMs was assessed using normal probability plot. The statistical analysis was performed using the statistical computing environment `R` (R Development Core Team 2010) [22] and the additional `R` packages `qcc` and `iqcc` for quality control charts [23, 24].

Ultrasound examination

The first-trimester ultrasound scans were performed at GA of 10 weeks$^{+6\ \mathrm{days}}$ - 13 weeks$^{+6\ \mathrm{days}}$ by seven sonographers, labeled A to G, of which sonographers A-F are the FMF accredited and undergoing the regular annual FMF audits and the sonographer G has not been accredited yet. The patients were alocated to the sonograprahers randomly.

The ultrasound machines Voluson E8 with 4-8-MHz 3D/4D transabdominal probe and Vivid 7 with 7-MHz transabdominal probe (both General Electric Medical Systems, Kretztechnik GmbH & Co, Austria) were used for all examinations. Fetal NT thickness was measured according to the current FMF guidelines [2]:

- the fetal CRL between 45-84 mm,
- the magnification of the image such that fetal head and thorax occupy the whole screen,
- the strict mid-sagittal view of the face, defined by the presence of the echogenic tip of the nose and rectangular shape of the palate anteriorly, the translucent diencephalon in the centre and the nuchal membrane posteriorly,
- the fetus in the neutral position,
- the widest part the translucency measured placing callipers on the inner border of the line that defies the NT.

Example of such an image is presented in Figure 1.

FMF and WIHRI methods

Two quality review methods based on NT measurement distributions were applied to the whole dataset and to each particular sonographer. The FMF method [2] includes the calculation of the proportion of NT measurements above and below the expected median (expected to be 50%, the acceptable range $40-60\%$), above the expected 95^{th} centile and below the expected 5^{th} centile (expected to be 5% with the acceptable range $4-6\%$).

WIHRI method [14] examines the median NT MoM (expected to be 1.0, acceptable range $0.90-1.10$), the SD $\log_{10}$ (NT MoM) (expected $0.08-0.14$) and the rate of increase of NT thickness with advancing gestational age (expected to be 20% per week, acceptable range $15-35\%$).

Statistical process control methods

Subsequently, selected SPC methods were applied to our dataset. This was done for each particular sonographer because the interpretation of the whole dataset would be difficult.

The quality characteric of interest is $\log_{10}$ NT MoMs which is assumed to be normally distributed. For designing of control charts we need to known the target process mean μ and standard deviation σ corresponding to the 'in control' process and on the basis of which CL, UCL and LCL are constructed. [15]

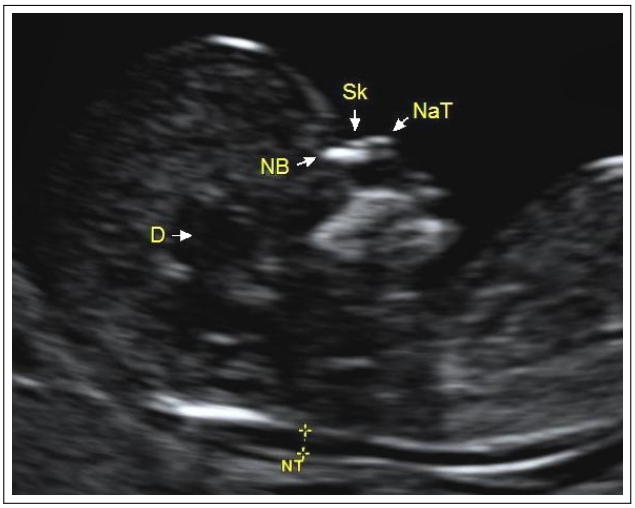


Figure 1: Nuchal translucency measurements according to the Fetal Medicine Foundation protocol. NT, nuchal translucency, Sk, skin, NaT, nasal tip, NB, nasal bone, D, diencephalon.

For this, our target mean is $\mu = 0$. The acceptable range outside which the process is considered to be 'out of control' was set to $0.90 - 1.10$ NT MoM, the same range used by Palomaki et al. [14] in WIHRI study and by Sabria et al. [9] and based on the extensive knowledge of the impact of inaccuracy on Down syndrome risk estimates using serum markers [25]. The NT MoM interval corresponds to the interval $\log_{10}$ NT MoM $\subset (-0.0458; 0.0414)$. However, having two different values mean that we are considering differently a process mean shift upward and downward and this is not consistent with the basic principles of SPC [15]. As underestimation is by far more common for NT measurements resulting in a drop in the detection rate of screening test, we decided to set up the acceptable $\log_{10}$ NT MoM interval symetrically to ± 0.0458, corresponding to $0.90 - 1.11$ NT MoM, which is entirely clinically acceptable.

The SD $\log_{10}$ (NT MoM) = 0.086, derived from our whole dataset, was markedly lower compared to the 13-year-old Nicolaides's one ($\sigma = 0.12$) [21]. However, if we

used the latter as a reference value, this would lead to control limits which are too loose and do not reflect the real situation. Moreover, if we refer to the literature, we will find that during the years the observed SD is gradually declining: 0.12 in 1998 (dataset $n = 95476$) [21], 0.105 in 2007 ($n = 23462$) [14], 0.079 in 2008 ($n = 38791$) [7]. We believe that the main reason for this fact is the rapid technical improvement of ultrasound equipment with higher resolution and better pre- and postprocessing technologies as well as better training and certification of sonographers participating in regular audits. Therefore, we decided to use as estimate of σ our SD $\log_{10}$ (NT MoM) = 0.086.

The performance of control charts is evaluated using the average run length (ARL). ARL0 (average run length under the null hypothesis) represents the number of procedures before an extreme value caused by natural variability of the 'in control' process is interpreted as being 'out of control' (corresponds to Type I error). Contrary, ARL1 (average run length under the alternative hypothesis) is the number of procedures which chart shows to be 'in control' despite the process being actually 'out of control' (corresponds to Type II error) [15, 16]. Naturally, the aim is to set control limits to minimize the ARL1 while maximizing the ARL0.

Shewhart control chart To control the process mean and variability we used the combination of Shewhart $\overline{x}$ and s control charts. Subgroups of N samples are used to calculate the sample mean $\overline{x}$ and standard deviation s and are succesively plotted into the diagram together with CL, UCL and LCL. In our case each sample consists of all measurements since the last sample was taken. The use of subgroup samples has the advantage that the variation of sample mean is by $\sqrt{N}$ lower than the variation in the population from which sample group comes from. Considering $\overline{x}$ chart, the CL and control limits are placed at [15, 16]

$$UCL = \mu + k\frac{\sigma}{\sqrt{N}} \quad (1)$$

$$CL = \mu \quad (2)$$

$$LCL = \mu - k\frac{\sigma}{\sqrt{N}} \quad (3)$$

where k is the distance of the control limits from the center line, expressed in standard deviation units. Usually $k = 3$ is selected for the probability of 0.9973 that 'in control' process will fall within these limits or $k = 2$ for the probalility of 0.9544. Considering our acceptable interval being ± 0.0458 and $\sigma = 0.086$, the minimal necesarry sample size N is from the formula 1: $0.0458 \geq 0 + k(0.086/\sqrt{N})$. Thus, for $k = 3$ is $N = 32$, which considering the average number of ultrasound scans would mean too long control interval. A more rational choice is $N = 15$ for $k = 2$, being aware that the Type I error is nearly 5%.

The limits for s chart are placed at [15]

$$UCL = c_4\sigma + k\sigma\sqrt{1 - c_4^2} \quad (4)$$

$$CL = c_4\sigma \quad (5)$$

$$LCL = c_4\sigma - k\sigma\sqrt{1 - c_4^2} \quad (6)$$

where c_4 is the bias correction constant for the sample standard deviation statistic defined as [16]

$$c_4 = \frac{\Gamma(\frac{N}{2})\sqrt{\frac{2}{N-1}}}{\Gamma(\frac{N-1}{2})} \quad (7)$$

where $\Gamma(.)$ is the gamma function. For $N = 15$ is $c_4 = 0.9823$. Similarly we selected $k = 2$.

To enhace the sensitivity of control charts Shewhart proposed a set of run rules to help to detect nonrandom patterns. In our study we used his rule of seven consecutive points plotted on one side of the center line, having the probability of accidental occuring $p = 0.5^7 = 0.0078$ [15, 23].

Exponencially weighted moving average chart Contrary to the Shewhart charts where the decision signal obtained depends largely on the last point plotted, using EWMA (sometimes also called 'moving geometric mean') charts the importance of various extent is given to all the previous points. An 'exponencially weighted mean' is calculated each time a new result becomes avalaible [26]:

$$Z_t = \lambda x_t + (1 - \lambda)Z_{t-1} \quad (8)$$

where λ is smoothing coefficient and $0 < \lambda \leq 1$, and the starting value of EWMA at time $t = 0$ is $Z_0 = \mu$ (the process target). The usually $\lambda = 0.2$ [26] or $\lambda = 0.25$ [16] are selected. $\lambda = 1$ corresponds to Shewhart control chart and the lower the λ, the lower is the reaction of Z_t to local changes in monitored process and the better is the tendency to emphasize systematic long-term changes [16]. Control limits are placed at [16]:

$$UCL = \overline{x} + k\frac{\sigma}{\sqrt{\lambda/(2-\lambda)}} \quad (9)$$

$$LCL = \overline{x} - k\frac{\sigma}{\sqrt{\lambda/(2-\lambda)}} \quad (10)$$

Due to the approximately normal distribution of Z_t, the choice of k is similar to the choice in Shewhart charts [16]. For EWMA we used $\pm 3\sigma$ limit. Since variability of NT measurements is significatly higher than is our 'in control' interval, we set $\lambda = 0.05$ in order to maximally eliminate local changes and to pick up systematic shift in the process. Using simulated random series of data we verified the eligibility of our settings.

Cumulative sum chart Similarly to EWMA, the CUSUM charts utilize all the information contained in-

side previous measurements. The CUSUM test computes, at each time t, a score S_t defined by

$$S_t = \max(0; S_{t-1} + W_t) \tag{11}$$

with $S_0 = 0$ and W_t the sample weight. W_t is a measure of the deviation of the observation from the target. At each t the CUSUM tests the null hypothesis that the process is 'in control' against the alternative one that the process is 'out of control'; practically it happens, if S_t is equal or greater than a decision limit h, the null hypothesis is rejected and the process is regarded to be 'out of control'. Until then the null hypothesis cannot be rejected and the process is considered to be 'in control' [17].

A modern CUSUM control charts use the cumulative sum of standardized deviations from target mean μ. Consider a standardized variable z_t [16]:

$$z_t = \frac{x_t - \mu}{\sigma} \tag{12}$$

and two cumulative sums $S_{H,t}$ to detect positive shift and $S_{L,t}$ to detect negative shift:

$$S_{H,t} = \max[0; (z_t - K) + S_{H,t-1}], \tag{13}$$
$$S_{L,t} = \max[0; (-z_t - K) + S_{L,t-1}] \tag{14}$$

where starting values $S_{H,0} = S_{L,0} = 0$. K is the reference value defined as $K = \delta/2$, where δ is the size of the shift one wants to detect in multiples of σ. The usual choice of K is $0.5 - 2.0$ $(1 - 4\sigma)$ [15, 16, 26].

Control limits are controlled by decision interval h. Usual choice is $h = 4$ or $h = 5$ providing a CUSUM that has good ARL properties against a shift of about 1σ. However, for different δ Biau et al. [17] recommends that h is best determined by simulation and compute different ARL0 and ARL1 values by varying h to obtain acceptable compromise between a very responsive test (short ARL1) and too many false alarms (short ARL0).

For our setting $\delta = \Delta/\sigma = 0.0458/0.086 = 0.533$. The decision interval h was determined using simulation. Thirty series of 20000 random measurements from a normal distribution without any deviation from the expected mean were generated to establish the ARL0 at different h limits. A futher thirty series of 1000 random measurements from normal distribution with the mean located at ± 0.0458 were simulated to estimate the median ARL1. The h was then selected to provide the best trade-off between early deviation detection and minimum number of false alarms.

3 Results

In total, 3578 NT measurements were eligible for designed analysis. The normal probability plot of $\log_{10}$ NT MoMs did not violate our assumption of normality. Figure 2 shows the whole data of NT measurements in mm plotted against CRL together with regressed medians expected from the reference distribution and our observed medians. The observed NT median is slightly below the FMF median, suggesting the overall trend towards underestimation in our dataset.

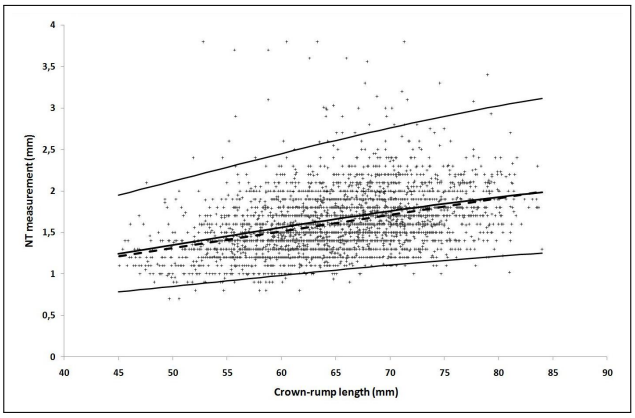


Figure 2: Nuchal translucency (NT) plotted against crown-rump length. ——— expected median and 5^{th} and 95^{th} centiles of the reference distribution; ------- observed median of our measured NT MoMs.

When FMF and WIHRI methods were applied to our whole dataset (see Table 1), the FMF percentage under the median showed underestimation only slightly below the acceptable lower range. However, the proportion below 5^{th} and above 95^{th} centile were markedly lower than expected, suggesting significantly lower variability within our dataset. All three WIHRI requirements were met. SD $\log_{10}$ NT MoM from our dataset of 0.086 is indeed markedly lower than σ of 0.12 in the FMF reference serie, thus explaining the results of FMF review method.

Regarding the selection of suitable decision interval h, the results of simulated random series of measurements are summarized in the Table 2. Value of $h = 10$ was selected, providing the estimated median number of measurements needed for a false alarm about 1500, and the median number of measurements needed to detect minimum desired deviation at 28.

The results obtained by the five quality review methods - ie. the two retrospective (FMF and WIHRI) and three SPC methods (Shewhart, EWMA and CUSUM charts) - applied to the seven particular sonographers are summarized in the Table 1. The three selected figures with Shewhart, EWMA and CUSUM charts demostrate three usual performance patterns: unsatisfactory performace (sonographer A in Figure 3), overall good performance with only temporal changes (sonographer C in Figure 4) and finally perfect performance (sonographer F in Figure 5).

Looking more thoroughly to the results, the sonographer A fulfilled the three WIHRI criteria having the median NT MoM at satisfactory 0.95 MoMs suggesting only slight underestimation. However, having 69.8% of measurements below median is slightly below lower FMF acceptable limit. A markedly lower expected number of

Table 1: Results of quality assessment methods of fetal nuchal translucency (NT) measurements.

Method	*Whole dataset*	*Sonographer*							*Acceptable range*
		A	*B*	*C*	*D*	*E*	*F*	*G*	
Number of cases	3578	576	893	641	496	541	320	111	
FMF									
< median (%)	**61.2**	**69.8**	56.2	59.8	49.4	**72.6**	51.6	**83.8**	40-60
> 95th centile (%)	**1.2**	**0.9**	**1.0**	**1.6**	**2.4**	**0.9**	**0.6**	**0.0**	4-6
< 5th centile (%)	**1.6**	**0.9**	**0.3**	**1.1**	**1.4**	5.4	**0.9**	**3.6**	4-6
WIHRI									
Median NT MoM	0.95	0.93	0.97	0.96	1.01	0.90	0.99	**0.83**	0.90-1.10
SD $\log_{10}$ NT MoM	0.086	0.077	0.081	0.084	0.094	0.086	0.087	0.075	0.08-0.13
Weekly increment (%)	19.9	18.4	15.2	26.5	19.0	28.4	21.3	**5.6**	15-35
Shewhart $\overline{x}$ chart									
Points beyond upper limit (%)	-	0	0	0	9.1	0	0	0	
Upper points violating runs (%)	-	0	0	0	0	0	0	0	
Points beyond lower limit (%)	-	31.6	6.8	2.4	3.0	58.3	4.8	100.0	
Lower points violating runs (%)	-	57.9	11.9	7.1	0.0	69.4	0.0	14.3	
Shewhart s chart									
Points beyond upper limit (%)	-	2.6	3.4	4.8	9.1	2.8	0	0	
Upper points violating runs (%)	-	0	0	0	0	0	0	0	
Points beyond lower limit (%)	-	7.9	3.4	0	3.0	5.6	0	0	
Lower points violating runs (%)	-	7.9	6.8	0	0	0	0	0	
EWMA chart									
Upper limit crossings (n)	-	0	0	0	1	0	0	0	
Lower limit crossings (n)	-	20	5	7	4	13	0	1(POL)‡	
Points beyond limit (%)	-	29.0	3.1	3.3	3.6	59.5	0.0	92.8	
CUSUM chart									
Upper limit crossings (n)	-	0	0	0	1	0	0	0	
Lower limit crossings (n)	-	1(POL)⋆	3	5	3	2(POL)†	0	1(POL)‡	
Points beyond limit (%)	-	84.0	7.4	10.0	11.3	67.0	0.0	90.0	

The values lying outside the acceptable range are in bold print. ⋆Out of limit from 93^{rd} measurement. †Out of limit from 177^{th} measurement. ‡Out of limit from 9^{th} measurement in EWMA and 12^{th} measurement in CUSUM.
NT, nuchal translucency, MoM, multiples of median, FMF, Fetal Medicine Foundation method, WIHRI, Women and Infants Hospital of Rhode Island method, EWMA, exponentially weighted moving average, CUSUM, cumulative sum. POL, persistently out of limits.

Table 2: Estimated median number of nuchal translucency (NT) measurements needed until false alarm or deviation detection occurs, for different decision intervals h, calculated from 30 simulated random series of measurements.

	Decision interval h						
	8	*9*	***10***	*11*	*12*	*13*	*14*
First false alarm	323	512	**1490**	3249	3421	6105	6947
± 0.0458 $\log_{10}$ NT MoM mean deviation detection	22	22.5	**28**	30	31	38.5	48

Our final choice of h is in bold print. NT, nuchal translucency, MoM, multiples of median.

measurements below the FMF 5^{th} and above 95^{th} centiles is present in all sonographers with only one exception in the proportion of below 5^{th} centile measurements of the sonographer E and will not be further commented. As mentioned previously it is related to the significantly lower variability in our series compared to the FMF one. In relation to the SPC methods (see Figure 3), Shewhart $\overline{x}$ chart shows underestimation starting from 6^{th} sample group onward and manifesting either by points below LCL or points violating the run rule. The $27^{th} - 29^{th}$ and further 31^{st} and 33^{rd} sample groups seem to be 'in control'. The EWMA chart produces a line fluctuating around CL up to the 93^{rd} point when it falls below LCL marking the underestimation and staying below until approximately 200^{th} measurement. Further course is more or less 'in control' with several temporal periods of underestimations. The CUSUM chart crosses the lower limit at the 93^{rd} measurement further gradually bottoming with occasional periods of horizontal course, corresponding very well to the periods in which EWMA shows the 'in control' line. Here it has to be pointed out that the horizontal course of CUSUM line indicates that the observations are scattered around target value although the line is located outside control limits. Shewhart s chart shows but one point above UCL. As the underestimation in the s chart representing the lower variability leads to an increase in the screening performance and is thus beneficial, we will not further comment on such results.

In the case of the sonographer B the FMF and WIHRI criteria were fulfilled. The SPC charts displayed the lines oscilating within the limits apart from the final observed period, when $\overline{x}$ chart presented 'out of control' state from 53^{rd} sample group, the EWMA chart from around 817^{th} and more profoundly from 872^{nd} and the CUSUM from 817^{th} observations onward. The other 2 points (12^{th}

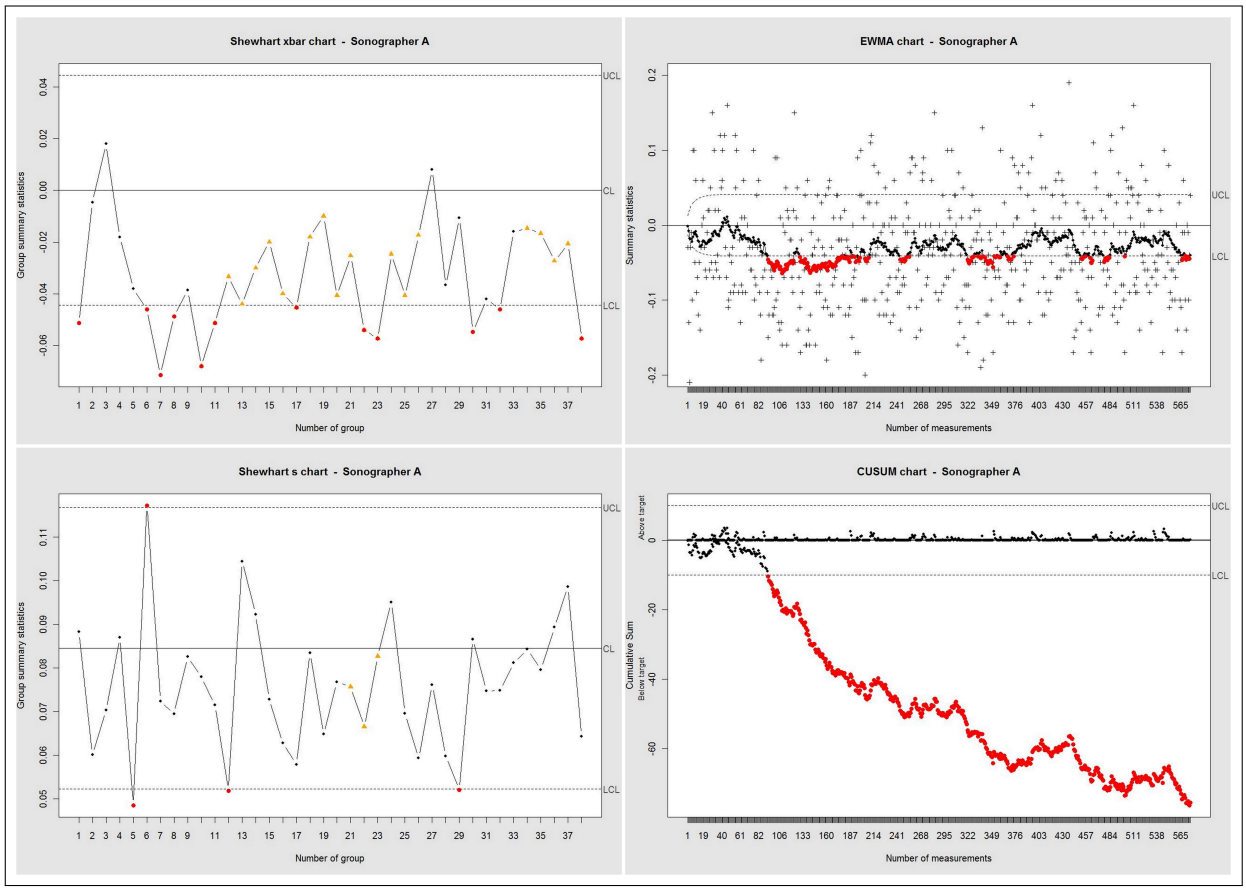


Figure 3: Shewhart $\overline{x}$ and s charts, exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) charts for dataset of sonographer A. • point within control limits, • point outside control limits, △ point violating runs. UCL, upper control limit, CL, center line, LCL, lower control limit.

and 26^{th}) in $\overline{x}$ diagram falling below LCL had the corresponding manifestations in EWMA and CUSUM diagrams where the lines almost touched the LCL. The s chart showed 2 points (3.4%) above the UCL.

Sonographers C (see Figure 4) and sonographer D proved the similar performance. Both satisfactory complied with the FMF and WIHRI limits. However, SPC methods revealed the periods of temporal 'out of control' state. As for the sonographer C the EWMA diagram presented the underestimation of the $38^{th} - 51^{st}$, $129^{th} - 135^{th}$ and $305^{th} - 308^{th}$ NT measurements. Corresponding response at CUSUM chart had the line sinking below LCL between 40^{th} and 75^{th}, 128^{th} and 142^{nd}, 305^{th} and 320^{th} measurements. The sonographer D overperformed between 236^{th} and 246^{th} and between 233^{rd} and 260^{th} measurements looking at EWMA and CUSUM charts respectively and underperformed between 440^{th} and 454^{th} according to the EWMA and between 435^{th} and 468^{th} measurements according to the CUSUM charts. In both sonographers the Shewhart $\overline{x}$ chart displayed in all these mentioned cases of suboptimal performance corresponding points beyond control limits. The proportion of points in s chart suggesting an increase in variability were 4.8% (2 points) and 9.1% (3 points) regarding the sonographer C and D, respectivelly.

