EJBI 2012 ISSN 1801 - 5603

European Journal for Biomedical Informatics

Volume 8 (2012), Issue 5

Special Topic

Semantic Interoperability in Biomedicine and Healthcare III

Editors

Štěpán Svačina, Jana Zvárová

An Official Journal of the European Federation for Medical Informatics

www.ejbi.eu

EFMI European Federation for Medical Informatics

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.

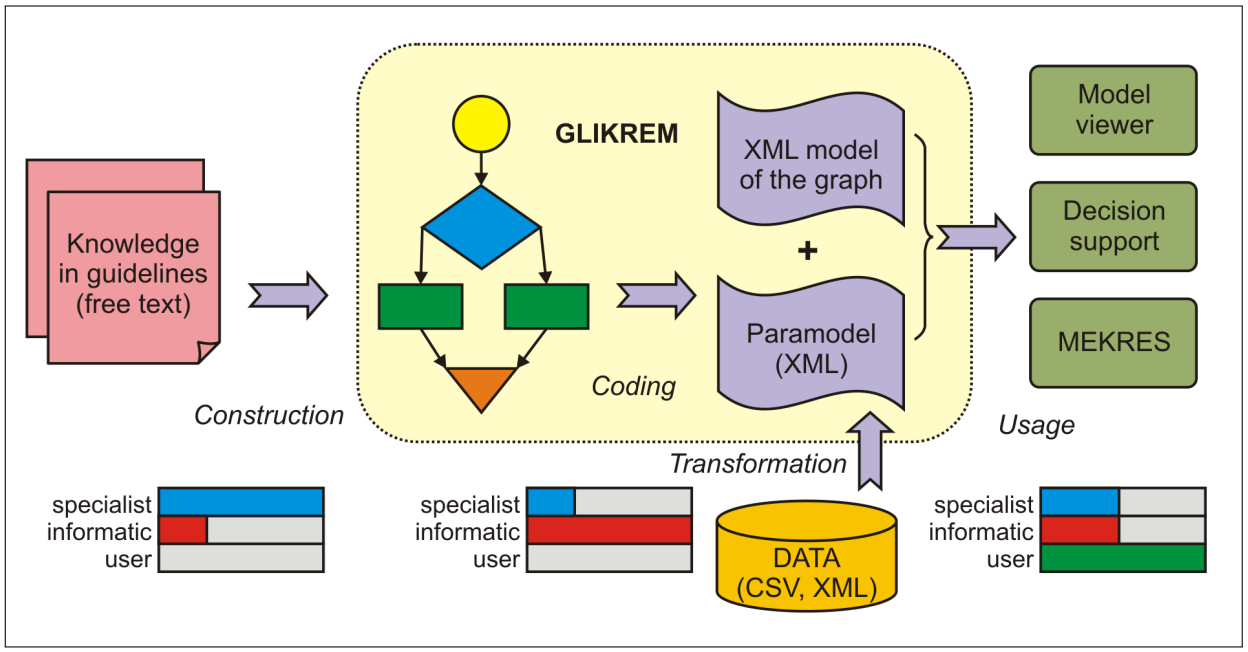

Figure 1: Construction, coding and use of GLIKREM.

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

en2 – en2 Semantic Interoperability in Medicine and Healthcare III
Svačina Š., Zvárová J.

en3 – en8 Health Records as an Object of Czech Personal Data Protection and Intellectual Property Law
Dostál O., Šárek M.

en9 – en18 How to Design an Integration Platform for Interoperable EHR?
Krsička D., Šárek, M.

en19 – en24 Behavioural Biometrics for Multi-factor Authentication in Biomedicine
Schlenker A., Šárek M.

en25 – en30 Stochastic Models for Low Level DNA Mixtures
Slovák D., Zvárová J.

en31 – en38 Mutation Analysis of the COL1A1 Gene in Czech Patients Affected by Osteogenesis Imperfecta
Šormová L., Mazura I., Mařík I.

en39 – en42 Obesity Treatment by Bariatric Surgery and Some of the Pharmacoeconomical Aspects
Telička Z., Svačina Š., Matoulek M.

en43 – en47 Amenability of Czech Medical Reports to Information Extraction
Zvára K., Svátek V.

Semantic Interoperability in Medicine and Healthcare III

Štěpán Svačina, Jana Zvárová

The special issue of the European Journal for Biomedical Informatics publishes selected peer-reviewed papers of students of the doctoral study at the 1st Faculty of Medicine of Charles University in Prague. These papers were also presented as lectures given by Ph.D. students during the third workshop on the topic Semantic interoperability in biomedicine and healthcare held on November 22nd, 2012 in Prague. The first workshop on the same topic was held on November 18th, 2010 in Prague, the second on November 24th, 2011 in Prague.

Semantic interoperability addresses issues of how to best facilitate the coding, transmission and use of meaning across seamless health services, between providers, patients, citizens and authorities, research and training. In essence the semantic interoperability goal is to work towards and support collaboration among human actors and stakeholders, rather than only interoperability among computers. 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, was 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, is to present selected terms from papers of students and to make their description in English and classification by SNOMED CT and ICD10. Then the translations of these findings to the Czech language are also presented at the workshop. Partial semantic interoperability is supported by creation of semantically sound and focused subsets of terms coded in SNOMED CT and ICD 10 that have immediate relevance to Ph.D. theses.