The sonographer E failed the FMF criteria due to the underestimation, but the WIHRI parameters were within, but close to, the lower limit. Shewhart $\overline{x}$ chart suggests the underestimation because from 12^{th} sample group onward all point are either below LCL or violating the run rule. The same pattern can be seen at the other two diagrams. The EWMA line crossed the LCL at 173^{rd} measurement further caterpillaring almost horizontally below LCL with only several very transient returns above LCL. The CUSUM presented a persistent decrement of the lower line from the 177^{th} observation, demonstrating similar response to underestimation. There was only one point above UCL in the s chart.

The sonographer F (see Figure 5) perfectly met the FMF and WIHRI criteria and in concordance with this all SPC charts presented lines rocking between control limits. The only exception was the last 21^{st} sample group falling below LCL at Shewhart $\overline{x}$ chart, which on one hand could be regarded as false positive signal, on the other hand it may be indication of the beginning of 'out of control' as both EWMA and CUSUM diagrams show a line bottoming down to LCL and we know that these charts although more sensitive than Shewhart ones react more slowly.

Contrary to the sonographer F, the sonographer G displayed totally unacceptable performance. There was a

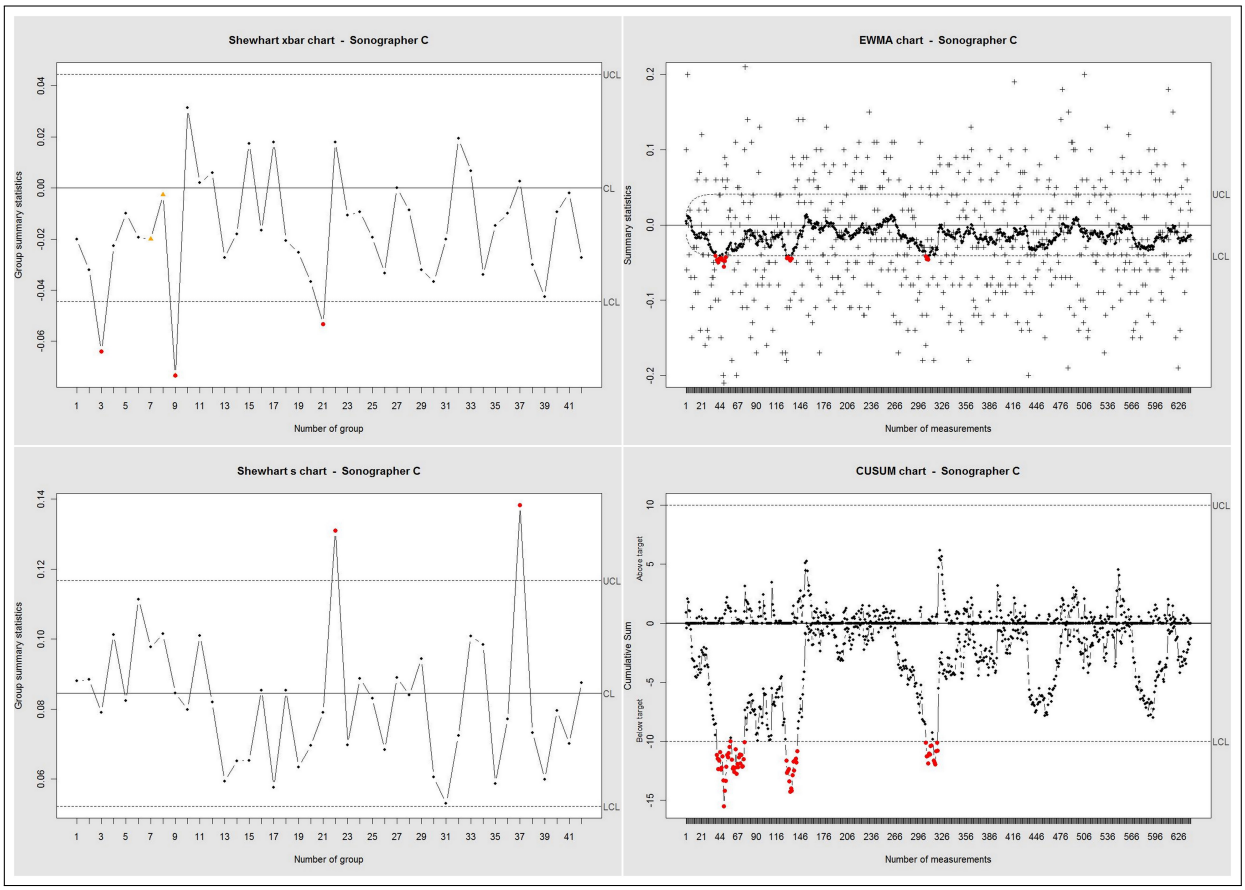


Figure 4: Shewhart $\overline{x}$ and s charts, exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) charts for dataset of sonographer C. • point within control limits, • point outside control limits, △ point violating runs. UCL, upper control limit, CL, center line, LCL, lower control limit.

heavy underestimation in FMF and WIHRI criteria. At the same time all $\overline{x}$, EWMA and CUSUM charts highlighted the pronounced underestimation, in $\overline{x}$ diagram from the very 1^{st} sample group and from the 9^{th} and 12^{th} in EWMA and CUSUM charts respectivelly presenting us how rapidly they are able to expose 'out of control' process. S chart complied within the limits. It has to be mentioned that the sonographer G has not been accredited by the FMF yet as she is still in the phase of learning the ultrasound scanning.

4 Discussion

The contribution of NT as a marker for chromosomal abnormalities is indisputable. Strict quality standards for NT measurements are difficult to follow even for a well-trained and experienced sonographers in tertiary centers [27].

Due to the fact that even minor deviations in NT measurements may have an impact on screening effectiveness (eg. underestimation by 25% may lead to the sensitivity decrease by 1.1%) [10], an establishing of an ongoing audit for NT screening is of a paramount importance [11, 28].

Quality review methods based on distribution parameters (FMF, WIHRI) are easily implemented due to the rather simple methodology. However, the main disadvantages include being retrospective and dealing with all measurements altogether. They are usually performed on the annual basis. Therefore, at the time of audit, a certain number of mothers may have been offered invasive testing unnecessarily (in the case of overestimating) being endangered by the risk of miscarriage which invasive procedures pose or on the other hand some fetuses with Down syndrome may have not been detected prenatally (in case of underestimating). Using the data altogether may lead to undetecting temporal, but significant changes as could be seen in our study for example in the sonographer C (underperforming between 40^{th} and 75^{th}, 128^{th} and 142^{nd}, 305^{th} and 320^{th} measurements) or similarly in the sonographer D (overestimating the $233^{rd} - 260^{th}$ measurements, underestimating the $435^{th} - 468^{th}$ measurements) but both perfectly fullfiling the criteria of WIHRI. Equally, it has to be pointed out that both methods need a critical minimum number of measurements in order not to be influenced by extreme values.

The presented SPC methods do not have these limitations. They can be used prospectively for ongoing NT audit. Prospectiveness represents an important advantage

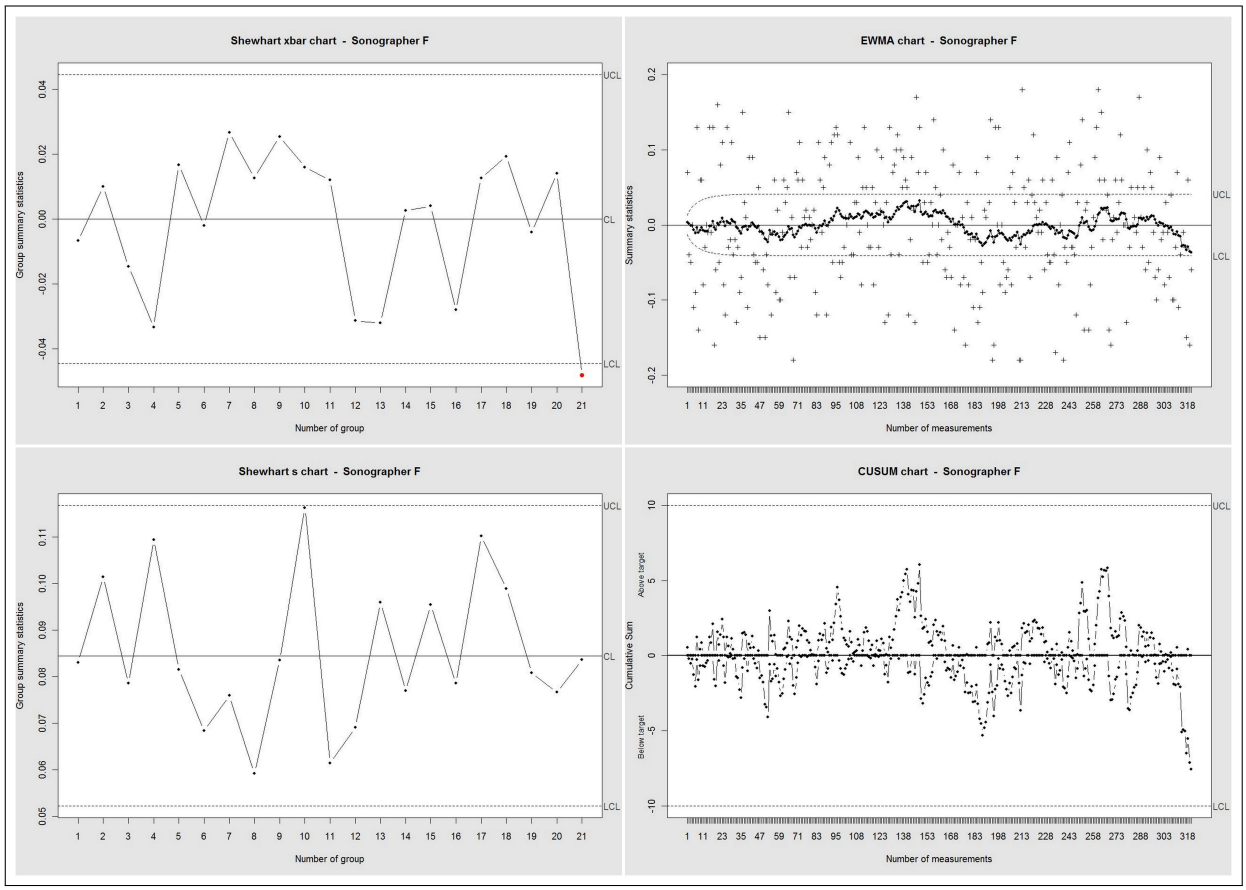


Figure 5: Shewhart $\overline{x}$ and s charts, exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) charts for dataset of sonographer F. • point within control limits, • point outside control limits, △ point violating runs. UCL, upper control limit, CL, center line, LCL, lower control limit.

as it allows the early detection of deviation from the target with prompt feedback and correction.

Shewhart control chart have not been to our best knowledge used for NT quality review yet. They are able to control the process accuracy (mean) as well as its precision (variability) showing their change in two separate diagrams. The methodology is simple and straightforward. They are more sensitive to rapid changes within a proces [26], however, relatively insensitive to small shifts on the order of about 1.5σ or less. To try to overcome these drawbacks, we can implement various kinds of run rules (Shewhart rules, Western Electric rules, etc.), yet making the interpretation more difficult. We observed that implementation of one Shewhart run rule has lead to better ability to detect 'out of control' state.

The other option is to increase the size of sample group. Nonetheless, in our case the desired 'in control' interval is about half the standard deviation and the necesarry sample size would be too large considering the average number of ultrasound scans performed by individual sonographer. Increasing the sample size would inevitably mean too long time interval when a subsequent point is added into control diagram. Our suggestion is to lower the control limits to $\pm 2\sigma$ allowing to have reasonable sample size but at the expense of increasing the risk of false positive signals at the same time, again making the interpretation more difficult.

EWMA charts have also not been reported to be used for NT quality review. Contrary to the Shewhart charts, EWMA gives the importance of various extent to previous measurements and is able to detect smaller shifts relatively quickly. The amount of weight attributed to the older data and to the last point is controlled by the smoothing parameter λ. We observed that an appropriate λ for NT quality review is 0.05 leading to the low reaction of the chart to local changes and very good ability to emphasize systematic long-term shifts. We have experienced that EWMA charts are capable to reveal the process being 'out of control' very effectively having a very low false positivity rate at the same time (0.27% using $\pm 3\sigma$ control limits).

Comparing the order of measurements from which the process was considered to be 'out of control' on the basis of EWMA and CUSUM charts the results were almost identical. One of the 'visual' advantage we have noticed is that when the 'out of control' process represented by the curve outside control limits is rectified to 'in control' state the EWMA curve returns back towards center line within the limits in contrast to CUSUM charts where the increasing or declining curve of 'out of control' process switches

to horizontal course remaining in the steady distance from the center line.

CUSUM charts have been repeatedly proved to be advantageous in medicine owing to its simple formulation and very intuitive graphical representation plotting two separate lines, the upper monitoring overestimation and the lower underestimation. It is a very sensitive method for small sustained changes which is able to detect quickly. The CUSUM method has already been proposed as a suitable prospective method for NT quality review; Biau et al. [17] suggested delta-based approach, while Sabria et al. [9] presented the MoM-based one. As most of the fetal medicine centers are using the MoM-based approach for calculationg the risk for Down syndrom, the MoM-based CUSUM model is in our opinion more appropriate. Similarly to conclusion of Sabria et al. [9], we have observed that CUSUM represents a very effective and sensitive technique for NT quality review, definetely superior to retrospective FMF and WIHRI methods. It allows to determine the exact time when inaccurate measurements start to occur allowing to find possible causes (eg. change in ultrasound equipment, in scanning routine, etc.) and take corrective actions.

5 Conclusion

In our study we have observed that prospective regular NT quality review is of a crude importance. SPC methods represent a powerful tool for this quality proces control. They allow a prospective evaluation with a graphically very instructive output. In our opinion the most suitable methods for NT quality control using MoM-based approach are CUSUM as well as EWMA charts as both have the ability to detect the process being 'out of control' very quickly, effectively and with low false positivity rate allowing for prompt correction of the technique when required.

Acknowledgements

The work was supported by the grant SVV-2011-262514 of Charles University in Prague. Authors would like to thank to Dr. Luca Strucca from Faculty of Economics, University of Perugia, the author of the `qcc` package for `R` for his advice with the quality control charts and `qcc` package.

References

[1] Nicolaides KH, Azar G, Byrne D, Mansur C, Marks K. Fetal nuchal translucency: ultrasound screening for chromosomal defects in first trimester of pregnancy. BMJ 1992; 304: 867-869.

[2] Fetal Medicine Foundation website. URL http://www.fetalmedicine.com/fmf/training-certification/certificates-of-competence/11-13-week-scan/nuchal/. [Accesses 1 September 2011]

[3] Snijders RJ, Noble P, Sebire N, Souka A, Nicolaides KH. UK multicentre project on assessment of risk of trisomy 21 by maternal age and fetal nuchal-translucency thickness at 10-14 weeks of gestation. Lancet 1998; 352: 343-346.

[4] Kagan KO, Wright D, Valencia C, Maiz N, Nicolaides KH. Screening for trisomy 21, 18 and 13 by maternal age, fetal nuchal translucency, fetal heart rate, free β-hCG and pregnancy-associated plasma protein-A. Hum Reprod 2008; 19: 1968-1975.

[5] Spencer K, Bindra R, Nix ABJ, Heath V, Nicolaides KH. Delta-NT or NT MoM: which is the most appropriate method for calculating accurate patient-specific risks for trisomy 21 in the first trimester? Ultrasound Obstet Gynecol 2003; 22:142-148.

[6] Nicolaides KH, Snijders RJ, Cuckle HS. Correct estimation of parameters for ultrasound nuchal translucency screening. Prenat Diagn 1998; 18:519-523.

[7] Wright D, Kagan KO, Molina FS, Gazzoni A, Nicolaides KH. A mixture model of nuchal translucency thickness in screening for chromosomal defects. Ultrasound Obstet Gynecol 2008; 31: 376-383.

[8] Kagan KO, Wright D, Baker A, Sahota D, Nicolaides KH. Screening for trisomy 21 by maternal age, fetal nuchal translucency thickness, free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-A. Ultrasound Obstet Gynecol 2008; 31: 618-624.

[9] Sabria J, Barcelo-Vidal C, Arigita M, Jimenez JM, Puerto B, Borrell A. The CUSUM test applied in prospective nuchal translucency quality review. Ultrasound Obstet Gynecol 2001; 37: 582-587.

[10] Kagan KO, Wright D, Etchegaray A, Zhou Y, Nicolaides KH. Effect of deviation of nuchal translucency measurements on the performance of screening for trisomy 21. Ultrasound Obstet Gynecol 2009; 33: 657-664.

[11] D'Alton ME, Cleary-Goldman J, Lambert-Messerlian G, Ball RH, Nyberg DA, Comstock CH, Bukowski R, Berkowitz RL, Dar P, Dugoff L, Graigo SD, Timor IE, Carr SR, Wolfe HM, Dukes K, Canick JA, Malone FD. Maintaining quality assurance for sonographic nuchal translucency measurement: lessons from the FASTER Trial. Ultrasound Obstet Gynecol 2009; 33: 142-146.

[12] Snijders RJM, Thom EA, Zachary JM et al. First-trimester trisomy screening: nuchal translucency maesurement training and quality assurance to correct and unify technique. Ultrasound Obstet Gynecol 2002; 19: 353-359.

[13] Wojdemann KR, Christiansen M, SUndberg K, Larsen SO, Shalmi A, Tabor A. Quality assessment in prospective nuchal translucency screening for Down syndrome. Ultrasound Obstet Gynecol 2001; 18: 641-644.

[14] Palomaki GE, Neveux LM, Donnenfeld A, Lee JE, McDowell G, Canick JA, Summers A, Lambert-Messerlian G, Kellner LH, Yebelman A, Haddow JE. Quality assessment of routine nuchal translucency measurements: A North American laboratory perspective. Genet Med 2008; 10: 131-138.

[15] Montgomerry DC, Runger GC. Applied statistics and probability for engineers. 3^{rd} ed. United States of America: John Wiley & Sons, Inc.; 2002.

[16] Meloun M, Militký J. Kompendium statistického zpracování dat. Praha: Academia; 2006.

[17] Biau DJ, Porcher R, Salomon LJ. CUSUM: a tool for ongoing assessment of performance. Ultrasound Obstet Gynecol 2008; 31: 252-255.

[18] Biau DJ, Resche-Rigon M, Godiris-Petit G, Nizard RS, Porcher R. Quality control of surgical and interventional procedures: a review of the CUSUM. Qual Saf Health Care 2007; 16: 203-207.

[19] Noyez L. Control charts, cusum techniques and funnel plots. A review of methods for monitoring performance in healthcare. Interact Cardiovasc Thorac Surg 2009; 9(3): 494-9.

[20] Sibanda T, Sibanda N. The CUSUM chart method as a toll for continuous monitoring of clinical outcomes using routinely collected data. BMC Med Res Methodol 2007; 7: 46.

[21] Nicolaides KH, Snijders RJM, Cuckle HS. Correct estimation of parameters for ultrasound nuchal translucency screening. Prenat Diagn 1998; 18: 519-523.

[22] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[23] Scrucca, L. Qcc: an R package for quality control charting and statistical process control. R News 2004; 4(1): 11-17.

[24] Recchia DR, Barbosa EP., Goncalves EJ(2010). IQCC: Improved Quality Control Charts. R package version 0.5. URL http://CRAN.R-project.org/package=IQCC. [Accesses 1 September 2011]

[25] Knight GJ, Palomaki GE. Epidemiologic monitoring of prenatal screenign for neural tube defects and Down syndrome. Clin Lab med 2003; 23: 531-551.

[26] Oakland JS. Statistical process control. 5^{th} ed. Cownwall: MPG Books Limited; 2003.

[27] Ville Y. Semi-automated measurement of nuchal translucency thickness: blasphemy or oblation to quality? Ultrasound Obstet Gynecol 2010; 36: 400-403.

[28] Evans MI, Van Decruyes H, Nicolaides KH. Nuchal translucency measurements for the first-trimester screening: the 'price' of inaccuracy. Fetal Diagn Ther 2007; 22: 401-404.

Comparison of EuroMISE Minimal Data Model for Cardiology and HL7 V3 DAM: Cardiology Rel. 2

Libor Seidl[1,2], **Petr Hanzlíček**[1]

[1] Centre of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic

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

Abstract

Background: The EuroMISE Minimal Data Model for Cardiology (MDMC) has been prepared by clinicians for clinical study in 2002. This model has been successfully implemented in an application for clinical data gathering. HL7 v3 Domain Analysis Model: Cardiology, Release 2 (HL7 DAM) has been published in HL7 September 2011 Ballot.
Objectives: The objective of this paper is to compare these two data models. The main motivations for the comparison are nearly identical ways of development, and the same format of both specifications.
Methods: HL7 DAM is much broader than EuroMISE MDMC. Thus I focus only on data elements present in MDMC but absent in HL7 DAM. Also different scales of elements present in both models are compared.
Results: I have found 25 elements out of 181 elements defined in MDMC which are not contained in HL7 DAM.

Ing. Libor Seidl

Conclusions: Results will be used for further discussion in HL7 Clinical Interoperability Council work group.

Keywords

Data model, EuroMISE MDMC, HL7 V3 DAM Cardiology, comparison

Correspondence to:

Ing. Libor Seidl
Institute of Computer Sciences,
Academy of Sciences of the Czech Republic
Address: Pod Vodarenskou vezi 2, Prague 8, Czech Republic
E–mail: seidl@euromise.cz

EJBI 2011; 7(1):34–37
recieved: November 1, 2011
accepted: November 8, 2011
published: November 20, 2011

1 Introduction

Standardization of information content of interfaces, EHR, or EMR is a huge challenge of nowadays biomedical informatics. According to The Generic Component Model [1] building of standardized healthcare IT environment consists of enterprise view, analysis of information content, computational view, engineering view and technology view.

In this article I focus on effort of standardization of a cardiology domain from the information point of view. I compare two data models, a previous work of my colleagues (EuroMISE MDMC) and a current release of a cardiology domain analysis model made by HL7 (HL7 DAM). This work will bring highlights and contributions to the further analysis of the cardiology domain. Both subjects of comparison are described in following chapters.

1.1 EuroMISE MDMC

The Department of Medical Informatics of the Institute of Computer Science AS CR, a part of the EuroMISE Centre, focuses mainly on the applications of advanced statistical methods, on the analysis of biomedical data and knowledge, on utilization of the structured electronic health record, methods for decision support and on data mining in biomedical databases [2].

The EuroMISE Minimal Data Model for Cardiology (MDMC) was prepared by EuroMISE clinicians for clinical study in 2002 [3]. This data model has been successfully adopted by system designers and implemented into an ADAMEK [4], a single-purpose application for clinical data gathering. In following years there was also an effort to map particular data elements of MDMC to ICD 10 and SNOMED-CT [5].

No special software tools were used in 2002 to capture the model. There is an Excel Spreadsheet where rows are data elements, columns are element attributes (name of the element, group, data type or enumeration, and extra attributes). The MDMC is set of 181 data elements (rows). Each element has an associated data type or an enumeration of possible values. Where it makes sense, additional attributes are also filled: data format, minimal value, maximal value, objectivity, reliability, importance, and economic aspect. Elements are also divided into groups according to parts of Czech medical record as defined by a Czech law:

- administrative dataset,
- family history (RA),
- social anamnesis (SA),
- personal anamnesis (OA),
- current difficulties (OB),
- therapy (Th.),
- physical examination,
- laboratory results (Labor.).

1.2 HL7 V3 DAM: Cardiology, Rel. 2

Health Level Seven International is a global authority on standards and interoperability of healthcare information technology. The development is based on voluntary work of experts around the world [6]. A HL7 Clinical Interoperability Council is a HL7 work group. It has developed HL7 V3 Domain Analysis Model: Cardiology, Release 2 (HL7 DAM). The specification has been submitted to a HL7 September 2011 Ballot Cycle [7].

The DAM contains the information analysis of clinical content related to the Acute Coronary Syndrome domain, heart failure, electrophysiology, vascular analysis and intervention. It also contains observations related to general cardiology. The DAM involves use cases, activity diagrams and UML class diagrams; Enterprise Architect is used as a maintaining tool. In total there is approximately 350 data elements which were adopted or harmonized from following sources [8]:

- the NCDR registries (National Cardiovascular Data Registry),
- the ACC/AHA Adult CV EHR Data Elements (American College of Cardiology/American Heart Association),
- the CDISC (Clinical Data Interchange Standards Consortium) standards Acquisition Harmonization elements,
- the FDA Cardiovascular Endpoint Data Elements,
- the NCI EVS Vocabulary (National Cancer Institute),
- other published cardiology standards (AHA, STS).

The ACC/AHA data standards cover broad areas within cardiology including outpatient care, heart failure, atrial fibrillation and cardiac imaging. The NCDR CATH-PCI registry is used for patients undergoing diagnostic and interventional catheterization; the ACTION registry is used for patients experiencing acute coronary syndrome, and the ICD registry is used for patients experiencing electrophysiology disturbances and heart failure. The STS Adult Cardiac Registry is used for patients undergoing cardiac surgeries [8].

As authors declare: "It is not intended for products to be derived from this model or directly implemented. Additional technical specifications and system requirements are necessary for implementation" [8].

2 Objectives

The main objective of my work is to find out whether MDMC is a subset of HL7 DAM. If it is not a subset the problem will appear in a future: ADAMEK will not be able to transmit all gathered data to another system using standardized HL7 messages. It will generate additional costs on our side. It is not a good idea to be data-locked with the only one software, even with our own.

If HL7 DAM is missing some elements, there should be an action taken, probably to join and participate with the Clinical Interoperability Council workgroup. In this case, this paper will serve as a starting point for discussions.

The second objective of this paper is to bring a short introduction of HL7 DAM to an attention of readers who are not familiar with HL7. Citizens of the Czech Republic have high average risk factor of cardiovascular disease. Our government invests lot of money to a Cardiovascular research [9]. Information Technology perspective of such research should be in concordance to the international state of the art.

3 Methods

Although definition of MDMC is captured by simple technology in contrast to HL7 DAM captured in Enterprise Architect, although the size of HL7 DAM is 3 times bigger than MDMC, these sets of data elements are very similar and easily comparable. I used a simple and

straight method of comparison: for each data element in MDMC I search for an equivalent in HL7 DAM. When found, the data type (or the enumeration) is compared, too.

Due to only a brief description of each data elements in HL7 DAM and no description of element in MDMC, there is no space for semantic discrepancies.

4 Results

According to the methodology, all data elements of MDMC were examined. MDMC data elements without a HL7 DAM equivalent can be divided into clinically significant discrepancies and elements of administrative purpose.

Clinically significant elements missed in HL7 DAM:

- SA: overall psychological stress faced (none, low, middle, high),
- SA: physical activity at work (none, low, middle, high),
- SA: physical activity at home (none, low, middle, high),
- SA: smoker (how many cigarettes per day),
- SA: alcohol (beer, wine, distillates),
- OA: body temparature,
- OA: body mass index (BMI),
- Th.: hypertension treatment (none, life style, antihypertensive agents),
- Th.: dyslipidemia treatment (none, life style, hypolipidaemic agents),
- Th.: Peripheral Arterial Disease therapy (none, conservative, PTA, stent, surgical),
- Th.: Renal Artery Disease therapy (none, conservative, PTA, stent, surgical).

Other missed elements of HL7 DAM can be seen as administrative, technical, or superfluous. These elements are probably out of scope of Clinical Interoperability Council because other attributes of HL7 V3 RIM [10] classes hold the information, or they are defined in other HL7 V3 domains. Values stored in those elements are still significant from clinical and informational point of view. Elements missed in HL7 DAM:

- patient: administrative gender,
- RA: family history of father, mother, brothers (elements of every clinical term have been assigned, but the attribute of family relationship is missing),
- SA: patient marital status,
- SA: patient education,
- AL: drug allergy (name of a drug, chemical name),
- OA: date of first observation of a disease (every indicator of HL7 DAM),
- Labor.: laboratory results (glycemia, uric acid, total cholesterol, HDL/LDL, triacylglyceride).

5 Discussion and Conclusions

When talking to semantic interoperability, I usually see lots of academic discussions how far proposed 2 concepts are identical, similar, or completely misleading. This gets much worse when concepts are not in the same language, which was my case, too (Czech vs. English).

Fortunately, HL7 DAM provides a brief description of each concept, whereas MDMC is just a list of data elements with scales but without any description. At this level of information, there were no such discussions and mapping has been done smoothly.

I have mapped 181 data elements of MDMC to HL7 DAM elements. I have identified 12 clinically significant elements missed in HL7 DAM. The relevance of these elements should be discussed with the HL7 Clinical Interoperability Council. I have also identified 13 elements of MDMC missed in HL7 DAM, but found somewhere else in HL7 V3. From the information point of view (in terms of Generic Component Model [1]) both groups are significant at the same level and require attention. The latter group can be easily addressed in terms of HL7 V3 classes.

I see one other huge source for contributions to the HL7 DAM. The Czech Medical Association of J.E. Purkyně, Cardiology section provides 40 textual Clinical Practice Guidelines listed in a catalogue of Czech published guidelines [11]. Therefore it seems that the Cardiology section of the Czech Medical Association has the sufficient credit and ability to contribute to HL7 DAM.

Acknowledgements

This work has been supported by Charles University in Prague grant no. SVV-2011-262-514 and by Czech Ministry of Education, Youth and Sport grant no. 1M06014.

References

[1] Diego M. Lopez, Bernd G.M.E. Blobel. A development framework for semantically interoperable health information systems. International Journal of Medical Informatics. 2009; 78: 83—103.

[2] EuroMISE *Homepage* [Internet]. 2011 [cited 2011 Oct 30]; Available from: http://www.euromise.org

[3] Tomeckova M et al. Minimal data model of cardiological patient. (in Czech). Cor et Vasa 2002; 4: 123.

[4] Mares R, Tomeckova M, Peleska J, Hanzlicek P, Zvarova J. User interface for patients' database systems – an example of

application for data collection using minimal data model of cardiological patient (in Czech). Cor et Vasa 2002; 4: 76.

[5] Petra Přečková. Language of Czech Medical Reports and Classification Systems in Medicine. European Journal for Biomedical Informatics. 2010; 6/1: 58–65.

[6] Health Level Seven International *Homepage* [Internet]. 2010 [cited 2011 Oct 30]; Available from: http://www.HL7.org

[7] Health Level Seven International *HL7 Version 3 Standard Ballot Site - September 2011* [Internet]. 2011 [cited 2011 Oct 30]; Available from: http://www.hl7.org/v3ballot/html/welcome/introduction/index.html

[8] Health Level Seven International, Clinical Interoperability Council. Cardiovascular Domain Analysis Model, Release 2 (HL7 V3 Standard Ballot Site - September 2011. 2011; page 4–8

[9] Czech Ministry of Health, *Ministry of Health Departmental Program for Research and Development III* [Internet]. 2010 [cited 2011 Oct 30]; Available from: http://iga.mzcr.cz/shareIGA/External_web/DPR_III.doc

[10] Hleath Level Seven International *Reference Information Model (RIM) Downloads* [Internet]. 2011 [cited 2011 Oct 30]; Available from: http://www.hl7.org/implement/standards/rim.cfm

[11] EuroMISE, *Katalog klinických doporučených postupů* [Internet, Czech only]. 2011 [cited 2011 Oct 30]; Avaiable from: http://neo.euromise.cz/kkdp

Biometric Methods for Applications in Biomedicine

Anna Schlenker[1,2], **Milan Šárek**[3]

[1] Center of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

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

[3] CESNET z.s.p.o.

Abstract

Objectives: The aim of this paper is to analyze current state of use of biometrics in computer security.
Methods: This paper provides an overview of the most commonly used anatomical-physiological and behavioral biometric identification methods.
Results: The result of the work will be a new set of methods, which allows reliable identification of the user in the most comfortable way.
Conclusions: These new principles of data security will be used to enhance the protection of specialized health record. This will contribute to expansion of generally conceived EHR MUDR concept to other application areas.

Ing. Anna Schlenker

Keywords

Biometrics, data security, EHR (electronic health record), fingerprints, hand geometry, face recognition, iris recognition, retinal scanning, keystroke dynamics, multi-factor authentication

Correspondence to:

Ing. Anna Schlenker
Center of Biomedical Informatics, ICS, AS CR, v.v.i.
Address: Pod Vodarenskou vezi 2, Prague 8, Czech Republic
E–mail: schlenker.anna@gmail.com

EJBI 2011; 7(1):37–43
recieved: September 15, 2011
accepted: October 24, 2011
published: November 20, 2011

1 Introduction

Biometrics, biometric identification and verification have been investigated since the early 80's of the last century. At the end of the 20th century first applications began to emerge, especially in forensic practice where biometrics was represented by automated processing of fingerprints and palm prints found at a crime scene.

Nowadays, biometric methods are irreplaceable both in the forensic sciences and in commercially available applications.

In this paper we analyze current state of use of biometrics in computer security, especially the possibilities of identification based on biometric data. Biometric characteristics can be divided into anatomical-physiological and behavioral [1, 2].

2 Anatomical-Physiological Biometric Characteristics

The most frequently used anatomical-physiological biometric characteristics in common practice are fingerprints, palm prints, geometry of hand shape and scanning of bloodstream patterns of the palm or the back of ones hand.

2.1 Fingerprints and Palm Prints

Fingerprints and palm prints are based on the uniqueness of ridge patterns [3]. Miniaturization of sensors and processors allows the fingerprint-based biometric identification for large commercial use.

In practice, fingerprints are often used for authentication of persons accessing to computers or communication devices, for enhancement of protection of identification or

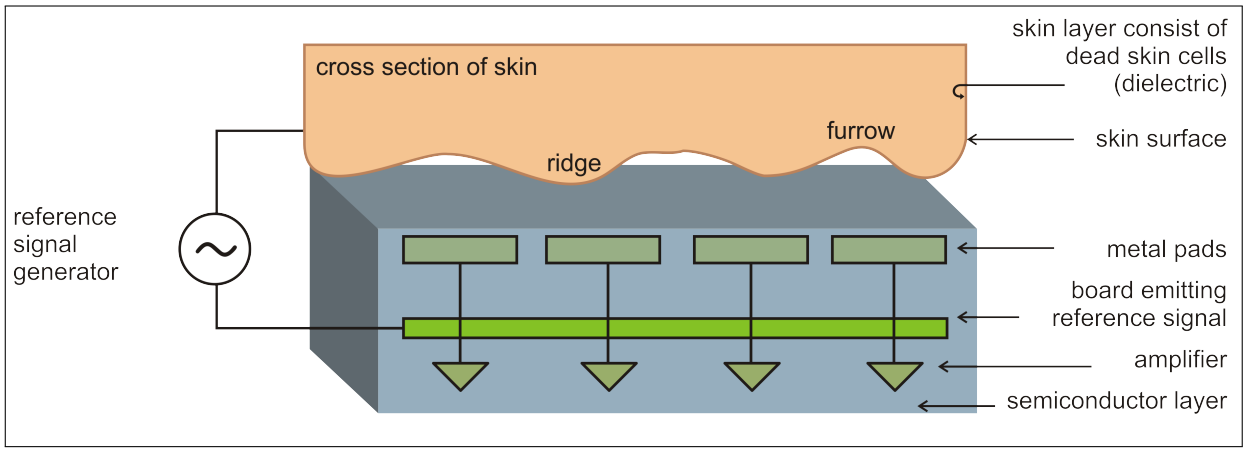


Figure 1: Simplified diagram of the electronic sensors (according to [5]).

credit cards, for authorization to access buildings and for protection of precious or dangerous devices from unauthorized use.

Interactive fingerprinting, which is now often implemented in a variety of technical equipment, is done by means of sensors. These sensors may be contact or contactless and their functions can be based on different physical principles [4].

2.1.1 Contact fingerprint sensors

Contact sensors include optical, electronic, optoelectronic, capacitive, pressure and temperature sensors. Some of these sensor types will be described in detail below. Main advantages and disadvantages of each method are clearly shown in Table 1.

Optical contact sensors Optical sensors are based on FTIR technology (Frustrated Total Internal Reflection). This means that a laser beam illuminates the bottom surface of a finger that touches a transparent sensor plate. Reflected light flux is then captured by a CCD (Charge-Coupled Device) element. The amount of reflected light depends on the depth of papillary lines and furrows. Papillary lines reflect more light than furrows.

Other optical sensors use a thick bundle of optical fibers that are perpendicular to the plane of the sensor. Here again, the method of exposure and reflection of light flow is applied. Another type of sensors uses CMOS technology (Complementary Metal-Oxide-Semiconductor).

Electronic contact sensors Electronic sensors operate on the principle of electric field between two parallel, conductive and electrically charged plates (see Figure 1). If the shape of the originally flat plate on top changes to wavy (papillary lines and furrows), the shape of the electric field changes too. The upper plate of the sensor is represented by surface of the skin that is connected to the source reference electrical signal. The main advantage of this sensor is that it does not scan only the surface of the skin but it scans deeper skin layers too. This means that this type of sensor is resistant to dirt and possible damage of the skin surface.

Optoelectronic contact sensors Optoelectronic sensors consist of two layers. The upper layer is in contact with the skin and it is able to emit light. This light is captured in the second glass layer in which photodiodes are sealed. These photodiodes convert the light into an electrical impulse.

Capacitive contact sensors Capacitive sensors capture fingerprint by measuring electrical capacity (see Figure 2). Scanning sensor is composed of a large number of scanning surfaces that are isolated from each other. By touching the sensor, papillary lines bridge the conductive pads while furrows act as isolators. The shape of papillary drawing, therefore, modulates voltage and capacitance drops between the conductive pads. These drops are measured and they form a digitalized picture of papillary drawing.

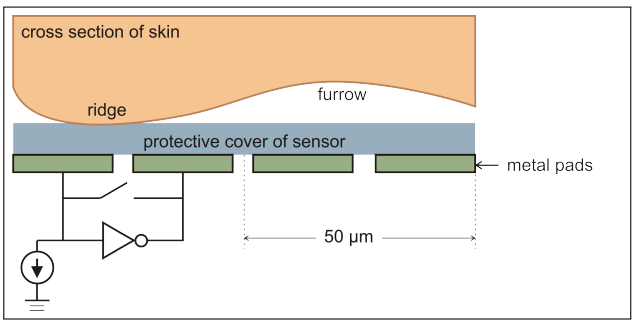


Figure 2: Simplified diagram of the capacitive sensors (according to [5]).

These sensors are highly nonresistant to various types of dirt that may significantly alter conductivity of the skin.

Pressure contact sensors Pressure sensors respond to a pressure of papillary lines on the surface of sensor. The

sensor surface is made of an elastic piezoelectric material that transforms the pressure into an electrical signal and thus creates a picture of fingerprint.

Temperature contact sensors Temperature sensors react to temperature differences between papillary lines and furrows. A great advantage of this sensor is that temperature is an important factor that can tell whether the scanned fingerprint belongs to a living person.

2.1.2 Contactless sensors for fingerprint

The best-known groups of non-contact sensors include optical and ultrasound sensors. The advantages and disadvantages of these sensors are also included in Table 1.

Optical non-contact sensors The principle of optical non-contact sensors is similar to the optical contact sensors described above with only one difference. The beam of light allows scanning from a distance of 3-5 cm.

The greatest advantage of this sensor is that it prevents contamination caused by contact with dirty fingers.

Ultrasonic non-contact sensors Ultrasonic sensors are based on a similar principle as the optical ones but instead of a light beam a beam of short mechanical waves (ultrasound) is being reflected from the skin surface (see Figure 3). This type of sensor eliminates all the disadvantages of previous types of sensors explained above [6].

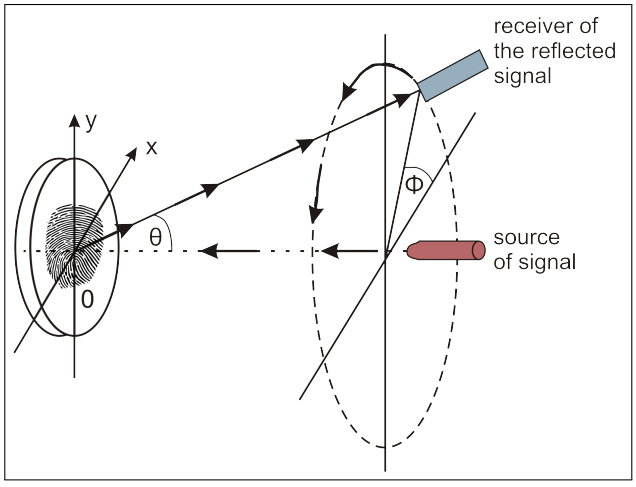


Figure 3: Simplified diagram of the ultrasonic sensors (according to [5]).

2.2 Geometry of a Hand Shape

Another frequently used method is the geometry of a hand shape, the essence of which is measurement of lengths and widths of fingers, bones or joints of the hand [7] (see Figure 4). The hand touches a horizontal scanner that has special fixation pins. These ensure that the hand is always in the same position. The scanner captures one image from the top (perpendicularly to the sensor board) and one image from the side. This generates two monochrome images of 'hand silhouette'.

At first, a user requiring evidence of his/her identity enters his or her identification number (PIN) via keyboard or he or she touches a magnetic stripe, a chip or a card to a reader. Then the user puts his or her hand to a specified position according to visual instructions that are on keyboard on the scanner [8]. Hand geometry scanners are now common in many areas including healthcare.

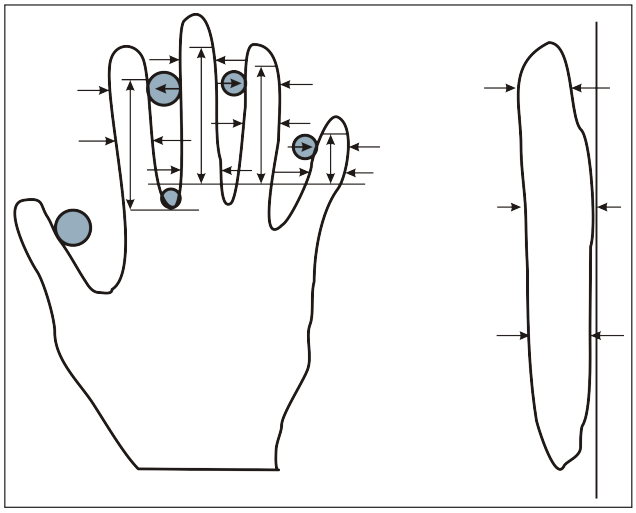


Figure 4: The basic principle of hand geometry (according to [5]).

2.3 Scanning of the Bloodstream of the Palm or the Back of Hand

Another method suitable for use in computer security is scanning of the bloodstream of the palm or the back of ones hand. A CCD camera, which is most commonly used in this case, takes a picture of the hand and a specific pattern of blood vessel distribution captured in the image is then used to identify the person.

An unquestionable advantage of this method is that it also verifies whether the tested object is alive. The scanning runs in infrared band which is sensitive to temperature. This method takes advantage of the fact that blood vessels in the body are warmer than their surroundings. The scanned image is further processed in a similar way as fingerprint (with the shape of vessels being compared).

Another advantage in comparison to scanning of hand geometry is that it is not necessary to place a hand in the scanner in the same position every time.

Other options for this method are to scan the bloodstream of the palm or to perform non-contact scanning of both the palm and the back of hand, which provides a high level of hygiene unlike hand geometry scanning or fingerprints [8].

2.4 Scanning of a Face and its Parts

Instead of hands a face or a part of the face can be used to identify a person as well. There are computer programs that can recognize the human faces like human brain does [9]. Face recognition is now typical especially in criminology and there are many different methods and algorithms used for these purposes.

This method can also be easily used to secure common computing and telecommunication systems. Any standard video camera, which can be already found integrated in many screens, is sufficient to take the image of the face. The face scan can replace traditional password entry.

A great advantage of this method is that there is absolutely no need for direct contact between the user and the sensor [10].

However, face recognition can be further improved in many ways. As an example, we can register signs of emotions.

An interesting application of this method in IT security suggests itself. Continuous face scanning during the work with computer would make it possible to evaluate whether it is still the same person accessing sensitive data. Not only that this method secures the system at the time of login, it can even protect the data later on, when the authorized user, for example, leaves the unlocked terminal for a period of time.

2.5 Scanning of the Iris or Retina

Recently, thanks to its simple implementation using only conventional video systems, scanning of the iris or retina is becoming a more widespread method of identification. Iris recognition is possible regardless of size, location and orientation but it requires a complicated algorithm [11]. This method is, therefore, usually used only to ensure a high level of security [8].

A light beam is used to map the bloodstream in the retina [12]. A part of the beam is absorbed by the retina while the other part is reflected. The special camera, that is required for the scanning, is expensive and the scanning process itself is not very user-friendly (many people are afraid of the technology) [8].

3 Behavioral Biometric Characteristics

Keystroke dynamics could be an interesting behavioral biometric characteristic for use in computer security not being widely used so far.

3.1 Keystroke Dynamics

Keystroke dynamics allows so-called continuous (dynamic) verification, which is based on the use of keyboard as a medium of continuous interaction between user and computer [13]. This offers a possibility of continuous control over the whole time the computer is being used. This method is useful in situations when there is a risk of leaving a computer without control for a while [14].

The most common characteristic is the time of pressing individual keys or the duration of the keypress. Another possibility is to measure typing speed, frequency of errors, style of writing capital letters or a force used to press the keys. This latter type requires a special keyboard that allows the force of the push to be measured. All other methods can be evaluated by a special program without any modification of hardware [15, 16].

Table 1: Comparison of contact and non-contact fingerprint sensors.

Sensor	Advantages	Disadvantages
Optical contact sensors	– very quick – user-friendly	– not resistant to dirt – not hygienic – don't recognize living tissue
Electronic contact sensors	– resistant to dirt – very quick – user-friendly	– not hygienic – don't recognize living tissue
Capacitive contact sensors	– very quick	– not resistant to dirt – don't recognize living tissue – not hygienic
Temperature contact sensors	– recognize living tissue – very quick	– not hygienic
Optical non-contact sensors	– resistant to dirt – hygienic – very quick	– don't recognize living tissue
Ultrasonic non-contact sensors	– resistant to dirt – hygienic – very quick	– don't recognize living tissue

Table 2: Comparison of anatomical-physiological and behavioral biometric characteristics.

Sensor	Advantages	Disadvantages
Geometry of hand shape	– resistant to dirt	– don't recognize living tissue – require scanning in the same position – not hygienic
Contactless scanning of bloodstream	– don't require scanning in the same position – recognize living tissue – hygienic – resistant to dirt	– no possibility of continuous control
Scanning of the face	– resistant to dirt – recognize living tissue – don't require scanning in the same position – possibility of continuous control	– time-consuming
Scanning of iris	– resistant to dirt – don't require scanning in the same position – user-friendly	
Scanning of retina	– resistant to dirt – don't require scanning in the same position	– not user-friendly – time-consuming
Keystroke dynamics	– user-friendly – possibility of continuous control – hardware-efficient	

4 Comparison of the Methods

Most of current data security systems verify user's authorization to access the system only at the time of login. In the case that the question of user identification is solved only on the basis of biometric data, only one biometric component (or just a few of them) is used for verification in most cases.

An optimal solution should preferentially include the methods mentioned in the introduction and emphasize those of them, which will prove themselves long-time stable or the least disturbing for staff. The method must be fast enough for the user. Hardware requirements and required processing power will also be considered.

Table 1 shows the main advantages and disadvantages of different types of contact and contactless sensors for fingerprinting. All sensors for fingerprinting are relatively quick and easy in comparison to other biometric methods.

The main differences are in resistance to dirt, which is important for the following two reasons.

- The first one is that the sensor should be able to work even when there is dirt on its surface or on the surface of the finger that is being scanned.
- The second reason is, of course, the hygienic aspect.

The greatest benefit is sensor's ability to distinguish living tissue from dead or synthetic material. Then it becomes very resistant to possible abuse.

Table 2 displays main advantages and disadvantages of other anatomical-physiological and behavioral characteristics. Besides the aspects mentioned above, we compared also the possibility of continuous authentication, the need for scanning in the same position and difficulty/ease of use.

Table 3: Comparison of methods in terms of stability of biometric characteristics and time-consuming.

Method	Stability of biometric characteristics	Time-consuming
	high = more than 80 %, medium = more than 60 %, low = less than 60 %	high = more than 3 sec, medium = less than 3 sec, low = less than 1 sec
Fingerprint	medium	low
Geometry of hand shape	medium	medium
Scanning of bloodstream	medium	medium
Scanning of the face	low	high
Scanning of iris	high	medium
Scanning of retina	high	high
Keystroke dynamics	low	low

Table 3 compares selected methods in terms of stability of biometric characteristics and time-consumption.

Stability of biometric characteristics means how much it changes over time. For example a human face can significantly change, either naturally or under the influence of a disease. In contrast to face, the retina does not change and contains features unique for each person.

Data in the table are not accurate readings but empirical estimates. The table shows that there is no method that would be "ideal", i.e. would offer high stability of biometric characteristics and low time consumption. Iris scanning, which is currently not used in everyday practice, is close to this ideal.

5 Application of Selected Methods in Electronic Health Record Security

The aim of this work is to propose a multifactor system that will verify a number of biometric features simultaneously, thus ensuring greater reliability of identification [17]. This will protect access to patient data in electronic health record, which is conceptually based on the proposal of Universal Electronic Health Record MUDR, see [18].

Security of patient data is one of the key issues in telemedicine. It may appear that this is a standard solution using the principles of electronic record EHR MUDR. But unlike our task, the concept of EHR MUDR record is designed with respect to ordinary patient data, accessed during everyday hospital operation.

Contrastingly, in the case of the electronic record of personal identification ERPI, there will be much more sensitive data related to the identification of individuals from different perspectives. For this reason there is also a demand for higher level of identification of persons accessing the data.

With regard to the nature of such data it appears necessary to use some set of DLP (Data Loss Prevention) allowing identification of the risks associated with the loss of sensitive data and possible dynamic reduction of these risks. Moreover, with regard to the type of sensitive identification data it is useful to have a resource that will allow consecutive audit of the data.

Commercial solutions such as RSA [19, 20] or Websense [21] are available. These sets are designed to reduce the impact of potential risks, irrespective of whether the data are stored in the datacenter, transmitted over the network (network DLP) or processed in user terminal equipment (DLP endpoint). This solution is particularly interesting because in the Czech Republic there has not yet been a deployment of DLP published in similar context.

6 Conclusion

The result should be a complex of new biometric identification methods that would allow reliable identification of users in the most comfortable form. The final application of these new principles of security will increase the level of protection of specialized health record. Furthermore, the EHR MUDR concept will expand to other application areas.

Acknowledgements

This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project 1M06014 and by the project SVV-2011-262514 of Charles University in Prague.

References

[1] Jain AK, Ross A. Introduction to Biometrics. In Jain AK, Flynn P, Ross A. Handbook of Biometrics. Springer, 2008. pp. 122. ISBN 978-0-387-71040-2.

[2] Denning DE. Cryptography and Data Security. Addison-Wesley, 1982. ISBN 0-201-10150-5.

[3] Herschel WJ. The Origin of Finger-Printing. Oxford University Press, 1916. ISBN 978-1104662257.

[4] Cravotta N. Looking under the surface of fingerprint scanners. EDN. 2000 [cited 2011 Aug 22]. Available from: http://www.edn.com/article/507025-Looking_under_the_surface_of_finger_print_scanners.php

[5] Rak R, Matyáš V, Říha Z. Biometrie a identita člověka: ve forenzních a komerčních aplikacích. Grada, Praha; 2008.

[6] Bicz W et al. Fingerprint structure imaging based on an ultrasound camera. 2005 [cited 2011 Aug 22]. Available from: http://www.optel.com.pl/article/english/article.htm

[7] Jain AK, Ross A, Pankanti S. A prototype hand geometry-based verification system. In Second International Conference on Audio and Video-based Biometric Person Authentication. Washington DC, USA, 1999. pp. 166171.

[8] Jain A, Bolle R, Pankarti S. Biometrics: personal identification in networked society. Kluwer Academic Publisher 1999. Norwell, Massachusetts, USA.

[9] Brunelli R, Poggio T. Face Recognition: Features versus Templates. IEEE Trans. on PAMI, 1993. (15)10:1042-1052.

[10] Zhang D. Automated biometrics: technologies and systems. Kluwer Academic Publisher, Boston; 2000.

[11] Daugman J. How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology, 2004. 14(1):2130.

[12] Lichanska A. Retina and Iris Scans. Encyclopedia of Espionage, Intelligence, and Security. The Gale Group, Inc. 2004.

[13] Bergadano F, Gunetti D, Picardi C. User authentication through Keystroke Dynamics. ACM Transactions on Information and System Security (TISSEC), 2002. 5(4):367-397.

[14] Gunetti D, Pikardi C. Keystroke analysis of free text. ACM Transactions on Information and System Security 2005. 8(3):312-347.

[15] Ilonen J. Keystroke Dynamics. Advanced Topics in Information Processing 2003 [cited 2011 Aug 22]. Lappeenranta University of Technology. Available from: http://www2.it.lut.fi/kurssit/03-04/010970000/seminars/Ilonen.pdf

[16] Monrose F, Rubin D. Keystroke dynamics as a biometric for authentication. Future Generation Computer Systems 2002. 16(4):351-359.

[17] Badr Y, Chbeir R, Abraham A, Hassanien AE (Eds.) Emergent Web Intelligence: Advanced Semantic Technologies. 1st Edition, 2010, XVI, 544.

[18] Hanzlicek P, Spidlen J, Nagy M. Universal electronic health record MUDR. Studies in health technology and informatics, Amsterdam: IOS Press 2004, 105:190-201.

[19] Young D. RSA Adaptive Authentication for Healthcare Environments, online: http://www.rsa.com/products/consumer/datasheets/10037_AAHC_DS_0611.pdf

[20] RSA, The Security Division of EMC: Security Solutions for Business Acceleration [Internet]. 2010 [cited 2011 Aug 22]. RSA Data Loss Prevention (DLP) Suite. Available from: http://www.rsa.com/node.aspx?id=3426

[21] Websense Security Survey: Security Pros and Cons, online: http://www.websense.com/content/websense-security-survey-security-pros-and-cons.aspx

Stochastic Models in the Identification Process

Dalibor Slovák[1], **Jana Zvárová**[1,2]

[1]Center of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

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

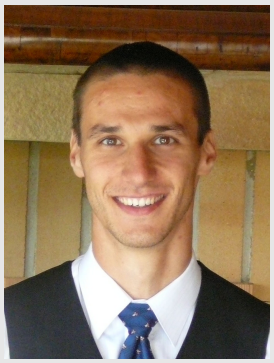
Abstract

Objectives: The DNA analysis is now accepted by the broad public as a completely standard and faultless procedure but in some circumstances its reliability can decrease substantially. This paper deals with the process of identifying and determining the weight of evidence against the suspect. Main stochastic approaches to identification are shown.

Methods: The weight-of-evidence formula was derived from Bayes theorem and its application in the model of the island problem was demonstrated. The beta-binomial formula derived from Dirichlet distribution was used for calculation of more complex situations.

Results: From many various complications in the model of the island problem there was shown how to work with uncertainty in a population size. The beta-binomial formula was used to include a subpopulation structure and in issues of DNA mixtures.

Conclusions: In particular, the influence of a population structure is now explored insufficiently. Using the results of H. Kubátová in this area, a new formula was derived.



Mgr. Dalibor Slovák

Keywords

Identification process, Weight-of-evidence formula, Coancestry coefficient, Beta-binomial sampling formula, DNA mixtures

Correspondence to:

Mgr. Dalibor Slovák
Center of Biomedical Informatics,
Institute of Computer Science AS CR
Address: Pod Vodarenskou vezi 2, Prague 8, Czech Republic
E–mail: slovak@euromise.cz

EJBI 2011; 7(1):44–50
recieved: September 20, 2011
accepted: October 24, 2011
published: November 20, 2011

1 Introduction

DNA profiling, discovered by Alec Jeffreys during the 1980s, has caused a revolution in criminology. DNA helps to convict the perpetrators of those crimes that once appeared irresolvable and also helps to prove the innocence of those who have already been convicted. The DNA analysis is now accepted by the broad public as a completely standard procedure, which reliably convicts the offender. Here, however, hides one of the main problems that results from using DNA, for even DNA evidence is not foolproof.

Several possibilities keep DNA from being completely reliable: for example there may be a false location of the trace (more specifically, the offender may have discarded a cigarette butt which had previously been smoked by someone else); the wrong take of a biological samples or damage to the samples could have occurred; or there may have been secondary transfer of a biological material. However, mathematicians do not deal with any of these things. Rather, they are faced with the following task: if all of the above options are excluded, what is the probability that a particular offender and a detained person are the same, given that the perpetrator's and the suspect's DNA profiles are available? As we will see, the answer depends mainly on the number of loci we use to DNA profiling, and the variability within each of them.

In forensic practice, genetic profiles consisting of the short tandem repeat polymorphisms (STRs) are currently used. STRs are known to vary widely between individuals by virtue of variation in their length and they are found only in the non-coding region of DNA, so they provide no information of medical or personal significance. Therefore,

STRs are very useful and convenient for identification purposes.

The numerical representation of DNA profile consists of two numbers of alleles at each locus examined, one allele inherited from the mother and the other one from the father, along with two letters (XX or XY) which show the result of the gender test. The number of examined loci varies from country to country, with the smallest being seven used in Germany and a maximum of sixteen used in the Czech Republic.

For example, the system of DNA profiling used in the UK is known as SGM Plus. It examines ten loci plus a gender test and produces a numeric DNA profile which may look like this:

```
15,18; 6,9; 11,13; 22,22; 31,32.2; 14,17;
17,20; 11,12; 13,16.3; 15,16; XY.
```

The number provides information about a feature of DNA at each locus we examine. The number of complete repeat units observed is designated by an integer. Variant alleles that contain a partial repeat are designated by a decimal followed by the number of bases in the partial repeat. For example, an 32.2 allele contains 32 complete repeat units and a partial repeat unit of 2 bases ([10]).

Although each person's DNA is unique (apart from identical siblings), there is a very small, but finite chance (less than 1 in a billion in SGM Plus) that two unrelated people could share the same DNA profile. For this reason it is not possible to convict a person on DNA evidence alone and there must be additional corroborating evidence available.

DNA left at a crime scene may also decompose over time because of bacteria, UV light, environmental conditions etc. Due to the quality of biological material and/or its amount it is not always possible to investigate all of the polymorphisms. An incomplete DNA profile may look like

```
15,  ; 6,9; 11,13;   ,  ; 31,32.2; 14,17;
  ,20;   ,12; 13,16.3; 15,16; XY.
```

If an incomplete DNA profile is obtained, the probability of unique identification drops accordingly. However, even very incomplete profiles can still be used to conclusively eliminate a person from an investigation.

In the following text we will assume the examination of one locus only. Assuming independence of loci, generalization to a larger number of loci can be performed using a product rule (i.e. multiplying the individual marginal probabilities).

2 Methods

Denotation

- E - evidence or information about the crime (i.e. the circumstances, witness testimonies, crime scene evidence, etc.),
- G - an event at which the suspect is guilty,
- I - an event at which the suspect is innocent,
- C_i - an event at which the culprit is a person i,
- $\mathcal{I}$ - the population of alternative suspects.

Our goal is to determine the conditional probability of $\mathsf{P}(G|E)$ that, given circumstances E, the suspect is truly the culprit of the investigated crime. According to Bayes theorem

$$\mathsf{P}(G|E) = \frac{\mathsf{P}(E|G)\mathsf{P}(G)}{\mathsf{P}(E|G)\mathsf{P}(G) + \mathsf{P}(E|I)\mathsf{P}(I)}. \tag{1}$$

However, the expression $\mathsf{P}(E|I)$ cannot be counted directly. The suspect is innocent if and only if there exists an index $i \in \mathcal{I}$ in which the event C_i occurs. Then the event I is equivalent to the event $\cup_{i\in\mathcal{I}} C_i$ and thanks to the disjunction of events C_i holds:

$$\mathsf{P}(I) = \mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right) = \sum_{i\in\mathcal{I}} \mathsf{P}(C_i).$$

Thus

$$\begin{aligned}
\mathsf{P}(E|I)\mathsf{P}(I) &= \mathsf{P}\left(E|\cup_{i\in\mathcal{I}} C_i\right)\mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right) = \\
&= \frac{\mathsf{P}\left(E\cap\left(\cup_{i\in\mathcal{I}} C_i\right)\right)}{\mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right)}\,\mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right) = \\
&= \mathsf{P}\left(\cup_{i\in\mathcal{I}}\left(E\cap C_i\right)\right) = \sum_{i\in\mathcal{I}} \mathsf{P}\left(E\cap C_i\right) = \\
&= \sum_{i\in\mathcal{I}} \mathsf{P}(E|C_i)\mathsf{P}(C_i).
\end{aligned}$$

Let define **likelihood ratio**

$$R_i = \frac{\mathsf{P}(E|C_i)}{\mathsf{P}(E|G)} \tag{2}$$

which expresses how many times the probability of evidence E is greater under the condition that the culprit is a person i than under the condition that the culprit is the suspect.
Further we define **likelihood weights**

$$w_i = \frac{\mathsf{P}(C_i)}{\mathsf{P}(G)}$$

which expresses how many times the prior probability of committing a crime by a person i is greater than the prior probability of committing a crime by the suspect.

Then

$$\mathsf{P}(G|E) = \frac{1}{1+\sum_{i\in\mathcal{I}} w_i R_i}. \tag{3}$$

The formula (3) is usually called **the weight-of-evidence formula**.

3 The Island Problem

The simplest application of the previous part is the "island problem". This is a model where a crime is committed on an inaccessible island which contains N people who are unrelated to each other. At the beginning, there is no information about the offender, so we assign to each of the islanders the same (prior) probability of committing a crime. Then the offender is found to possess a certain characteristic Υ (it can be an allele, or a pair of alleles respectively, in the appropriate locus) and the suspect is also found to have that characteristic, Υ. The question becomes, to what extent can we be sure that we have found the suspect who is truly the culprit.

First we calculate the likelihood ratio using the formula (2). Let p be the frequency of the Υ in the population. We suppose that the evidence consists only of the information that the suspect's and the culprit's DNA profiles are the same. If the hypothesis G holds, both profiles come from the same individual and thus the denominator equals 1. The numerator of R_i, $\mathsf{P}(E|C_i)$, can be estimated by p. Because $w_i = 1\ \forall i \in \mathcal{I}$, using the formula (3) we get

$$\mathsf{P}(G|E) = \frac{1}{1 + N \cdot p}\ . \tag{4}$$

For example if $p = 0.01$ and $N = 100$ then $\mathsf{P}(G|E) = 1/2$.

The previous result can be modified for more complex (and realistic) situations. Let's see which situations is this simple model inadequate for:

- *Typing and handling errors*
  As the test may give erroneous results in a small percentage of cases, errors caused by a human factor must also be considered: contamination or replacement of a sample from which the Υ-status is investigated; incorrect evaluation of the results, or even intentional misrepresentation.

- *The population size*
  Often the population size N is only estimated and furthermore, if there is migration in the population, then it is necessary to account for greater uncertainty within the population size.

- *The probability of occurrence Υ in the population*
  The value of p is usually unknown and is therefore estimated on the basis of relative frequency of the Υ in a smaller sample or in a similar population, about which we have more information. However, this auxiliary data may be outdated or may only partially describe the ivestigated population.

- *Suspect searching*
  The suspect is not usually chosen randomly from the population but on the basis of other circumstantial evidence which increase the probability of guilt. Another possibility is to choose the suspect by testing individuals from the population for the presence of Υ. In this way, people who are not Υ-bearers can be excluded and thus the population size of alternative suspects is reduced.

- *Relatives and population subdivision*
  If the suspect (or other individual being tested) is a Υ-bearer and some of his relatives are included in the population too, then in the case of DNA profile increases the probability of other individuals having Υ due to inheritance. Similarly, unusually high relative frequency of a rare character usually occurs within the same subpopulation due to its shared evolution history.

- *The same prior probability of committing a crime*
  Although this requirement intuitively corresponds with the general presumption of innocence, we can asses varying prior probability (i.e. based on the distance from the scene, time availability, or a possible alibi).

We will analyze some of these cases in detail in the following sections.

4 Uncertainty about the Population Size

The uncertainty about the size of the population of possible alternative suspects affects the prior probability of $\mathsf{P}(G)$. Consider the population size $\tilde{N}$ is a random variable with mean N. The prior probability of guilt, given value $\tilde{N}$, is

$$\mathsf{P}(G|\tilde{N}) = 1/(\tilde{N} + 1)$$

but since $\tilde{N}$ is not known, we use the expectation:

$$\mathsf{P}(G) = \mathsf{E}\left[G|\tilde{N}\right] = \mathsf{E}\left[\frac{1}{\tilde{N}+1}\right].$$

The function $1/(\tilde{N}+1)$ is not symmetric but it is at least convex on the interval $(0, \infty)$. Jensen's inequality for convex functions ($\mathsf{E}[f(x)] \geq f(\mathsf{E}[x])$) implies

$$\mathsf{P}(G) = \mathsf{E}\left[\frac{1}{\tilde{N}+1}\right] \geq \frac{1}{N+1}$$

because $\mathsf{E}[\tilde{N}] = N$.

Thus the failure to uncertainty about the value of N tends to favor defendant. Moreover, this effect is usually very small, let it show in a concrete example.

For $\varepsilon \in (0, 0.5)$ we put

$$\tilde{N} = \begin{cases} N-1 & \text{with probability } \varepsilon \\ N & \text{with probability } 1-2\varepsilon \\ N+1 & \text{with probability } \varepsilon. \end{cases}$$

Then

$$\mathsf{P}(G) \quad = \quad \mathsf{E}\left[\frac{1}{\tilde{N}+1}\right] = \frac{\varepsilon}{N} + \frac{1-2\varepsilon}{N+1} + \frac{\varepsilon}{N+2} =$$

$$= \frac{1}{N+1} + \frac{2\varepsilon}{N(N+1)(N+2)} \geq \frac{1}{N+1}$$

and if we put $\varepsilon = 0.25$ and $N = 100$ then $\mathsf{P}(G)$ is greater than $1/(N+1)$ by only 0.000000485.

Let's see what the population size uncertainty causes in formula (4):

$$\begin{aligned}
\mathsf{P}(G|E) &= \frac{1}{1+\sum_i R_i \frac{\mathsf{P}(C_i)}{\mathsf{P}(G)}} = \frac{1}{1+p\frac{1}{\mathsf{P}(G)}\underbrace{\sum_i \mathsf{P}(C_i)}_{=1-\mathsf{P}(G)}} = \\
&= \frac{1}{1+p\frac{N(N+1)(N+2)}{N^2+2N+2\varepsilon}(1-\frac{N^2+2N+2\varepsilon}{N(N+1)(N+2)})} = \\
&= \frac{1}{1+Np\frac{N^3+2N^2-2\varepsilon}{N^3+2N^2+2N\varepsilon}} = \\
&= \frac{1}{1+Np\left(1-2\varepsilon\frac{N+1}{N^3+2N^2+2N\varepsilon}\right)}.
\end{aligned}$$

Substituting again $\varepsilon = 0.25$ and $N = 100$ we receive $\mathsf{P}(G|E) = 0.5000124$ which value, despite the high value of ε, differs from the original result of 50 %, at which we calculate with N fixed, in an order of just one thousandth of a percent. If we want to still count with uncertainty about N,

$$\mathsf{P}(G|E) \approx \frac{1}{1+Np\,(1-2\varepsilon/N^2)}$$

is very good approximation to take. In our example this approximation gives $\mathsf{P}(G|E) = 0.5000125$, i.e. 50.00125 %.

Balding in [1] uses an approximation order of magnitude worse than

$$\mathsf{P}(G|E) \approx \frac{1}{1+Np\,(1-4\varepsilon/N^3)}$$

which gives in our example the value $\mathsf{P}(G|E) = 0.5000003$, i.e. 50.00003 %.

5 DNA Database

DNA profiles as a sequence of alphanumeric data allow relatively easy storage in the database, therefore national databases are created from late 1990's. Currently there are three major forensic DNA databases: CODIS (Combined DNA Indexing System), which is maintained by the United States FBI; the ENFSI (European Network of Forensic Science Institutes) database; and the ISSOL (Interpol Standard Set of Loci) database maintained by Interpol.

All systems mentioned above divide the DNA database into two subdatabases. In *the crime scene database* biological samples collected at the scene are stored, in *the convicted offender database* figure genetic profiles of individuals convicted in the past. These two databases are compared with each other and eventual match of profiles is examined by qualified professionals.

The type of offenses for which DNA is stored differs among countries and states. Initially, these databases contained only samples from violent offenders, those convicted of aggravated assault, rape, or murder. However, the value of obtaining DNA from offenders of less severe crimes has been recognized, as many small time criminals become repeat offenders and also more violent offenders. The power of a large bank of DNA samples extends to the possibility of it acting as a deterrent. A match of DNA evidence from a crime scene (which would then be logged in the crime scene database) to one in the convicted offender database rapidly solves the crime, saving time, effort, and money ([3]).

The absolutely largest national database is the US National DNA Index System (NDIS). It contains almost ten million offender profiles and over 380 000 forensic profiles as of July 2011 ([7]). The oldest and relatively largest database is the national DNA database of UK (NDNAD) which currently consists of over six and a half million profiles.

After the creation of DNA databases the number of solved crimes in the UK has increased from 24 % to 43 %. The success of this approach is also confirmed by the fact that a new crime scene DNA profile being loaded to the DNA database had a 45 % chance of matching a persons DNA profile in 2002/03 against 60 % in 2008/09 ([8]). Thus the database system has the support of public. On the other hand, from DNA very sensitive personal information can be obtained and therefore it is necessary to ensure a thorough protection of databases against abusing.

The Czech national database was created in 2002. Then there was a rapid development of the database and it currently contains approximately 90 000 genetic profiles.

6 Relatives and Population Structure

Alleles, which are identical and come from a common ancestor, are called *ibd* (identical by descent). A common recent evolution history of two individuals, whether relatives or members of the same subpopulation, increases the probability of occurrence of ibd alleles. Therefore, as the degree of relatedness within subpopulations is used, the *coancestry coefficient* θ indicated the probability that two randomly selected alleles on fixed locus are ibd. Neglecting the influence of kinship and population structure leads to overestimation of posterior probability of the suspect's guilt. Ignoring this tends to suspect's disfavour, so this topic is given considerable attention.

Balding and Nichols in [2] proposed a method which allows to calculate probability of observing considered genotype in structure population via coancestry coefficient. More detailed mathematical derivation of method including several corrections was provided by Kubátová in [6].

Let's denote p_A and p_B frequencies of alleles A a B in the whole population, k proportion of the subpopulation in the general population and θ coancestry coefficient in the subpopulation. The probability of observing homozygous genotype can be calculated as

$$\mathsf{P}(AA) = p_A\left(\theta + (1-\theta)\,\frac{p_A - \theta k}{1-\theta k}\right) \tag{5}$$

and similarly heterozygous genotype:

$$\mathsf{P}(AB) = 2p_A p_B \frac{1-\theta}{1-\theta k}\,. \tag{6}$$

Balding and Nichols do not use variable k in their derivation, we get their results by putting $k = 1$. Thus, probabilities of homozygous genotypes decreased and conversely, probabilities of heterozygous genotypes increased.

7 Beta-binomial Formula

To get formulas (5) and (6), we can use also a more general approach proposed by Wright ([11]). Consider on given locus J alleles $A_1, \ldots, A_J$ having probability of occurrence in the population $p_1, \ldots, p_J$, $\sum_{i=1}^{J} p_i = 1$. Allele proportions in the subpopulation can be modelled by the Dirichlet distribution with parametres $\lambda p_i, \lambda = \frac{1-\theta}{\theta(1-k)}$. Thus the probability of observing m_i alleles A_i ($\sum_i m_i = n$) is given by

$$\mathsf{P}(m_1, \ldots, m_J) = \frac{\Gamma(\lambda)}{\Gamma(\lambda + n)} \prod_{i=1}^{J} \frac{\Gamma(\lambda p_i + m_i)}{\Gamma(\lambda p_i)}. \tag{7}$$

Putting $m = (m_1, \ldots, m_J)$ we can adjust formula (7) to

$$\mathsf{P}(m) = \frac{\prod_{j=1}^{J} \prod_{i=0}^{m_j - 1} [(1-\theta)\,p_j + \theta i\,(1-k)]}{\prod_{i=0}^{n-1} [1 - \theta + \theta i\,(1-k)]}. \tag{8}$$

The formula (8) is usually called **beta-binomial sampling formula** and applies to ordered samples. If we want to work with unordered samples, it is necessary to multiply the result by $\frac{n!}{m_1! \cdots m_J!}$.

From the formula (8) we can also deduce the probability of observing certain combination of alleles. For $J = 2$, $m_A = 2$ and $m_B = 0$ we have

$$\begin{aligned}
\mathsf{P}(AA) &= \frac{(1-\theta)\,p_A\,[(1-\theta)\,p_A + \theta\,(1-k)]}{(1-\theta)\,[1 - \theta + \theta\,(1-k)]} = \\
&= p_A\left[\frac{(1-\theta)\,p_A + \theta - \theta k}{1-\theta k} + \theta - \frac{\theta - \theta^2 k}{1-\theta k}\right] = \\
&= p_A\left[\theta + \frac{(1-\theta)\,p_A + \theta - \theta k - \theta + \theta^2 k}{1-\theta k}\right] = \\
&= p_A\left[\theta + \frac{(1-\theta)\,p_A - \theta k\,(1-\theta)}{1-\theta k}\right] = \\
&= p_A\left[\theta + (1-\theta)\,\frac{p_A - \theta k}{1-\theta k}\right],
\end{aligned}$$

which is in agreement with (5).

Similarly putting $J = 2$, $m_A = 1$ and $m_B = 1$ in the formula (8) we get

$$\mathsf{P}(AB) = 2\frac{(1-\theta)\,p_A\,(1-\theta)\,p_B}{(1-\theta)\,(1-\theta+\theta\,(1-k))} = 2p_A p_B \frac{1-\theta}{1-\theta k},$$

which agrees with formula (6).

8 Aplication of Beta-binomial Formula

Using the formula (8) we can also deduce the probability of observing certain allele given by our knowledge of previous alleles observing:

$$\mathsf{P}(m_j{+}1|m_1, \ldots, m_j, \ldots, m_J) = \frac{(1-\theta)\,p_j + m_j\theta\,(1-k)}{1 - \theta + n\theta\,(1-k)}. \tag{9}$$

Denote G_C and G_S genotype of culprit and suspect respectively, and generally G_i genotype of person i. Likelihood ratio (2) can be rewritten as

$$\begin{aligned}
R_i &= \frac{\mathsf{P}\,(G_C = G_S = D|C_i)}{\mathsf{P}\,(G_C = G_S = D|G)} = \frac{\mathsf{P}\,(G_i = G_S = D)}{\mathsf{P}\,(G_S = D)} = \\
&= \mathsf{P}\,(G_i = D|G_S = D)\,.
\end{aligned}$$

Suppose first that the suspect has a homozygous profile A_jA_j and with this knowledge calculate the probability that the suspect has the same homozygous profile:

$$\begin{aligned}
R_i &= \mathsf{P}(G_i = A_jA_j|G_S = A_jA_j) \equiv \mathsf{P}(A_j^2|A_j^2) = \\
&= \mathsf{P}(A_j|A_j^3) \cdot \mathsf{P}(A_j|A_j^2)
\end{aligned}$$

We know how to calculate these conditional probabilities by using (9). First we put $m_j = n = 2$ and then $m_j = n = 3$, that is

$$R_i = \frac{[(1-\theta)\,p_j + 2\theta\,(1-k)]\,[(1-\theta)\,p_j + 3\theta\,(1-k)]}{[1 - \theta + 2\theta\,(1-k)]\,[1 - \theta + 3\theta\,(1-k)]}.$$

Similarly we proceed for a heterozygous profile A_jA_k:

$$\begin{aligned}
R_i &= \mathsf{P}(G_i = A_jA_k|G_S = A_jA_k) \equiv \mathsf{P}(A_jA_k|A_jA_k) = \\
&= \mathsf{P}(A_k|A_j^2A_k^1)\mathsf{P}(A_j|A_j^1A_k^1) + \\
&\quad + \mathsf{P}(A_j|A_j^1A_k^2)\mathsf{P}(A_k|A_j^1A_k^1).
\end{aligned}$$

To quantify both expressions on the bottom line we put $m_j = 1, n = 2$ and $m_k = 1, n = 3$; $m_k = 1, n = 2$ and $m_j = 1, n = 3$ respectively. In total

$$R_i = 2\frac{[(1-\theta)\,p_j + \theta\,(1-k)]\,[(1-\theta)\,p_k + \theta\,(1-k)]}{[1 - \theta + 2\theta\,(1-k)]\,[1 - \theta + 3\theta\,(1-k)]}.$$

9 DNA Mixtures

If the DNA sample is found to have more than two alleles at one locus, it is clear to be a mixture. The number of contributors to the mixture can be known or estimated, usually as $\lceil \frac{n}{2} \rceil$ where n is the maximum number of alleles detected. Because of the large number of situations that may have arised we show for illustration only the case when the victim (V) and one other individual contribute to the mixture.

The likelihood ratio R_i defined by formula (2) can be rewritten as

$$\begin{aligned} R_i &= \frac{\mathsf{P}\left(E_C, G_S, G_V | C_i\right)}{\mathsf{P}\left(E_C, G_S, G_V | G\right)} = \\ &= \frac{\mathsf{P}\left(E_C | G_S, G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} \cdot \frac{\mathsf{P}\left(G_S, G_V | C_i\right)}{\mathsf{P}\left(G_S, G_V | G\right)} = \\ &= \frac{\mathsf{P}\left(E_C | G_S, G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} = \frac{\mathsf{P}\left(E_C | G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} \end{aligned} \quad (10)$$

Four alleles mixture

First we look at the case where the mixture consists of four alleles.

Suppose the following conditions apply:

1. None of the individuals are considered relatives to each other.
2. The population is homogeneous (i.e. $\theta = 0$).
3. The population follows Hardy-Weinberg equilibrium.

Let the mixture be made up of alleles A, B, C, D with known probabilities of occurrence in the total population p_A, p_B, p_C, p_D and let the suspect have alleles A, B and the victim C, D respectively. The denominator in the formula (10) is equal to one, the numerator is equal to the probability of observing the individual with alleles A, B which is under the above assumptions $2p_Ap_B$. Therefore, the likelihood ratio equals to

$$R_i = 2p_Ap_B.$$

Suppose now that all considered individuals have the same degree of relatedness to each other expressed by coancestry coefficient θ. Then according to (9)

$$\begin{aligned} R_i &= \mathsf{P}\left(AB|ABCD\right) = \\ &= \frac{2\left[(1-\theta)\,p_A + \theta\,(1-k)\right]\left[(1-\theta)\,p_B + \theta\,(1-k)\right]}{\left[1-\theta+4\theta\,(1-k)\right]\left[1-\theta+5\theta\,(1-k)\right]}. \end{aligned}$$

Three alleles mixture

In the case of three alleles in the sample, assuming at least two contributors to the mixture is also necessary. Consider alleles A, B, C with probabilities p_A, p_B, p_C. If the victim is homozygous for allele C, we get the same results as in the case of a mixture of four alleles.

Let's assume that the victim is heterozygous with alleles A, B. Let the suspect be homozygous for allele C and conditions 1 to 3 are fulfilled. The denominator of the formula (10) is again equal to one, the numerator equals to the probability of observing an individual who has the allele C and does not have a different allele than A, B or C. Therefore

$$\begin{aligned} R_i &= \mathsf{P}(AC) + \mathsf{P}(BC) + \mathsf{P}(CC) = \\ &= 2p_Ap_C + 2p_Bp_C + p_C^2\,. \end{aligned} \quad (11)$$

To include the population structure we use the formula (9) again:

$$\begin{aligned} R_i &= \mathsf{P}\left(AC|ABCC\right) + \mathsf{P}\left(BC|ABCC\right) + \\ &+ \mathsf{P}\left(CC|ABCC\right) = \\ &= \frac{2\left[(1-\theta)\,p_A + \theta\,(1-k)\right]\left[(1-\theta)\,p_C + 2\theta\,(1-k)\right]}{\left[1-\theta+4\theta\,(1-k)\right]\left[1-\theta+5\theta\,(1-k)\right]} \\ &+ \frac{2\left[(1-\theta)\,p_B + \theta\,(1-k)\right]\left[(1-\theta)\,p_C + 2\theta\,(1-k)\right]}{\left[1-\theta+4\theta\,(1-k)\right]\left[1-\theta+5\theta\,(1-k)\right]} \\ &+ \frac{\left[(1-\theta)\,p_C + 3\theta\,(1-k)\right]\left[(1-\theta)\,p_C + 2\theta\,(1-k)\right]}{\left[1-\theta+4\theta\,(1-k)\right]\left[1-\theta+5\theta\,(1-k)\right]} \\ &= \frac{\left[(1-\theta)\,p_C + 2\theta\,(1-k)\right]}{\left[1-\theta+4\theta\,(1-k)\right]} \times \\ &\times \frac{\left[(1-\theta)\,(2p_A + 2p_B + p_C) + 7\theta\,(1-k)\right]}{\left[1-\theta+5\theta\,(1-k)\right]}. \end{aligned}$$

We assumed in the previous calculation that the suspect is homozygous for allele C. If he is heterozygote with alleles A and C, or B and C respectively, formula (11) remains unchanged under conditions 1 to 3. If the population structure is included, we get in both cases the likelihood ratio

$$\begin{aligned} R_i &= \frac{\left[(1-\theta)\,p_C + \theta\,(1-k)\right]}{\left[1-\theta+4\theta\,(1-k)\right]} \times \\ &\times \frac{\left[(1-\theta)\,(2p_A + 2p_B + p_C) + 8\theta\,(1-k)\right]}{\left[1-\theta+5\theta\,(1-k)\right]}. \end{aligned}$$

10 Conclusion

We derived the weight-of-evidence formula and its simplest applications. To include the uncertainty about the population size we proposed a better approximation than Balding in ([1]). We showed how to include the subpopulation structure into the model. Here we used new results from ([6]) which we plan to investigate in more detail in future.

Acknowledgements

This work was partially supported by the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic and by the project SVV-2011-262514 of Charles University in Prague.

References

[1] Balding D.J.: *Weight-of-evidence for forensic DNA profiles*, John Wiley & Sons, Ltd, 2005, pp. 15-63

[2] Balding D.J., Nichols R.A.: *DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands*, Forensic Science International **64**, 1994, pp. 125-140

[3] eNotes. World of Forensic Science. *DNA Evidence, Social Issues* [online]. 2011 [cit. 2011-9-15]. Available at www.enotes.com/forensic-science/dna-evidence-social-issues.

[4] Slovák Dalibor: *Stochastic Approaches to Identification Process in Forensic Medicine and Criminalistics*, in Doktorandské dny '11, Matfyzpress, Praha, 2011

[5] The office for personal data protection. *Otevřete ústa, prosím...* & *Databáze DNA* [online, in czech]. February 2007 [cit. 2011-9-15]. Available at www.uoou.cz/uoou.aspx?menu=287 &submenu=288.

[6] Kubátová H., Zvárová J. (supervisor): *Statistical methods for interpreting forensic DNA mixtures*, MFF UK, Praha 2010, pp. 20-26

[7] The Federal Bureau of Investigation. *CODISNDIS Statistics* [online]. July 2011 [cit. 2011-9-15]. Available at www.fbi.gov /about-us/lab/codis/ndis-statistics.

[8] The National Policing Improvement Agency. *The National DNA atabase* [online]. 2010 [cit. 2011-9-15]. Available at www.npia.police.uk/en/8934.htm.

[9] Slovák D., Zvárová J. (supervisor): *Statistické metody stanovení váhy evidence v procesu identifikace jedince*, MFF UK, Praha, 2009

[10] The Applied Biosystems. *AmpFℓSTR SGM Plus. PCR Amplification Kit. User's Manual* [online]. 2011 [cit. 2011-9-15]. Available at www3.appliedbiosystems.com/cms/groups /applied_ markets_support/documents/generaldocuments /cms_041049.pdf, pp. 178.

[11] Wright S.: *The genetical structure of populations*, Ann. Eugen. **15**, 1951, pp. 323-354

Jitter Effect on the Performance of the Sound Localization Model of Medial Superior Olive Neural Circuit

Pavel Šanda[1,2]

[1]Institute of Physiology, Academy of Sciences of the Czech Republic

[2]3rd Medical Department, First Faculty of Medicine, Charles University in Prague, Czech Republic

Abstract

Objectives: Additional properties of the stochastic neural circuit model suggested in [1] were studied.
Methods: The performance of the whole circuit when the system employs a different jitter was studied by extensive simulations. By performance we mean the time needed to obtain a reliable estimate of ITD.
Results: It was found that the relation between jitter and performance is nonlinear and we estimated a plausible range of jitter values for the model.
Conclusion: To conclude, there exists an upper bound of the timing jitter since the number of neurons needed to compensate the injected noise grows exponentially and above certain jitter values becomes unrealistically high.

Mgr. Pavel Šanda

Keywords

Medial superior olive (MSO), stochastic model, timing jitter, interaural time difference (ITD)

Correspondence to:

Mgr. Pavel Šanda
Institute of Physiology, Academy of Sciences of the Czech Republic
Address: Videnska 1082, 142 20 Prague 4, Czech Republic
E–mail: sanda@biomed.cas.cz

EJBI 2011; 7(1):51–54
recieved: September 10, 2011
accepted: October 31, 2011
published: November 20, 2011

1 Introduction

The way mammalian brain localizes sound azimuth remains a matter of discussion. The current textbook view is based on the theory of delay lines proposed a long time ago by [2].

Although there is a strong experimental evidence that delay lines implemented by the branching pattern of neuronal fibers are present in the Nucleus Laminaris in birds [3], experimental evidence for such branching pattern in the Medial Superior Olive (MSO - counterpart of bird's NL) in mammals remain weak [4] and alternative theories have been proposed [5].

In a specific variant of the slope-encoding model [6] proposed in [1] the interaural time difference (ITD) is encoded by the firing rate of the first binaural neuron. This rate is driven by coincidence detection of the action potentials coming from time locked ipsi- and contralateral inputs shifted by ITD and additional jitter added to the system. Under certain conditions each ITD value corresponds to a unique value of the firing rate, thus the imaginary observer monitoring output of such a neuron is able to estimate ITD only by interpolation from its firing rate.

The role of noise in this model is ambiguous. On the one hand it allows a finer distribution of recognized ITD values, on the other hand higher values deteriorate the estimation performance of the circuit.

This performance decline was indicated in [1] for two circuits with different jitter. The aim of this report is to extend the previous result and show quantitatively how jitter affects performance of the whole range of circuits defined by different jitter values.

2 Methods

The circuit operates at an abstract level of description without explicit membrane potential regarding spikes as single time point events and consists of several consecutive processing stages (see Fig. 1):

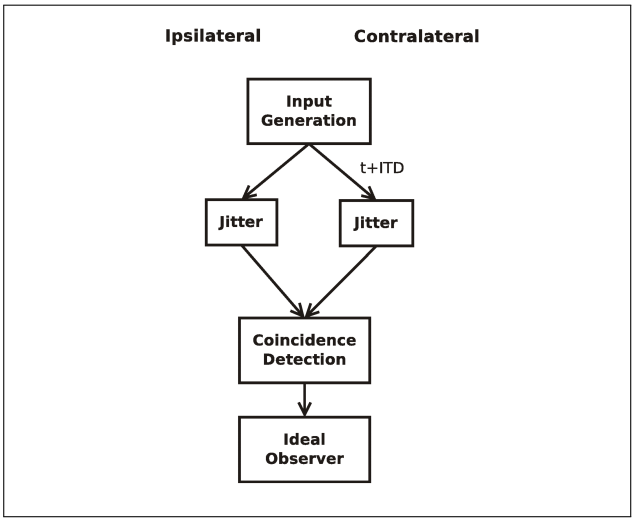


Figure 1: Scheme of successive processing stages of the circuit.

- A generator of action potentials simulates time-locked inputs, impulses from the contralateral side are shifted in time by ITD value. The frequency of generation is set to 140 Hz as in the previous study.
- A jitter generator which represents noise occurring in the circuit during the signal transmission along the auditory pathway. It is parametrized by a single value. It should be noted that each different parameter value defines a different circuit since it changes the characteristic ITD *interpolation curve* used for interpolation. Together with the spike generator they can be considered as a very simplified counterpart to the auditory pathway up to the MSO (where the signal from the left and right ear used for sound localization based on low sound frequencies converges). In this stage each spike is shifted in time by small random jitter Δ which is parametrized by jitter magnitude J, more precisely $\Delta = J(B(2,4) - 0.5)$, where $B(a,b)$ is a random variable from the beta distribution with parameters a, b.
- A coincidence detector representing the first binaural neuron. It generates a new spike only in case two input spikes occur within a short time window and in a specific order when contralateral spike precedes the ipsilateral one.
- An observer which collects output of the previous processing stages and estimates the ITD value computed by the circuit. It can be seen as a counterpart of higher processing stages which measure how much information can, in principle, be obtained from the rate coded presented by a single binaural neuron.

Details of the stages above are identical to those in [1] except for one important feature. Fixed parameters of the circuit define the ITD interpolation curve as seen in Fig. 2. In our previous study this curve was carefully fitted to a fixed sinusoidal function and the inverse of this function was used to interpolate ITD from estimated firing rate.

In Fig. 2 we can see how jitter J dramatically changes this curve. Since we will use the whole range of different jitter values we cannot rely on the fitted function anymore and we shall use directly this interpolation curve. Conceptually, this is not adding anything new, however, it leads to additional computational difficulties - for each jitter value a circuit ITD curve must be recomputed anew and an inverse mapping from firing rate to ITD must use a more elaborate interpolation mechanism since the curve is not locally strictly monotonous.

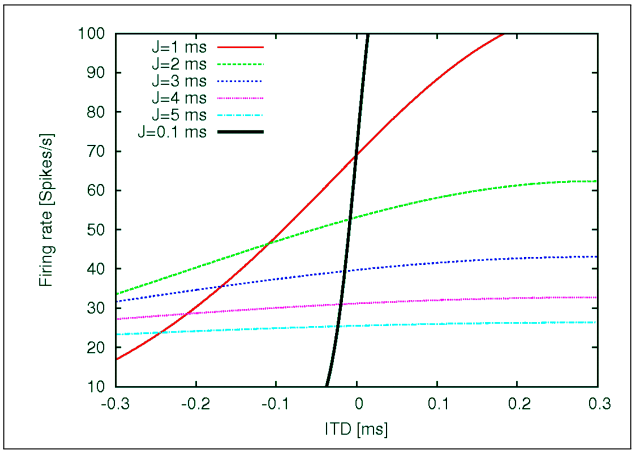


Figure 2: ITD interpolation curves for circuits with a different jitter value. Each firing rate value corresponds to an ITD value and is uniquely determined in case the function is strictly increasing in the ITD values under scrutiny. We see that increasing jitter leads to smaller slopes of the interpolation curve and we expect a deteriorated circuit performance for higher jitter values.

3 Results

Each jitter value defines a new circuit and after computing its interpolation curve we let the circuit estimate a single ITD value while observing how the estimate develops in time. This way we obtain asymptotic behaviour for each circuit, see Fig. 3.

From psychophysical experiments we know that the precision of azimuth estimation in a human is approximately 4° in the head-on direction [7]. We define that the time needed for reliable estimation of ITD is identical with the last-passage-time (LPT) of the 4° precision region, see the area delineated by horizontal dotted lines in Fig. 3.

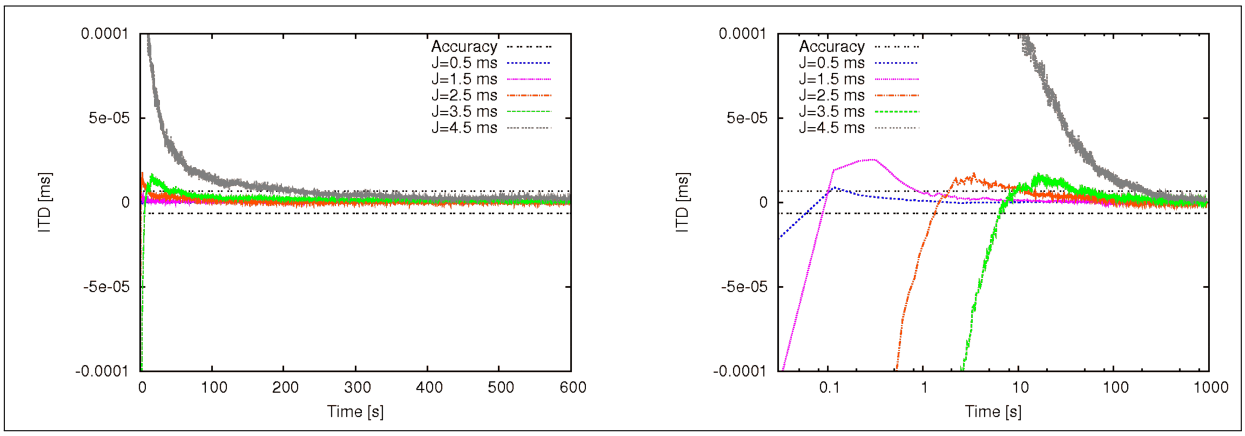


Figure 3: Asymptotic behaviour of ITD estimation produced by observer for selected values of jitter J. The original azimuth was selected as $ITD = 0$. Horizontal lines delineate the region when desired precision of ITD estimate was achieved ($\pm 2^\circ$). For each line we can define the last passage time (LPT) when the function enters the region and remains inside of it. We see that increasing jitter leads to the increase of LPT value. Each line is an averaged function from 1000 simulation runs.

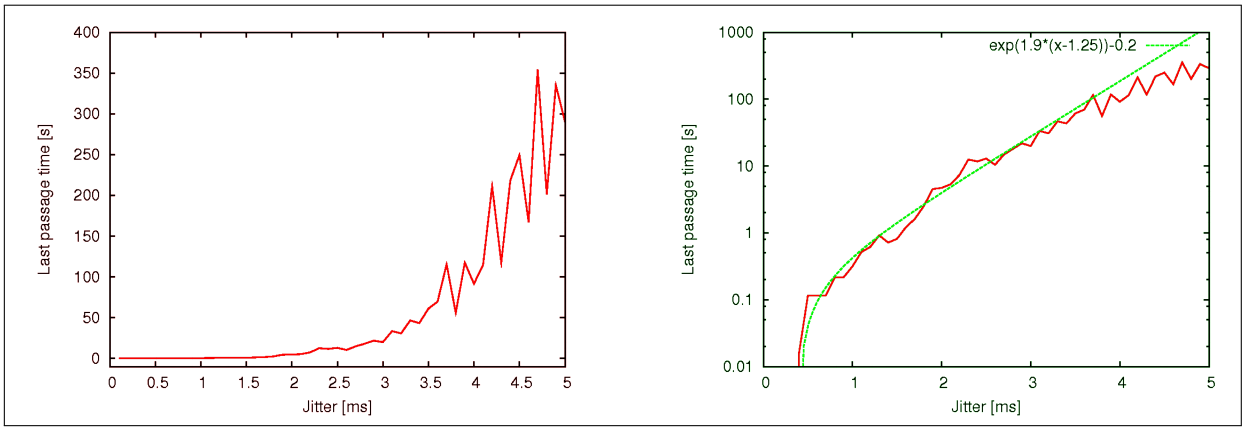


Figure 4: Dependence of last-passage-times on different jitter values. On the right hand side the same plot in logarithmic scale. We fit $f_{fit}(x)$ in such a way to be as close as possible in the interval of 0.1 - 10 s of LPT. This will be subsequently used for relating plausible jitter ranges, see text.

In this way we obtain unique LPT for each circuit with specific jitter value, as plotted in Fig. 4. As we can see, the functional dependence is nonlinear and can be approximately fitted by $f_{fit}(x) = e^{1.9(x-1.25)} - 0.2$.

Obtained circuit's LPT time t_A corresponds to the processing time of a single binaural neuron in MSO needed to estimate ITD. Because the auditory pathway consists of many parallel fibers and processing of the signal is simultaneous, we used the ergodic hypothesis in our previous study.

In short, we assume that when a single neuron of this type requires the time t_A, n neurons working in parallel need the time t_A/n to produce equivalent information subsequently used in higher stages of the pathway (represented by the concept of the observer).

The number of binaural neurons working in parallel is difficult to estimate but does not exceed hundreds of units. Next, we know from psychophysical experiments that the time t_A needed for azimuth estimation ranges around 150 - 300 ms in human subjects [8].

This allows us at least to connect specific jitter value J with the required number of neurons n in order to obtain t_A (let us fix $t_A = 0.2$ s). By employing the ergodic hypothesis we get $t_A = \frac{LPT(J)}{n}$ and from fitting $f_{fit}(J)LPT(J)$, hence

$$n = \frac{f_{fit}(J)}{t_A} = \frac{e^{1.9(J-1.25)} - 0.2}{0.2} \,. \qquad (1)$$

To sum up, we obtain that the physiologically plausible range of simultaneously working neurons $n \in [1; 100]$ corresponds to jitter range $J \in [0.7; 2.8]$, which also implies plausible jitter values for the canonical set of parameters of this model.

4 Discussion

Irregularities in spike timings observed in physiological recordings were originally thought to be the result of neuronal cells unreliability and it was assumed that the

firing-rate neural coding scheme is used because of its robustness against the noise present in neuronal activity. Later decades have shown that what was often considered as erratic behaviour was rather a misunderstanding of the transmitted code [9] and it turned out that neurons are capable of reliable and precise spike timing [10] needed for so-called temporal coding. Coincidence detection of precisely timed input spikes is an important concept in theories of binaural hearing and we suggested one variant of such a model in a stochastic neural circuit in [1].

This time we focused specifically on the role of jitter. In the previous study the jitter parameter was fixed to $J = 1$ ms which is in a good agreement with experimental findings [11]. Here we took a further step and estimated a range of possible values based on circuit performance. We should, however, note that this analysis is bound to the canonical set of basic circuit parameters. For example, the spike generator frequency also has an impact on the overall performance of the circuit; in the previous study we employed a more detailed model of the auditory periphery [12] and we could observe a decrease of overall performance of the circuit. This result cannot be, however, so easily incorporated since one processing stage (bushy cells layer) is missing. There are indications that this layer is able to provide better time locking and consequently improve coincidence detection in binaural neurons — that can be another example of a somewhat unexpected observation that higher processing stages of neural circuity increase the accuracy of phase locking [13].

Another problematic point is that the number of parallel circuits employed in ITD estimation is not experimentally known. This parallelism would have a strong impact on the overall performance as well, and we have at least shown the correspondence between jitter and the required number of neurons (or vice versa). By employing the ergodic hypothesis we can conclude that due to (1) the number of neurons needed to compensate the injected noise grows exponentially and above certain jitter values becomes unrealistically high. This gives us an approximate upper bound of jitter allowed for this type of circuit.

Acknowledgments

The work was supported by the grant SVV-2011-262 514 of Charles University in Prague.

References

[1] Sanda P, Marsalek P. Stochastic InterpolationModel of the Medial Superior Olive Neural Circuit. Brain Res. 2011;in press.

[2] Jeffress LA. A place theory of sound localization. J Comp Physiol Psychol. 1948;41(1):3539.

[3] Carr CE, Konishi M. Axonal delay lines for time measurement in the owls brainstem. Proc Natl Acad Sci USA. 1988;85(21):83118315.

[4] Grothe B. New roles for synaptic inhibition in sound localization. Nat Rev Neurosci. 2003;4(7):540-50.

[5] Jennings TR, Colburn HS. Models of the Superior Olivary Complex. In: Meddis R, Lopez-Poveda EA, Fay RR, Popper AN, editors. Computational Models of the Auditory System. Springer, New York; 2010. p. 6596.

[6] McAlpine D, Jiang D, Palmer AR. A neural code for low-frequency sound localization in mammals. Nat Neurosci. 2001;4(4):396401.

[7] Mills AW. Auditory Localization. In: Tobias JV, editor. Foundations of Modern Auditory Theory. New York: Academic Press; 1972. p. 303348.

[8] Middlebrooks JC, Green DM. Sound Localization by Human Listeners. Annu Rev of Psychol. 1991;42(1):135159.

[9] Barlow HB. Single units and sensation: a neuron doctrine for perceptual psychology. Perception. 1972;1(4):371394.

[10] Mainen ZF, Sejnowski TJ. Reliability of spike timing in neocortical neurons. Science. 1995;268(5216):15031506.

[11] Oertel D, Bal R, Gardner SM, Smith PH, Joris PX. Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. Proc Natl Acad Sci USA. 2000;97(22):1177311779.

[12] Meddis R. Auditory-nerve first-spike latency and auditory absolute threshold: A computer model. J Acoust Soc Am. 2006;119(1):406417.

[13] Carr CE, Heiligenberg W, Rose GJ. A time-comparison circuit in the electric fish midbrain. I. Behavior and physiology. J Neurosci. 1986;6(1):107.

Physiological Model Creation and Sharing

Jan Šilar[1,2], **Jiří Kofránek**[1]

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

[2] 3rd Medical Department, First Faculty of Medicine, Charles University in Prague, Czech Republic

Abstract

Background: Mathematical modelling in medicine (physiology) takes place in research, clinical practice and learning. There are several suitable modelling languages and tools for physiological model implementation. Graphical simulators are created in order to make models used in teaching more illustrative.
Objectives: It is important to make model developers' cooperation easier. For this reason it should be possible to interconnect models written in different languages. E-learning simulators must be well available for students and teachers. Necessity to install the application can be an obstruction. Methods for sharing models written in Modelica language should be enhanced.
Methods: Standardised interface FMI allows interconnection of models written in different languages. E-learning models are developed on .NET platform. It enables them to run in web browser. The editor for models written in Modelica is being developed. It will run in browser and will be interconnected with web based model repository.
Results: Simulation runtime, solver and simulation centre was implemented in .NET languages. They are used in web e-learning applications. They will be also used together with web-based model repository, that is being developed now.

Ing. Jan Šilar

Conclusions: New tools will simplify model creation and interconnection. They will also improve collaboration of model developers.

Keywords

Model, simulation, physiology, differential equations, solver, domain specific language, block oriented / acausal language, simulator, .NET, teaching, web-based model repository

Correspondence to:

Ing. Jan Šilar
Institute of Pathological Physiology, First Faculty of Medicine,
Charles University in Prague, Czech Republic
Address: U Nemocnice 5, 128 53 Prague 2, Czech Republic
E–mail: jansilar@jansilar.cz

EJBI 2011; 7(1):55–58
recieved: September 15, 2011
accepted: November 8, 2011
published: November 20, 2011

1 Introduction

Many scientists and physicians in the whole world deal with physiological modelling. Modelling is used particularly in areas like genetics, proteomics, pharmacokinetics, physiology of organ systems etc.

It is important for scientists to be able to cooperate well and share results of their work. Sharing models by lengthy descriptions in journal articles is not always sufficient. It is impossible to describe the complex model with all details in the article. Sharing implemented models ready to be simulated on a computer is better. It is important to choose a language that can describe the model's mathematical relations in an accurate as well as understandable manner. The goal is to make even a complex and complicated model easy to understand.

2 Languages and Tools

From the mathematical point of view, models can consist of algebraic equations, ordinary differential equations (ODE) and partial differential equations (PDE). There are also stochastic models, agent models, neuron networks, etc. All listed options can also be combined. In this arti-

cle we concentrate on models based on systems of ordinary differential equations and algebraic equations (DAE). Several convenient languages exist to implement these models.

Implementing simulation runtime in some general programming languages (usually FORTRAN or C), that combines the model and a numerical solver is an old approach but it is sometimes still utilised. This makes the creation of a solver tailored for the particular model possible. Various numerical methods, different time steps etc. may be used for each equation of the model to make simulation faster and more precise. This approach was convenient in times when computer performance was significantly lower than today. This was actually the only possibility when no specialised simulating tools and languages were yet developed. The disadvantage is that the model and the solver are merged in one code. It is necessary to implement numerical methods to solve equations with each new model. This is very labour-intensive and demands knowledge of numerical mathematics. The main disadvantage is a lack of clarity. The model equations are being lost in the clutter of the code.

It is more suitable to use one of the domain specific languages (i.e. a language focused on the specific issue) for the DAE model description. These languages use either a block oriented or an acausal approach or their combination [2, 5].

Block oriented languages (also called signal or causal languages) describe a model using functional blocks. These blocks have inputs and outputs that are used for their interconnection. Individual blocks may for example represent common mathematical operations such as addition, multiplication, integration or many other mathematical functions. The signal passes through these blocks, where it is being modified. Like this the equations are already clearer. The functional blocks stand for equations and the connections between them stand for the variables.

If we find a feedback (i.e. the input is being fed by an output dependant on this input) in the block oriented model, we call this situation an *algebraic loop*, see Figure 1.

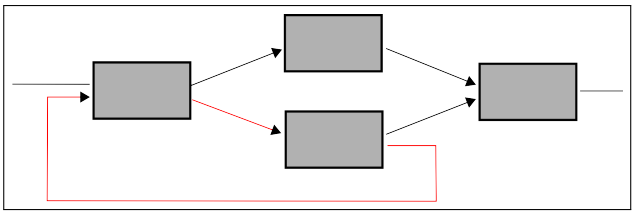


Figure 1: Algebraic loop.

Some algebraic loops can be solved by simulation tools, some represent very complex computational issues and some are insoluble altogether.

The most used and also the most refined block oriented[1] simulation language, also used in physiology [3], is *Simulink* by Mathworks company[2]. The obvious disadvantage is its dependency on the commercial Simulink/Matlab tool. Another example of a commercial signal modelling tool also utilised in the physiology field is *LabView* [7, 4].

Acausal languages work with equations. Unlike in the block oriented approach, there is no need to identify which variable is calculated from which equation. The equations constitute a system of equations, which is being solved all at once. This best represents the mathematical approach. The equations are clearly visible in the implemented model. The following languages are closer to the acausal approach.

There are several open languages directly focusing on biological simulations. The first is the *SBML*[3] [1] (Systems Biology Markup Language), a markup language based on XML. MathML is used for the description of equations. The models being created can be put together from components (submodels), which increases clarity. The next advantage is the possible usage of one component in various models. There are many tools available for model creation and simulation in this language. The SBML language is particularly designed for chemical reaction modelling (genetics, proteomics, enzyme chemical reactions, metabolic pathways). Another very similar language is *CellML*[4] and is designed for cell simulation. It is being developed at the Auckland university. The CellML project contains a storage space, where models by different authors are shared[5] [10].

JSim [9] is a modelling tool, which came into being as a part of the Physiome Project[6]. It uses its own text language *MML* (Mathematical Modeling Language) for the model description. The tool can also cooperate well with SBML and CellML languages. JSim can work with models based on ODE, PDE and discrete[7] equations. JSim is running partially on the client computer and partially on the server. The Physiome project also includes a web-based model repository.

Modelica is an acausal object-oriented language aimed at complex physical system modelling. It was developed by an international organisation Modelica Association. The language can model hybrid systems described with algebraic, differential and discrete equations. Libraries with predefined blocks that model several physical and technical elements (e.g. electrical, mechanical and heat components, mathematical functions etc.) are a part of the language. Well known commercial modelling and simulation environments are Dymola and MathModelica. MathModelica enables direct interconnection with the integrated computational system Mathematica.

[1] Acausal libraries have been arising in Simulink recently, e.g. Simscape.
[2] http://www.mathworks.com/products/simulink/
[3] http://www.sbml.org
[4] http://www.cellml.org
[5] http://models.cellml.org/cellml
[6] http://www.physiome.org
[7] Discrete equations are often used to describe phenomena where values of certain quantities or even whole equations are being altered depending on some conditions (usually inequalities).

OpenModelica is the most widely used open environment. It is developed by Open Source Modelica Consortium in cooperation with Linköping University and other 10 universities and 14 companies. Part of the OpenModelica software package are several tools: OMEdit – graphical tool for creation of models, OMC – compiler of Modelica, OMShell – tool for simulation and other actions with models from command line, OMNotebook – for teaching Modelica (creates similar notebooks as Mathematica.), OMOptim – for model optimisation and others.

It is possible either to write the model textually from scratch or use and connect blocks from Modelica libraries or to create your own blocks. The complete model is translated in C. It is further compiled with runtime and numerical solver into an executable application. Everything is of course done automatically.

OMEdit also contains a simulation centre, that can run models, plot results into graphs etc. OMEdit can even simulate interactively, i.e. it allows a user to change model parameter values during simulation and directly observe effects of these changes.

3 FMI

It is obvious that many different simulation languages and tools are used. It is often purposeful, because to describe certain problems, some languages are more suitable than others. Standardised interface *FMI*[8] (Functional Mock-up Interface) was designed in order to make cooperation of different simulating tools and languages possible. The interface enables creating a model in one tool and running it in another. It also enables co-simulation, i.e. creating a model consisting of several interconnected submodels written in different simulation languages.

FMI defines interface in C language to be implemented by functions that represent specific model equations. Interface implemented with a particular model is called *FMU* (Functional Mock-up Unit). Functions of FMU are called by the numerical solver in every time step of the calculation. Models are distributed in zip archive, that contains in addition to FMU files also standardised XML file with description of the model. The FMU files in the archive might be either in source code in C or translated into DLL.

Many simulation tools support FMI. The most common are e.g. Dymola, Matlab/Simulink, JModelica and CATIA. FMI support in OpenModelica is in development.

4 Models on Web

Creation of e-learning applications for medicine [6] based on physiological models is one of the main goals of the Laboratory of Biocybernetics and Computer Aided Teaching at First Faculty of Medicine, Charles University in Prague. These applications are intended mostly for students of medicine, but applications for biology teaching at elementary and high schools will be developed, too.

Representation of the model should be as instructive as possible. Description with equations and visualisation of results in graphs is not suitable. The model is therefore supplemented with graphical user interface, that represents individual components of the model (usually organs), their mutual interaction and current state, see Figure 2. Students can understand the physiological meaning of the model and the principle of the demonstrated effects far better this way. The interconnected model and graphical interface is called *simulator*.

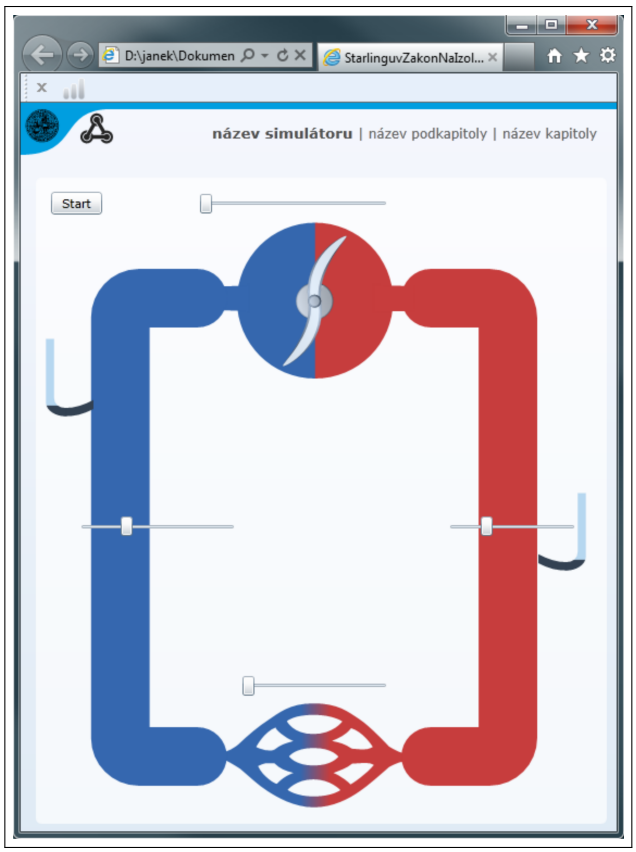


Figure 2: A simple systemic circulation simulator.

Modelica language is used to implement models in this laboratory. In order to remove barriers in simulator utilisation for students and teachers, they are being developed on the *.NET* platform. This enables for them to be run directly in the web browser [8], using plugin *Silverlight*, which is already a part of Windows[9].

To make this possible, the OMC (Open Modelica Compiler) has been expanded in such a way, that it enables model translation also into the C# language and the runtime and numerical solver have been implemented in the F# language. C# is directly supported by Silverlight. For F# support the F# runtime needs to be installed.

[8] http://www.modelisar.com/

[9] For Linux and Mac OS there is an alternative opensource plugin *Moonlight*.

Subsequently, all models run directly in the web browser without a need for further installations.

A long term goal of the laboratory is to build up a sophisticated web based repository of physiological models. A simulation centre was implemented to make possible solving any model of the repository directly in a web browser. This application allows user to set parameters and initial conditions of model and run the computation.

Selected quantities are plotted into a graph during computation. It is even possible to change parameters of the model on the fly. Creation of the web based Modelica editor has been proposed. Implementation of the new Modelica compiler in .NET languages would be too labour-intensive, therefore the translation will be realised on the server side.

This web-based simulation tool will be directly interconnected with the model repository. This will enable anybody to run the models as well as modify them or create new ones. By doing that the repository will continually be growing.

5 Conclusion

There are many languages suitable for physiological model description. Some of them are directly aimed at biological and medical simulations. Some languages and simulation tools are compatible and can cooperate. Increasingly more modelling tools support the FMI interface, which enables such compatibility. In order to make simulation tools and models even more widely available, a creation of an innovated model editor in Modelica language has been proposed. This editor will run in a web browser and will have a direct access to the web model repository. For this purpose, Modelica compiler was already extended and new simulation runtime, solver[10]and simulation centre was implemented.

Acknowledgments

The paper was supported by the project SVV-2011-262514 of Charles University in Prague.

[10]Author participate on implementation of the simulation runtime and solver.

References

[1] Anrew Finney, Michael Hucka, Benjamin J. Bornstein, Sarah M. Keating, Bruce M. Shapiro, Joanne Matthews, Benjamin K. Kovitz, Maria J. Schilstra, Akira Funahashi, John Doyle, and Hiroaki Kitano. System Modeling in Cellular Biology, From Concepts to Nuts and Bolts. The MIT Press, 2006.

[2] P.A. Fritzson. Principles of object-oriented modeling and simulation with Modelica 2.1. IEEE Press, 2004.

[3] Kofránek J. and Velan T. Components for Golem. Simuling and Control Web utilisation for creation of physiological functions simulator. (Komponenty pro Golema. Využívání simulinku a control webu při tvorbě simulátoru fyziologických funkcí. In Objekty 1999, pages 189–204, Praha, 1999. Česká zemědělská univerzita.

[4] M. E. Jackson and J. W. Gnadt. Numerical simulation of nonlinear feedback model of saccade generation circuit implemented in the labview graphical programming language. Journal of Neuroscience Methods, 87(2):137–145, 1999.

[5] J. Kofránek, M. Mateják, and Privitzer P. Causal or acausal modeling: labour for humans or labour for machines. In Technical Conmputing Prague, pages 1–16, 2008.

[6] J. Kofránek, S. Matoušek, J. Rusz, P. Stodulka, P. Privitzer, M. Matěják, and Tribula M. The atlas of physiology and pathophysiology: web-based multimedia enabled interactive simulations. In Computer Methods and Programs in Biomedicine, volume 104(2), pages 143–153, 2011.

[7] G. Lipovszki and P. Aradi. Simulating complex system and processes in labview. Journal of mathematical Sciences, 132:629–636, 2006.

[8] P. Privitzer, J. Šilar, M. Tribula, and J. Kofránek. From model to simulator in a web browser (Od modelu k simulátoru v internetovém prohlížeči). In MEDSOFT, pages 149–169, 2010.

[9] G. M. Raymond, E. A. Butterworth, and J. B. Bassingthwaighte. Jsim: Mathematical modeling for organ systems, tissues, and cells. The FASEB Journal, 21(6):827, 2007.

[10] Tommy Yu, Catherine M. Lloyd, David P. Nickerson, Michael T. Cooling, Andrew K. Miller, Alan Garny, Jonna R. Terkildsen, James Lawson, Randall D. Britten, Peter J. Hunter, and Poul M. F. Nielsen. The physiome model repository 2. Bioinformatics, 27(5):743–744, 2011.

Osteogenesis Imperfecta Type I-IV, the Collagenous Disorder of Connective Tissue in Czech Population

Lucie Šormová[1], **Ivan Mazura**[2]

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

[2]Center of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic

Abstract

Background: Osteogenesis imperfecta is an inherited disorder particularily of a human connective tissue. It is a worldwide extensive disorder regardless of age, gender or ethnic group. At present the disease includes nine clinically different types. Typical clinical features are brittle bones, high frequency of fractures and bone deformities. The other observed signs are blue sclera, dentinogenesis imperfect and otosclerosis. The first four types of the disease arise from mutations in collagen type I genes, composed from COL1A1 and COL1A2 chains. A result of these mutations is the production of shortened or structurally defective protein. Individuals affected by OI forms V to IX have mutations in proteins encoded by following genes: CRTAP, LEPRE1, PPIB, FKBP10. Collagenous types of the illness exhibit a broad range of severity depending on type and mutation localization in the structure of the collagen type I.

Objectives and Methods: The aim of this study is the description of the clinical forms of the disease, identifying mutations and polymorphisms of genes of the collagen type I by a molecular genetic analysis of genomic DNA of Czech OI patients.

Results: Currently in the Czech population there are described mutations and polymorphisms only of MLBR2 region, namely exons 31, 33 and 36 and introns 32 and 39, of the COL1A1 gene of 25 OI patients.

Mgr. Lucie Šormová

Conclusion: It is important to perform a further molecular genetic analysis of both collagen type I genes for the detection of the widest possible mutational spectrum for determination of possible genotype phenotype relationship of affected individuals.

Keywords

Osteogenesis imperfecta, collagen type I, COL1A1, COL1A2, MLBR, mutations

Correspondence to:

Mgr. Lucie Šormová
Address: Vančurova 2686,
54401 Dvůr Králové n. L., Czech Republic
E–mail: black.luca@seznam.cz

EJBI 2011; 7(1):59–64
recieved: September 15, 2011
accepted: October 24, 2011
published: November 20, 2011

1 Introduction

Osteogenesis imperfecta, type I-IV, is an inherited disorder of the connective tissue formation, especially of bones, joints and skin. Clinical features of this disease are bone fragility, high frequency of fractures, bone deformity, joint hypermobility, subnormal or short stature, dentinogenesis imperfecta (DI), bluish/greyish hue of sclera, hearing lose in adulthood, vascular, neurological and pulmonary complications and some other [30]. Presentation of these characters vary according to the type of a disease, but also within the same type of an illness. The incidence of non-lethal forms of the disease, so called OI tarda involving types I, III and IV, is reported in the range of 1: 25 000 to 1: 40 000 live births. The incidence of lethal type of this disorder - OI type II (known as OI congenita),

is featured in the ratio 1: 60 000 live births [10]. Currently, OI is classified into nine clinically different types (I – IX). Only the first four types are associated with collagen type I mutations. Collagen type I is the major protein of bone, tendon and skin. It is composed of two alpha1(I) chains (encoded by COL1A1 gene) and one alpha2(I) chain (encoded by COL1A2 gene). The mutations of these two genes have the result in decreased production of the protein or in synthesis of structurally defective collagen molecules [13].

2 Osteogenesis Imperfecta Classification

Based on clinical signs the first OI classification from David Sillence (created in 1979) distinguished four types of the disease (I-IV). In the past, related to the development methods of analysis, such as molecular-genetic techniques and histological findings, new forms were identified in the IV group of OI - OI type V-IX [13]. The disease exhibits a wide spectrum of clinical and radiological signs and varies in severity from mild to perinatal lethal forms. Inheritance is mainly autosomal dominant (AD), but there are some types with the autosomal recessive (AR) type of the inheritance [16].

2.1 OI Type I

This autosomal dominant type is the mildest form of OI. Patiens do not have deformed bones and achieve normal or smaller growth. Fracture frequency is constant during childhood, decreases after puberty and then increases after menopause of women and after the sixth decade of men. Mild scoliosis resulting from vertebral fractures is very common for this type [18]. Another usual but not characteristic sign is blue sclera (intensity does not change with age) [30]. Affected individuals can have dentinogenesis imperfecta, mild joint hypermobility, tendency to bruising and suffer from partial or total hearing loss [20]. Based on the presence of DI we distinguish OI type IA (absence DI) and OI type IB (presence DI). Patients diagnosed with the type IB may have mild bowing of long bones of limbs. OI type I is a result of COL1A1 or COL1A2 genes mutations [13, 15].

2.2 OI Type II

OI type II is the lethal type of OI with high frequence of stillborns and perinatal mortality (up to 80% infants die during the first week of life). Survival of the perinatal period is rare [4]. These individuals offen die of a lung failure. Their bones are severely deformed, multiple fractures occur already in the perinatal period. The extremities are significantly short. Infants have a triangular face, blue or grey sclera and extremly large and soft cranium [13]. Based on radiographic features we distinguish type IIA (characterized by short and deformed long bones of lower limbs, short, deformed and continuously expanded ribs, dark blue sclera, and macrocephaly), IIB (which is similar to the type A, but individuals have small head circumference, shallow orbit and white or bluish sclerae) and IIC (this type differs by the presence of deformities and low bone density especially of ribs and long bones of limbs) [1, 25]. OI types IIA and IIC are inherited by the AD manner and are caused by COL1A1 or COL1A2 genes changes [4]. OI type IIB results from CRTAP gene mutations. It is the autosomal recessive form of OI type II [1].

2.3 OI Type III

OI type III is the most severe form of OI. The first fractures occur in uterus and at birth. Patients have subnormal stature with short extremities compared to the body, deformed short and long bones. Other distinctive signs include a triangular face, DI, blue sclera which usually turn to white with age, barrel shaped chest, weak muscles and severe scoliosis [16]. Radiographic findings of infants demonstrate undermineralized calvarium with Wormians bones, of adult detect osteopenia and popcorn-like calcification, especially metaphyseal and epiphyseal. This calcification disrupts the growth plate and reduces growth of long bones, especially femurs. Metaphyses of long bones are broad, diaphysis are thin. Osteopenia and joint hypermobility often lead to kyphoscoliosis. Basilar impression can occur in some cases. The patients require a wheelchair and crutches. Genetic transmission is autosomal dominant. This OI type is caused by mutations in collagen type I genes [13].

2.4 OI Type IV

OI type IV is the highly heterogeneous form of this disease. There is considerable intra- and interfamilial heterogenity. Individuals can be mildly to severely affected. Their stature is variable short. The first fractures can occure at birth. Bones are mildly to severely deformed with popcorn-like calcification that is less common as in the type III. Patients have white sclera, although bluish and grayish color is also described.

Otosclerosis occurs in some cases. Based on presence of DI, OI type IV is divided into two types: type IVA (absence of DI) and IVB (presence of DI). Osteoporosis and scoliosis are common radiographic signs. Determination of basilar impression is more typical than in the type III. This type of OI disease results from autosomal dominant COL1A1 or COL1A2 mutations [12].

OI types V, VI, VII, VIII and IX are newly-described forms of this disorder. Their origin is not in mutations of collagen type I genes. The molecular nature of these types are mutations of genes FKBP10 (OI type VI), CRTAP (OI type VII), LEPRE 1 (OI type VIII) and PPIB (OI type IX). Genetic nature of osteogenesis imperfecta type V is

unknown. All of these types (V-IX) are inherited in an autosomal recessive manner [13].

3 Collagen Type I

Type I collagen is the most abundant protein of extracellular matrix in connective tissue, primarily in bones. It is a heterotrimer which is composed of two alpha1(I) chains, encoded by COL1A1 gene which is located at 17q21.3-q22 position of chromosome 17, and of one alpha2(I) chain that is encoded by COL1A2 gene on chromosome 7 at locus 7q21.3-q22.1. COL1A1 gene is consisted of 51 exons, the coding sequence of COL1A2 consists of 52 exons. Even so, genetic information of the alpha chains is the same size in both of these two genes because the amino acids 568-603 encode exon 33 of COL1A1 gene, in the COL1A2 gene there are the same amino acid residues encoded by exons 33 and 34 [7]. Structurally we distinguish three areas of collagen genes: *promoter*, the 5 part of the gene which includes the signal sequence for binding the RNA polymerase and which contains binding sites of transcription factors, *coding sequence* carrying the genetic information of the alpha chain, and *terminator*, the 3domain where the polyT sequence and the termination codons, the termination signals for DNA transcription, are situated (Fig. 1).

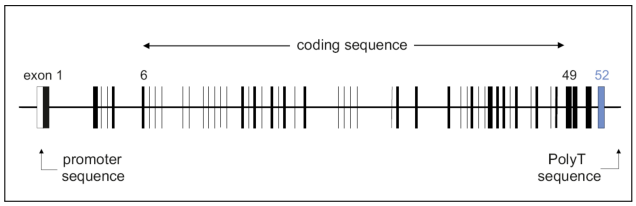


Figure 1: Structure of collagen of the type I genes. Vertical lines represent positions of exons. The exon number 52 (marked in blue) is found only in the COL1A2 gene.

There are some crucial areas of collagen type I genes promoters that influence DNA transcription. We rank to them transcription factors binding sites, activating proteins binding sites (such as YY1, c-Krox, IF1, IF2, AP1 and so on) stimulating or suppressing the transcription and CpG rich sequences of promoter (and of exon 1 and intron 1) whose methylation is prevented binding of transcription factors [9].

The first form of alpha chain, prepro-alpha chain, is produced by fibroblasts, osteoblasts or odontoblasts [9]. There are three domains within the structure of prepro-alpha chain: N-terminal propetide, encoded by exons 1-5 and part of exon 6, helical domain encompassing exons 6-49 and C terminal propeptide encoded by exons 50 and 51 and part of exon 49 [7]. The N-propeptide structure further consists of signal peptide, von Willebrand factor binding site and Col 2 binding site of cell-specific proteins (Fig. 2). Pro-alpha chains arise in the endoplasmic reticulum by splitting the signal peptide and then they are assembled into procollagen molecules. The folding process proceeds from the C– to the N–terminus [5]. The subsequent extracellular N– and C–propeptides separation (by N– and C–peptidase) creates the final molecule collagen type I, which is subject to other posttranslational modifications such as glycation or hydroxylation of amino acid residues [9, 23]. The final collagen monomer is terminated by N– and C–telopeptides [7].

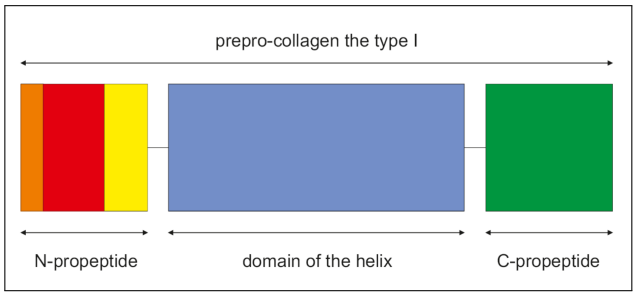


Figure 2: Structure of prepro-collagen typ I molecule. There are three domains within the N-propeptide structure: orange: area of the signal peptide - dysplayed in orange, von Willebrand factor binding site shown in red, and Col2 binding site yellow box of the N-propeptide of the prepro-collagen type I molecule.

The triple-helical region of alpha chains is composed of 338 repeating Gly-X-Y sequences, in which Gly is glycine, X is frequently proline and Y is often hydroxyproline. It follows that the amino acid glycine is crucial for correct folding of collagen monomers due to inter-chains links production [28]. The main function of the proline and hydroxyproline is to stabilize the elongated nature of the alpha chains and to increase denaturation temperature of the protein [3, 24].

Collagen monomers are assembled into the collagen microfibrils and these into collagen fibrils. The basic repeat structure of collagen fibrils is called D-period. This period contains whole sequence of the monomer. It is 67nm long and is composed of one overlap and one gap zone (Fig. 3) [8].

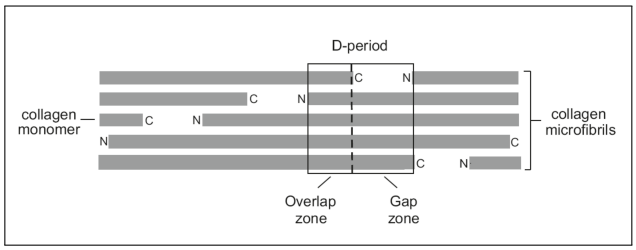


Figure 3: D-period of the type I collagen fibril.

3.1 Multi Ligand Binding Regions (MLBR) of the Collagen Type I Protein

There are several ligand binding sites situated at the level of collagen monomer. There are three hot spots on the alpha1 and alpha2 chains defined by codons 81200

(MLBR1), 682830 (MLBR2) and 9211040 (MLBR3) [26]. These sequences bind *integrines* that bind extracellular matrix molecules [26], *keratan sulphate proteoglycans* and *dermatan sulphate proteoglycans*, regulators of the fibrogenesis and formation of the inter-fibrils interactions and protectors of the fibrils against proteolytic damage [20]. *COMP* (Cartillage oligomeric matrix protein), *fibronectin* ensuring collagen type I molecules adhesion [8], *Hsp47* (heat-shock protein 47) that serves to thermal stabilization of the helix during procollagen synthesis, helps to folding and assembly of procollagen molecules and participates in the transport of structurally unaffected molecules from the endoplasmatic reticulum [27], belong also to ligands of collagen of the type I. Finally, the other important extracellular matrix proteins interacting with the collagen type I molecules are *phosphophoryn* inducing dentin matrix mineralization [6], *osteonectin* that binds extracellular matrix proteins, regulates production and storage of some extracellular matrix molecules or inhibits cell cycle [2, 11], *von Willebrand factor*, protein affecting platelet function [22], and some more.

Generally, binding of extracellular matrix molecules to molecules of collagen of the type I increases strenght and elasticity of bone tissue [26].

4 Molecular Basis and Genotype-Phenotype Correlation of Osteogenesis Imperfecta, Type I-IV

Types I-IV of osteogenesis imperfecta are the result of collagen type I genes (COL1A1 and COL1A2) mutations. Essentially, mutations of these two genes manifest in two ways: 1) synthesis of a decreased number of alpha chains, 2) production of structurally defective protein.

Production of a decreased amount of collagen fibrils is associated with nondeforming OI type I. This type of OI disease results from null mutations, a single nucleotide substitutions, that lead to the STOP codons formation (presence of STOP codons terminates DNA transcription). Decreased expression of the protein also may result from genetic changes in splice-sites of pre-mRNA, provided that these one result in intron retention in mRNA or in STOP codon formation [16].

Deforming types of osteogenesis imperfecta, types II, III and IV, arise based on mutations affecting the structure of collagen. These changes are in 80% missense mutation (glycine substitution), the remaining 20% are frameshift mutations including deletion/insertion of one and more nucleotides (number not divisible by three) and splice-site mutations resulting in exon skipping or in new splice-sites production [13].

In terms of genotype-phenotype relationship, several links were defined. Severity of the disorder increases from N to C terminus of alpha chains. This relationship applies to MLBR regions. Specifically, MLBR 1 mutations result in mild to severe forms of OI while clinical pictures of MLBR 2 and 3 changes are primarily OI II and III types. Furthermore, there are eight lethal regions of the alpha2 chain (all of them are located in proteoglycans binding sites of the collagen fibrils) [26]. Regardless, lethal mutations are located rather in the alpha1 chain (about 35.6% glycine substitutions cause lethal OI) than in the alpha2 chain (only 19% are lethal) [13].

5 Current Knowledge of Osteogenesis Imperfecta Treatment

Treatment of patients diagnosed by OI includes to using medicaments, surgery, orthotic treatment and rehabilitation. Recently, in the medical treatment the most widely used medicament are bisphosphonates. Effect of the bisphosphonates is reduction of the bone turnover with subsequent increase of the bone density but not the improving of the structure of the collagen type I molecules [13, 19].

Surgical treatment is mainly performed to correct deformities and to reduce the bone brittleness as the result of bad bowing and to improve the physical condition of the individual. It includes osteotomy, intramedullary fixation due nails, wires, pins and other or spinal fusion with Harrington rod [20, 29]. Surgical intervention is also one of the possible solutions of otosclerosis in which the patients undergo stapedectomy (surgical removal of the stapes). The non-invasive therapy involves the use of orthotic trunk and limbs orthoses to correct scoliosis and mild limb deformations, such as genua valga or vara. Individuals suffering from hearing loss use a cochlear implant to improve its hearing. Patients are also advised to perform light physical activity (swimming, walking in water, Nordic walking) to strengthen the weakened muscles [17].

Recently gene and cell therapies are a current issue in the treatment of this disease. The essence of cell therapy is the transplantation of bone marrow matched donors. Normal osteoblasts formed by marrow donors have the ability to replace the mutant osteoblasts. OI patients who have undergone a cell therapy show an increase in a bone mineral content and an increase in body height. The aim of a gene therapy is to prevent expression of mutated alleles. This is achieved by binding of complementary antisense DNA/RNA fragments or hammerhead ribozymes to abnormal pre-mRNA. The intention is to prevent translation of this pre-mRNA and its subsequent degradation. A gene therapy using these mechanisms results in the conversion of severe types of OI in mild forms of the disease. Another approach in a gene therapy can be modification of mesenchymal stem cells of OI patients in vitro and consequent returning of these cells to the individuals. Using a gene therapy to treat OI currently has one obstacle. That is the small number of known mutations of genes of collagen type I. Thus, the creation of fragments, the use of

rRNA, or modification of stem cells of affected individuals is currently very difficult because of the high mutation spectrum of OI [13, 14].

6 Conclusion

Osteogenesis imperfecta is a heterogeneous disorder with a wide spectrum of clinical characters and a large genetic diversity. Determination of genotype-phenotype relationship is a permanent problem because the same mutation may present different phenotypes among unrelated individuals also in members of one family with the same form of this disorder. The reason is the wide spectrum of clinical signs of identical mutations. Currently, 10% of all mutations that alter the glycine codon is described. In the future it is important to detect a lot of collagen of the type I mutations using the molecular genetic analysis to determination of the genotype-phenotype correlation in patiens diagnosed by OI types I-IV. For this purpose the method of laser microdissection of affected tissue can be also used. This method can detect a specific mutations affecting bone formation.

The analysis of genes of the collagen type I should be aimed primarily at multi ligand binding regions (MLBR1-3), because changes in these sequences may prevent the creation of intra- or extramolecular bonds important for the quality of a bone structure, and at COL1A1 and COL1A2 regions of participating genes, important for initiation and process of the transcription.

Recently, the molecular genetic analysis (comprising polymerase chain reaction (PCR) and double-sided sequencing) focused on the part of the COL1A1 gene, containing MLBR2 region, in 25 Czech patients affected by OI type I-IV had carried out. We observed mutations in DNA samples of seven Czech patients. Four of these patients are affected by OI type IA, one patient suffering from OI type III and two patients were diagnosed with OI type IVB. All determined changes in our sample collection are single nucleotide mutations that result in either amino acid substitution, STOP codon production or the mutations do not alter the reading frame. Mutations of coding sequence were observed in exons 31, 33 and 36. Of these mutations only Gly523Cys, Gly526Cys and Arg519STOP were described in literature. Two non-coding sequence modifications were observed in introns 32 and 39. Both of these two intronic mutations were detected in two patients affected with OI type IA and in one patient diagnosed with OI type IVB.

Currently, we collect a biological material (venous blood and bone grafts) of Czech OI patients for the molecular genetic analysis of other important regions of the COL1A1 gene and for subsequent COL1A2 gene analysis. In the future it is important to perform the molecular genetic analysis of complete sequences of both collagen type I genes and subsequently to compare the clinical manifestations of the disease in patients having the same form of OI and the same change in DNA. This is crutial for the correct diagnosis of the type of the disease and for provision of timely treatment of affected patients to limitation potential health problems associated with osteogenesis imperfecta.

Acknowledgments

Acknowledgements belong to Doc. MUDr. I. Mařík, CSc. and to MUDr. Olga Hudáková, Ph.D. for the provision of useful informations regarding the clinical description of individual forms of the disease. The work was supported by the CBI project No. IM06014 of Ministry of Education, Youth and Sport CR and by the project SVV-2011-262514 of Charles University in Prague.

References

[1] Barnes AM, Chang W, Morello R, Cabral WA, Weis M, Eyre DR, et al. Deficiency of cartilage associated protein in recessive lethal osteogenesis imperfecta. New Eng J Med. 2006; 355: 2757-2764.

[2] Bradshaw AD, Graves DC, Motamed K, Sage EH. SPARC-null mice exhibit increased adipozity without signifiant differences in overall body weight. PNAS. 2003 May 13; 100(10): 6045-6050.

[3] Burjanadze TV, Veis A. A thermodynamic analysis of the contribution of hydroxyproline to the structural stability of the collagen triple helix. Connect. Tissue Res. 1997; 36: 347-365.

[4] Byers PH, Tsiopouras P, Bonadio JF, Starman BJ, Schwarz RC. Perinatal lethal osteogenesi imperfekta (OI type II): a biochemically heterogenous disorder usually due to mutations in the genes for the type I collagen. Am J Hum Genet. 1988; 42: 237-248.

[5] Cabral WA, Chang W, Barnes AM, Wies MA Scott MA, Leikin S, et al. Prolyl 3-hydroxylase 1 causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta. Nat Genet. 2007 Mar; 39(3):359-365.

[6] Dahl T, Sabsay B, Veis A. Type I collagenphosphophoryn interactions: specificity of the monomermonomer binding. Journal of Structural Biology. 1998 Oct; 123(2): 162-168.

[7] Dalgleish R. The human type I collagen mutation database. Nucleic Acids Res. 1997; 25:181 187.

[8] Di Lullo GA, Sweeney SM, Krkk J, Ala-Kokko L, San Antonio JD. Mapping the ligand binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. J Biol Chem. 2002 Feb 8; 277(6): 4223-4231.

[9] Ghosh AK. Factors Involved in the Regulation of Type I Collagen Gene Expression: Implication in Fibrosis. Exp Biol Med. 2002; 227(5):301-314.

[10] Hudáková O, Mařík I, Zemková D, Šedová M, Mazura I, Kuklík M. Osteogenesis imperfecta se zaměřením na antropologickou charakteristiku onemocnění a diferenciální diagnostiku jednotlivých typů. Pohybové ústrojí. Pokroky ve výzkumu, diagnostice a terapii. 2007; 14(3-4), Supplementum: 321-324.

[11] Jorgensen LH, et al. Secreted protein acidic and rich in cysteine (SPARC) in human skeletal muscle. Journal of Histochemistry Cytochemistry. 2009; 57(1): 29-39.

[12] Kashyap RR, Gopakumar R, Gogineni SB, Sreejan CK. Osteogenesis imperfecta type IV. Kerala Dental Journal. 2009 Jan; 32(1): 47-49.

[13] Marini JC. Osteogenesis imperfecta. 2010. Available at: http://www.endotext.org/parathyroid/parathyroid17/parathyroid17.pdf. (Revised 1 March 2010).

[14] Niyibizi C, Wang S, Mi Z, Robbins PD. Gene therapy approaches for osteogenesis imperfecta. Gene Therapy. 2004; 11: 408-416.

[15] Paterson CR, McAllion S, Miller R. Heterogenity in osteogenesis imperfecta type I. J Med Genet. 1983; 20: 203-205.

[16] Primorac D, Rowe DW, Mottes M, Barii Antii Mirandola S, Lira MG, Kalajzi Kuec V, Glorieux FH. 2001. Osteogenesis Imperfecta at the Beginning of Bone and Joint Decade. Croatian Medical Journal. 2001; 42(4): 393-415.

[17] Rauch F, Plotkin H, Zeitlin L, Glorieux FH. Bone mass, size, and density in children and adolescent with osteogenesis imperfecta: effect of intravenous pamidronate therapy. Journal of Bone and Mineral Research. 2003; 18(4): 610-614.

[18] Rauch F, Glorieux FH. Osteogenesis imperfecta. Lancet. 2004; 363: 1377-1385

[19] Rodan GA, Fleisch HA. Bisphosphonates: mechanism of action. J Clin Invest. 1996 Jun; 97(12): 2692-2696

[20] Roughley PJ, Rauch F, Glorieux FH. Osteogenesis imperfecta clinical and molecular diversity. European Cells and Materials. 2003; 5: 41-47.

[21] Roughley PJ. The structure and function of cartilage proteoglycans. European Cells and Materials. 2006; 12: 92-101

[22] Ruggeri ZM. Von Willebrand factor. Vascular biologi. 2003 Mar; 10(2): 142-149.

[23] Shegg B, Hlsmeier AJ, Rutschmann Ch, Maag Ch, Hennet T. Core Glycosylation of Collagen is initiated by two 1-O)galactosyltransferases. Mol Cell Biol. 2009 Feb; 29(4): 943-952.

[24] Shoulders MD, Raines RT. Collagen structure and stability. Annu Rev Biochem. 2009; 78: 929-958.

[25] Sillence DO, Barlow KK, Garber AP, Hall JG, Rimoin DL. Osteogenesis imperfecta type II: delineation of the phenotype with reference to genetic heterogeneity. Am J Med Genet. 1984; 17:407-423.

[26] Sweeney SM, Orgel JP, Fertala A, McAuliffe JD, Turner KR, Di Lullo GA, et al. Candidate cell and matrix interaction domains on the collagen fibril, the predominant protein of vertebrates. J Biol Chem. 2008 Jul 25; 283(30): 21187-21197.

[27] Tasab M, Batten MR, Bulleid NJ. Hsp47: a molecular chaperone that interacts with and stabilizes correctly-folded procollagen. The EMBO Journal. 2000; 19(10): 2204-2211.

[28] Vilím V. Imunochemické možnosti sledování degradace kolagenu typu II. Česká revmatologie. 2007 Mar; 15(1): 3-12.

[29] Vyskočil V, Pikner R, Kutílek S. Effect of alendronate therapy in children with osteogenesis imperfecta. Joint Bone Spine. 2005 Oct; 72(5): 416423.

[30] Wollina U, Koch A. Osteogenesis imperfecta type I and psoriasis a report on two cases. Egyptian Dermatology Online Journale. 2006 Jun; 2(1):15.

Utilization of Custom-Made Databases in Both Medical Research and Patients' Treatment

Zdeněk Telička[1], **Jan Jiskra**[1], **Josef Kubinyi**[2]

[1]3rd Medical Department, Charles University and General Faculty Hospital in Prague, Czech Republic

[2]Department of Nuclear Medicine, Charles University and General Faculty Hospital in Prague, Czech Republic

Abstract

Background: Hospital Information Systems widely used in departments of university hospitals are not sufficient for both storing data about patients treatment and long-term research. Clinicians often use custom-developed applications which are maintained without any cooperation with the management of the hospitals and mostly break law in the Czech Republic.
Objectives: This article describes using such an application in cooperation with the Hospital Information System. It also describes an example of a research of cost effectiveness thyroid gland diseases treatment using Radioiodine ^{131}I in outpatient regime compared to hospitalization or an alternative surgery.
Methods: The database application is developed in Visual Basic. The research studies the treatment by the Radioiodine ^{131}I in 45 patients. We evaluated the financial cost of radioiodine therapy in the outpatient regime and hospitalization compared with a surgery.
Results: The financial costs for 1 patient is 114 EUR, it means 16% if compared with the same treatment in a hospital and only 25% of the possible alternative operation.

Mgr. Zdeněk Telička

Conclusion: This study describes that the treatment by outpatient regime can be a motivating alternative compared to the treatment by ^{131}I at a hospital or even the surgery.

Keywords

Thyroid gland, cost effectiveness, Radioiodine, hospital information system, database

Correspondence to:

Mgr. Zdeněk Telička
3rd Internal Department of the 1st Faculty of Medicine,
Charles University in Prague and General Faculty Hospital in Prague
Address: U Nemocnice 1, 128 08 Prague 2, Czech Republic
E–mail: zdenek@telicka.cz

EJBI 2011; 7(1):65–68
recieved: October 17, 2011
accepted: November 4, 2011
published: November 20, 2011

1 Introduction

At the General University Hospital in Prague, there is used one Hospital Information System (HIS) for managing the patients' health documentation and for storing the laboratory results. The largest problem of the HIS is in its universality, because it has limitations when used at university hospitals, where the physician also needs to use some information system for long-term research purposes.

Since the HIS stores the progress of the patients' treatment in a form of text-based medical records, nearly each physician at various departments maintains its own duplicate databases in various forms, from easy made spreadsheets tables to some small file-based databases. Advance of these custom-made databases is in its structure.

The data are stored in particular fields and clinicians find this solution much more useful for evaluating some research data in patients watched in long-term studies. Also we see two disadvantages:

- Maintaining of custom databases at hospitals by clinicians is often without any cooperation with the

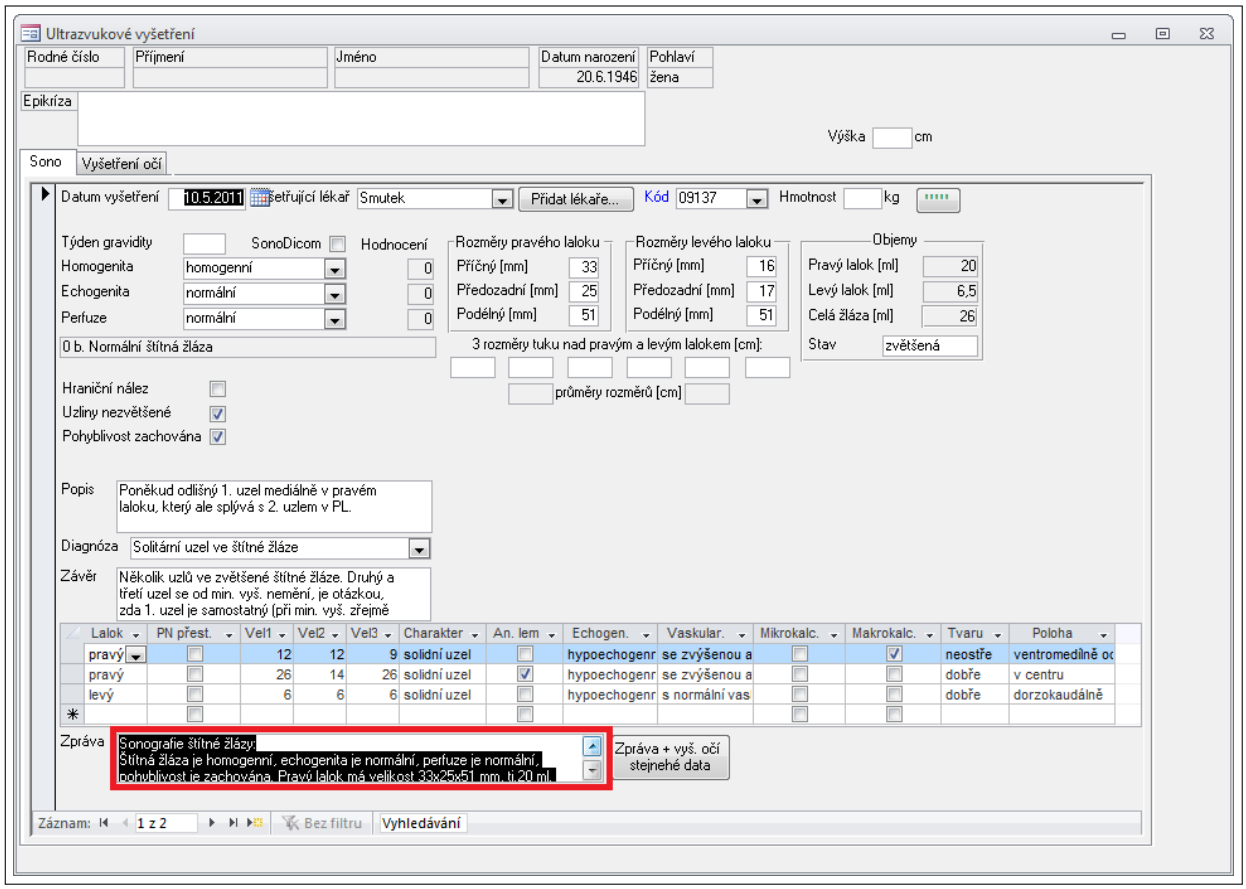


Figure 1: Ultrasonography module.

management and the IT department of the hospitals. This may lead to violation of law [1].

- Duplicity of the physician's work. The physician is using two applications simultaneously - Hospital Information System and his/her own application. Those applications mostly are not connected in any way, so each record must be manually typed twice, etc.

In this work I would like to show that custom-made databases can be very helpful for both research and patients treatment without any large requirements for clinicians work with PC.

2 Example of Using the Application During Medical Treatment

One of the examples is the module developed in the custom application, the ultrasonography module. During the ultrasonography examination, the clinician uses mainly the developed module which is connected with the Graphical client of the Hospital Information System. We used the standard programming interface of the Windows API. This gives the module ability to read data from opened record in the Hospital Information System [2]. This provides a clinician a comfort in using these two applications together. The clinicians simply open the patient's record in the HIS and the developed module can automatically search this record in its own database or create a new one.

Also, this application has a secured database. It is saved in a hospital data-store system which is accessible using the Windows Domain. So only the selected physicians have access to the data. This was performed with the cooperation with the IT Department of the hospital.

For clinicians, the advantage of this custom-made module is in two important things:

1. Data of the graphic examinations are stored in quantitative-based structural information. These results can be easily searched for any criteria. This is not currently possible in the HIS since the medical record is based on natural text-based sentences.

2. The clinician also needs to write the medical report in the HIS. This is performed automatically in the developed module. Using the Visual Basic programming language, there is automatically generated medical record in a natural language with data based from the application. We used fact, that the ultrasonography examination is mostly a repetitive work, mostly with rotation of given terms. Of

Table 1: Cost of particular treatments (Prices in EUR using 25 CZK/ 1 EUR exchange rate).

	^{131}I	^{131}I	Surgery
	outpatient regime	hospitalization	
Accumulation test	6 EUR	6 EUR	
Scintigraphy of thyroid using ^{98m}Tc	29 EUR	29 EUR	
Therapeutic 550 MBq ^{131}I	79 EUR	79 EUR	
Hospitalization TOTAL		585 EUR	314 EUR
Surgery (total thyroidectomy)			143 EUR
TOTAL	114 EUR	699 EUR	457 EUR

course, this is not applicable each time. So the clinician has always ability to complete the medical record as he/she wants. See Figure 1 including the generated medical report from the structured data.

3 Example of Using the Application for Long-Term Studies

We found that the database application can be used in types of research which requires long time (more than 2 years) of getting data or observation of the patients for such a research. And what is important, when the observed cases or methods of treatment are not very common, for example when the incidence of cases is in a number of 20 per year. We found very hard to maintain the data of those patients in the HIS. Mostly the data are unsearchable in the HIS after years when the department treats hundreds of patients per year.

3.1 Thyroid Treatment

The thyroid treatment by application of radioiodine ^{131}I in the thyroid is clinically used from the 40th of the last century. Indications for the treatment are for example differentiated carcinoma of the thyroid gland or thyroid hyperfunction [3]. This type of a treatment is widely used around the world and is certified as a save method by multiple studies [4] but in the Czech Republic, however, this method is still indicated relatively little and still is realized only in hospitals. The most used method of treatment is the surgery and the number of the noninvasive treatment is increasing very slowly [5]. The 3rd Medical Department together with the Department of Nuclear Medicine at 1st Faculty of Medicine is cooperating on patients' treatment by application of radioiodine ^{131}I without any hospitalization. We tried to evaluate the cost of the treatment together with the effectiveness of the treatment. Because the number of patients is very low even during 2 years of the study, we used the custom-made database application to store the research data. The reason was due to a large number of patients treated at the Endocrinology Department and after some years, it is not possible to find out all the observed patients in hundreds of patients.

3.2 Methods

This research took 22 months during the years 2008-2010. The treatment by Radioiodine ^{131}I was offered to 45 patients and finally it was performed on 39 patients (31 women, 8 men). We evaluated the financial cost of the radioiodine therapy in the outpatient regime and hospitalization compared with surgery.

The diagnostic and treatment by Radioiodine ^{131}I consist of those steps:

- Laboratory examination of the thyroid gland (TSH - thyroid stimulating hormone, FT4 - free thyroxine, FT3 - free triodothyronine, TPOAb - antibodies against thyroid peroxidase and TRAK - TSH receptor antibodies in serum using chemiluminescence analyzer by Centaur).
- Ultrasonography of thyroid nodules in patients with fine needle biopsy (FNAB) to exclude malignancy.
- Patients were given a fixed activity of 550 MBq in the form of Radioiodine ^{131}I capsules.
- Clinical and laboratory monitoring of thyroid parameters for 4-6 weeks and 2-6 months depending to actual health status of the patient.

Table 2: Contraindications using ^{131}I treatment.

a.	Pregnancy and breast-feeding
b.	Middle and hard endocrine orbitopathy
c.	Pregnancy sooner than 6 months before the ^{131}I application
d.	Malignant node in thyroid

We retrieved the cost of particular treatments, see Table 1.

3.3 Results

For the costs of treatments, we counted the Radioiodine accumulation test, thyroid scintigraphy with ^{99m}Tc Pertechnetate, the price of ^{131}I therapy capsules (according to the current price list of suppliers of radiopharmaceuticals) and in hospitalized patients, cost of hospitalization. Similarly, the cost of a thyroid surgery was evaluated according to the current tariff paid by insurance companies plus the cost of hospitalization (see Table 1). We used 25 CZK for 1 EUR exchange rate. By comparing the financial costs for 1 patient, we found that the cost of an outpatient radioiodine therapy (114 EUR) represents only 16% of the same treatment in a hospital and only 25% of the possible alternative operation.

3.4 Discussion

The treatment using ^{131}I has also some side effects, such as the post-radiation thyroiditis or development of the Graves-Basedow disease is up to 5% of patients. In patients with light orbitopathy the side effect can be treated by the low amount of prednisone, but in patients with developed orbitopathy it is determined as contraindication [6]. Other contraindications are described in the Table 2. The Radiodine treatment is not widely developed in the Czech Republic. Mostly there is a large distrust at this type of treatment in our country. This study describes that the treatment can be motivating alternative compared to the treatment by ^{131}I at a hospital or even the surgery. Another motivating aspect can be the effectiveness of the treatment, which is comparable with foreign countries (88 %) [7].

Acknowledgments

Research was supported by the project SVV-2011-262514 of Charles University in Prague and the grant from the Czech Health Ministry IGA NS10595-3.

References

[1] Law of the Czech Republic n. 480/2004 from 29th July 2004: Zákon o některých službách informační společnosti a o změně některých zákonů (zákon o některých službách informační společnosti)

[2] Roman, S.:Win32 API programming with Visual Basic, 1999: O'Reilly Media, p. 120-130

[3] Vlček P. Karcinom štítné žlázy, pooperační sledování nemocných: Vlček P, Neumann J. Karcinom štítné žlázy, Praha, Maxdorf, 2002: p. 220.

[4] Ross DS. Radioiodine therapy for hyperthyroidism. N Engl J Med, 2011, 364: 542-50.

[5] Horáček J. Terapie benigních onemocnění štítné žlázy radiojódem: Límanová et al. Štítná žláza, Trendy soudobé endokrinologie. Svazek 2, Praha, Galén, 2007, p. 281-297.

[6] Nygaard B, Faber J, Veje A, Hegedus L et al. Transition of nodular toxic goiter to autoimmune hyperthyroidism triggered by 131I therapy. Thyroid, 1999; 9: 477-81.

[7] Nygaard B, Hegedus L, Nielsen KG et al. Long-term effect of radioactive iodine on thyroid function and size in patients with solitary autonomously functioning toxic thyroid nodules. Clin Endocrinol (Oxf), 1999; 50: 197-202.

Cardio Online Reader – Conjunctions in Cardiology Knowledge

Miroslav Zvolský[1,2], **Vendula Papíková**[1], **Arnošt Veselý**[1]

[1]Centre of Biomedical Informatics, Institute of Computer Science, Academy of Sciences of the Czech Republic

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

Abstract

Background: The progress of scientific knowledge produces an excessive amount of scholarly literature. For effective clinical use it is necessary to use time, quality and relevance filtering.
Objectives: We decided to create an independent web based interface for searching the PubMed Database for recent evidence-based articles in the domain of cardiology.
Results: The Cardio Online Reader (COR) application (http://neo.euromise.cz/cor) offers an easy way to reach valid and relevant articles filtered by keywords, MeSH terms, authors, date of publication and type of publication. Results can be commented, rated, stored as citations and shared using most popular web sharing services.
Conclusion: Using COR can positively affect clinical decisions of physicians bringing them relevant and up-to-date scientific information.

MUDr. Miroslav Zvolský

Keywords

Electronic databases, Web 2.0, medical subject headings, decision support

Correspondence to:

MUDr. Miroslav Zvolský
Department of Medical Informatics, ICS, AS CR, v.v.i.
Address: Pod Vodarenskou vezi 2, Prague 8, Czech Republic
E–mail: zvolsky@euromise.cz

EJBI 2011; 7(1):69–74
recieved: October 20, 2011
accepted: November 1, 2011
published: November 20, 2011

1 Introduction

The progress of scientific knowledge in the field of medicine forces physicians in the clinical practice to keep in touch with recent scholarly literature. The most important communication medium for transfer the research results into the practice are scholarly articles in specialized expert journals. The amount of these articles grows in time and in last years the growing accelerates. Ways how to cope with this trend are not new [1, 2, 3], but the need of their practical use in the clinical practice is more and more urgent. Doctors in clinical practice need relevant information quickly and easily accessible [4].

Electronic databases are well arranged sources of information about biomedical articles published in an electronic or printed form. They collect metadata about articles published all around the world in the large number of journals.

The most important free accessible resource of biomedical science articles is the PubMed database, which is one of key services provided by the National Center for Biotechnology Information (NCBI). On the August 1, 2011 the PubMed database contained 21 067 999 article citations.

For acquiring the right (requested) result filtering from so big amount of articles it is necessary to describe the content of the article in right categories of metadata. On the other hand we require an easy to use and complex tool for searching through the database.

The web interface of the PubMed database [5] is accessible via http://www.ncbi.nlm.nih.gov/pubmed . It offers one html form field for searching the key terms in the

database. It puts the accent on the syntax of search query, for the definition should be defined precisely. The search engine returns often tens of thousands results when the search query consists of one or two key terms. The result list is sorted by the time of adding into the PubMed in a descendent order (by default), so the most recent articles come at first, but this order says nothing about qualitative parameters of articles.

The NCBI web pages also offer an advanced search tool for the PubMed database http://www.ncbi.nlm.nih.gov /pubmed/advanced. Users can define the search query in the PubMed Advanced Search in 39 parameters, which stores the PubMed database. Advanced Search saves a history of searched queries for each user. These queries can be repeatedly retrieved.

The web interface of Advanced Search is more complicated to use then one form field in the basic search. The definition of the query is time consuming and it needs an experience with search query formulation for obtaining high-quality results.

2 Objectives

Clinicians and healthcare participants should obtain the requested information for clinical decisions quickly and in an easy way. Our purpose was to simplify the process of obtaining searched articles in the stressing and time lacking situation of clinical practice. We wanted to allow any clinical worker or other interested person to gain high-quality results with no experience in large bibliographical database search. The only requested ability is to manage Internet browser.

Our aim is to provide a multidimensional view on the medical knowledge and its usability in the clinical practice.

As a data source we choose the PubMed database and we established an objective to create a new independent interface to filter out a specific part of resources (Evidence-based Medicine Categories, specific medical domain).

In the structure of metadata indexed in the PubMed database we selected those that represent one of dimensions we consider to be crucial in filtering relevant clinical information.

For the pilot project we assumed to concentrate on one medical domain cardiology. Articles focused on this domain are filtered in the PubMed search query definition by a specific set of Medical Subject Headings (MeSH) terms.

MeSH terms intermediate relations between articles each MeSH term of an article can link to other articles with the same MeSH keyword.

The type of publication representing the strength of evidence supporting the objective of the article is another dimension of the article. We limited publication types only to randomized controlled trials, systematic reviews and guidelines obtaining the strongest clinical evidence [6].

In order to bring the most relevant information and follow the dimension of time we want to order the list of results from the most actual articles.

For the future development we assume including some kind of regional relevance, that gives a new dimension to the knowledge collected all around the world. Some articles can be specific for concrete populations or healthcare systems. The native language of the article and the availability in the language of clinical user plays a very important role for usability of the paper in a particular clinical case or for a target person.

For demonstrating and easy understanding of the search results we want to use mostly graphical representation of the range and parameter specification and search results. In the future development we will prefer graphical MeSH cloud, map of the regional relevance or diagram of the most popular languages used in the papers rather than plain textual results.

3 Results

As a practical result of our efforts we developed the Cardio Online Reader (COR) web application that offers fully functional search interface for articles from Cardiology indexed in the PubMed database.

The actual version of the COR application is freely accessible on http://neo.euromise.cz/cor .

The database of citations and abstracts of biomedical science articles is the main part of the COR application. This database uses a MySQL database engine. The main data source for our project is the PubMed database, that can be used free of charge. Import of the data was realized by the query to the PubMed database defining the domain of Cardiology by using the most important MeSH terms from Cardiology. We filtered off articles, which do not fulfill our qualitative criteria from the Evidence Based Medicine point of view.

We selected only these types of articles:

- Randomized Controlled Trials,
- Systematic Reviews,
- Systematic Reviews with Meta-Analysis,
- Guidelines,
- Practical Guidelines.

The result of this query was saved in the XML file. Exported XML file was parsed by one-purpose PHP import script and selected data fields (title, authors, MeSH terms, abstract, unique identifier PMID, date of the abstract publication in the PubMed database, link to the full text, journal title) were saved to the COR database.

The actualization of the COR database proceeds daily by an automatically started PHP script, which uses tools from the Entrez Programming Utilities [7] for gathering metadata for each article published since the last actualization.

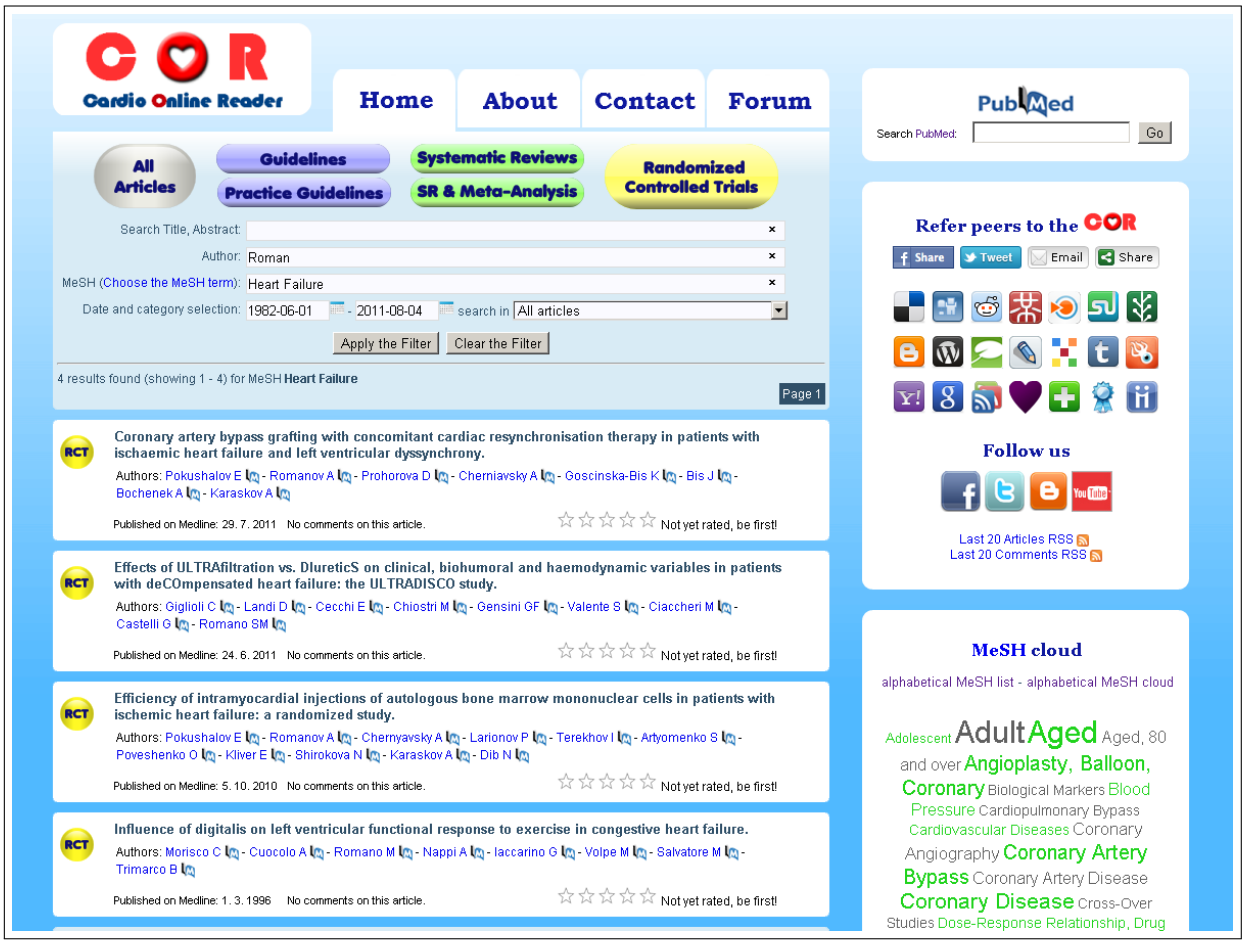


Figure 1: An example of COR application web interface showing the list of articles retrieved for parameters set in the filter form fields above.

3.1 The Web Interface of the COR

Contrary to the original PubMed interface we concentrated on the fastest way to reducing the number of search results preserving the focus on results important for the clinical practice.

Users can limit search results by one mouse-click to one category of high quality evidence. Users can also use six form fields of the filter on the home page of the COR application for entering search criteria. Individual form field stands for entering a part of text in the article title or abstract. Another form field specifies the requested author or authors. The third large form field stands for entering parts or exact full terms of Medical Subject Headings (MeSH) thesaurus. Users can write down requested MeSH terms or they can choose them from a generated

Table 1: Dimensions of clinical information and matching PubMed metadata used in Cardio Online Reader.

Dimension of clinical information	**Matching PubMed metadata category**
Medical Domain	filtered using MeSH
Related Medical Terms	MeSH
Evidence Based Medicine Categories	Publication Type
Time of Relevance	Publication Date
Regional Relevance	Country of the Journal Affiliation Country of the Author ? (external data) specific population description ? (external data)
Full text Availability	DOI, Full Text Link Out
Language	Language

Figure 2: An example of COR application web interface showing the detail of an article record.

MeSH Cloud or MeSH List, where terms are displayed in relation to their appearance in articles or sorted alphabetically.

In these form fields it is possible to use logical operators AND and OR. It is also possible to use a dynamically generated autocomplete function in these three fields to simplify entering exact phrases.

Users can limit the list of search results by setting the lowest and the highest date of publishing the article in the PubMed database in next two form fields. The dates can be set manually or chosen from the JavaScript date picker.

The last form field stands for the manual choice of the category of EBM quality of evidence.

For the fast choice of most frequented MeSH terms and their insertion to the filter, there is a "MeSH cloud" in the right part of the application web page, where enlisted terms differ in the text size displaying frequency of each term in the database. Users can use the list of last search queries.

3.2 Search Results

The COR application display search results matching entered parameters below the filter. Search results are in descending order sorted by the date of publication in the PubMed database. The simple list of results shows an article title, names of authors and the date of publication in the PubMed database. There can be a maximum of 15 results on one page. User can browse through the result pages. The category of an article is indicated by a graphical icon.

By clicking on the article title in the list of results user can navigate to the detail page of the article record. The detail page shows the article title, names of authors, list of assigned MeSH terms, PMID identifier, link to the original record in the PubMed database, link to the full text of the article (if available on the Internet), article abstract, journal title and the date of publication in the PubMed database. The User of the COR application can rate the helpfulness of the article in the scale from one to five star symbols. This rating is linked to the IP address, so one user can rate a single article only one time. User can also attach a comment to the article.

3.3 Web 2.0 Social Functions

The second big task for the COR application is to allow easy sharing of clinically important results with colleagues, friends and professional community. The detail of

the article including abstract and bibliographical data can be shared via email, Facebook, Twitter and other social networks contained in the Share this web service [8].

Excepting a particular scientific article detail, the COR offers an easy way to share a link to itself via Share this service, Facebook, Twitter, e-mail or one of 21 most common social and bookmarking services like Digg, Delicious, Reddit, Yahoo! or Google Bookmarks.

Users can follow a CORs own profile on Facebook and Twitter, Blogger account and YouTube channel. Users can subscribe to RSS channels with last 20 articles or last 20 comments generally or individually for each type of article.

3.4 Future Plans and Improvements

We plan further improvements and simplifications in the web interface of the COR application in the future. One thing which can speed up using the filter and make the work more illustrative is to place a graphical slider and the time plot showing numbers of articles published in the discrete time periods and their selection in the filter.

A long term problem is to optimize the autocomplete function in three form fields in the filter to help users in inserting key terms in the easiest way. This process should be evaluated in the cooperation with common users.

Geotagging can help to make search results more regionally-oriented. Metadata contained in the PubMed database can show in which country the article was published. Geographical information in the field Affiliation is even more interesting. It is possible to find out where the article was created and what population is in the article described. We can draw this information in the map or allow its limitation in the filter.

We assume an individualization of the web interface for registered users in the future development of the COR application. After the registration process and logging in a user could browse the history of own search queries, create lists of favorite articles, let the system send him notification on some events in the database or define own RSS channels or add authorized comments and ratings.

The COR application not only can serve users of the web interface or RSS readers. The same problem as physicians have today with obtaining relevant clinical information will rise in the future for decision support systems (DSS) that can improve practitioner performance [9]. Even DSS algorithms can easily complete well-structured PubMed search query and ask directly the source database, we should prepare computer readable results of the COR application as a simple and independent alternative of the PubMed Entrez Utilities.

By creating an XML data interface we can connect the clinical information system and send it search results or record details on its demands. Possible service for hospital information systems could offer relevant document for the concrete clinical situation defined by MeSH and geographical terms.

4 Discussion

The widely accepted PubMed database of biomedical bibliographies of biomedical publications has a free accessible web interface with a basic or advanced version of the search. We can use other web services for searching biomedical databases for scientific articles by clinical terms or other parameters. These services are more general (Google, Google Scholar) or focused on natural sciences (Scopus).

Why to create another search tool?

The amount of scientific articles indexed in electronic databases increase steeply. A recent question "where to find" has been replacing by questions "how to search" and "how to search the easiest way". The COR offers simple and fast way how to search the PubMed database for articles in the field of Cardiology and with the focus on highest evidence. It copes with only one thousandth of the PubMed database and provides easy-to-use tools for setting the search query that can acquire small amount of articles appropriate to the clinical need.

Other web services are specialized on searching in large databases of scientific bibliography (Vivisimo, Tripdatabase, Pubmeddy - discontinued). The COR is unique in its focus on one domain (Cardiology), on few defined EBM categories most important for clinical practice and in the simplicity of use.

The key question for the progress of the COR application will be the interest of expert medical community. The COR contains tools for sharing scientific information between experts, tools for subjective rating of their quality and tools for expert discussion. Experts could be motivated by functions for registered users, individualized search functions, the ease of use and the fact, that this application is free of charge. The COR application is designed for Cardiology specially.

Filters used for extraction from the PubMed database are firmly defined. The same technology could be used for another one purpose (one expert domain) web portal. Similar filters could be also individually set for registered users, so the scope could be widened to other domains.

5 Conclusion

As an alternative to the widely accepted PubMed search web tool we created the Cardio Online Reader web application. It concentrates on the domain of Cardiology. The COR uses the free accessible PubMed database of biomedical scientific articles and ads an easy-to-use search web interface and functions for sharing, rating and commenting articles.

This application is freely accessible via http://neo.euromise.cz/cor .

Acknowledgments

This work was supported by the project 1M06014 of the Ministry of Education Youth and Sport CR and by the project SVV-2011-262514 of Charles University in Prague.

References

[1] Huth EJ. The information explosion. Bull N Y Acad Med. 1989 Jul-Aug;65(6):647-61; discussion 662-72.

[2] Moore M. Battling the biomedical information explosion: a plan for implementing a quality filtered database. Med Ref Serv Q. 1989 Spring;8(1):13-9.

[3] Eysenbach G, Diepgen TL. Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. BMJ. 1998 Nov 28;317(7171):1496-500.

[4] Smith R. What clinical information do doctors need? BMJ 1996 Oct 26: 313 : 1062

[5] PubMed: MEDLINE Retrieval on the World Wide Web United States National Library of Medicine 2002-06-07 http://www.nlm.nih.gov/pubs/factsheets/pubmed.html

[6] Eddy DM. Evidence-based medicine: a unified approach. Health Aff (Millwood). 2005 Jan-Feb;24(1):9-17.

[7] D. Wheeler,T. Barrett,D. Benson,S. Bryant,K. Canese,V. Chetvernin,D. Church,M. DiCuccio,R. Edgar,S. Federhen,L. Geer,Y. Kapustin,O. Khovayko,D. Landsman,D. Lipman,T. Madden,D. Maglott,J. Ostell,V. Miller,K. Pruitt, G. Schuler,E. Sequeira,S. Sherry,K. Sirotkin,A. Souvorov,G. Starchenko,R. Tatusov,T. Tatusova,L. Wagner, E. Yaschenko Database resources of the National Center for Biotechnology Information, Oxford University Press 2006.

[8] ShareThis http://www.sharethis.com

[9] Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005 Mar 9;293(10):1223-38.