Health Records as an Object of Czech Personal Data Protection and Intellectual Property Law

Otto Dostál[1], **Milan Šárek**[2]

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

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

Abstract

Objectives: The handling of health records is closely tied with in the last years very much discussed topic of personal data protection. It is still possible to encounter fears if the legal regulation of personal data protection allows some of these deployments and in which way. Less often, but still, it is possible to encounter concerns also regarding possible intellectual property claims. In the light of these questions the authors decided to do an analysis of the existing legal framework.

Methods: It this article we analyse the relevant content of Czech Personal Data Protection Act (though as this area is already highly harmonized by EU directives, the demonstrated principles can be applied more generally, not only in the context of the specific country). In similar way we analyse also the Czech Copyright Act.

Results: When comparing both regulations we see that their principles and the subjects they concentrate on are largely different and the personal data protection is more prominent in our context, but the intellectual property regulation can also apply in some cases and complements the regulation. Legal frameworks we discussed here can be judged as developed and relatively mature. This appears to be the result of the harmonisation by EU directives and other supranational legislation.

Conclusions: Legal regulation discussed in this article seems to be generally ready for development and deployment of e-health services. This does not, however, meant, that the described regulation should not be a major concern of health care providers. Quite the opposite. The data Protection Act prescribes critical obligations, such the adoption of measures preventing unauthorised access to personal data. Also for certain types of databases the intellectual property rights cannot be ignored.

Keywords

Health record, database, legal framework, personal data, intellectual property

Correspondence to:

Otto Dostál
First Faculty of Medicine, Charles University in Prague
Address: Kateřinská 32, 121 08 Prague 2, Czech Republic
E–mail: ottodostal@gmail.com

EJBI 2012; 8(5):3–8
recieved: August 16, 2012
accepted: September 26, 2012
published: November 22, 2012

1 Introduction

The keeping of health records is closely tied with in the last years very much discussed topic of personal data protection. It is a critical aspect which is necessary to have in mind in the process of deploying various e-health applications and which still arouses various questions. It is still possible to encounter fears if the legal regulation of personal data protection allows some of these deployments and in which way. Less often, but still, it is possible to encounter such questions also regarding possible intellectual property claims.

The authors are dealing with this topic in the context of the Czech system of law. Nevertheless, as this area is already highly harmonized by EU directives, the demonstrated principles can be applied more generally, not only in the context of the specific country.

The goal of this article is to provide a review of the legal framework in this area and of obligations prescribed by it, to identify possible issues and to help understood its role in the regulation of health records and other medical data.

2 Personal Data Protection

The Act n. 101/2000 Sb. on the Protection of Personal Data (hereinafter referred only as "Personal Data Protection Act" or "Act") [1] is a reflection of the article 10 subarticle, 3 of the Czech Charter of Fundamental Rights and Basic Freedoms [2] according to which *"Everyone has the*

right to be protected from the unauthorized gathering, public revelation, or other misuse of his/her personal data." It also reflects the Council of Europe's Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (also known as the "Convention 108") [3]. And from the point of European law it implements [4, 5] the Directive n. 95/46/ES on the protection of individuals with regard to the processing of personal data and on the free movement of such data [6].

2.1 The Scope of the Personal Data Protection Act

The Personal Data Protection Act is in the field of personal data protection lex generalis, that is, it will be used if there is no special law with different rules (lex specialis). In this article we are going to concentrate on situations where this lex generalis will be used (the Czech regulation lex specialis in the field of health records has already been discussed by the authors in their previous article [7]).

The scope of the Personal Data Protection Act is large; it covers the processing of personal data by both the public authority bodies (the state authorities and territorial self-administration bodies) and by the natural and legal persons. It also applies to personal data processing both by automatic or other means. However, the Personal Data Protection Act does not cover all processing; outside its scope are the cases of personal data processing carried out by a natural person for personal needs exclusively and of accidental personal data collection, if these data are not subject to further processing.

The Act defines the term „personal data". According to its definition *"personal data shall mean any information relating to an identified or identifiable data subject. A data subject shall be considered identified or identifiable if it is possible to identify the data subject directly or indirectly in particular on the basis of a number, code or one or more factors specific to his/her physical, physiological, psychical, economic, cultural or social identity".*

In literature, there is no consensus if the Act does regulate only the personal data of the living people or also the personal data of the deceased [4, 5]. The Act itself does not explicitly states anything about it. The authors though consider logical that the interpretation of the law should be such that the protection granted by the Act should be enjoyed, especially in the field of healthcare, even by the deceased.

2.2 The Status of Health Records

The data about health status of a person the Act considers to be "sensitive personal data". Their processing is nevertheless allowed in healthcare by the § 9 letter c) of the Act, according to which it is possible to process sensitive personal data *"if the processing in question is in relation with ensuring health care, public health protection, health insurance, and the exercise of public administration in the field of health sector pursuant to a special Act, or it is related to assessment of health in other cases provided by a special Act."*

The legal framework in this context does not distinguish the keeping of health records in paper and electronic form. The important thing is, that it must be a processing in relation with ensuring health care or some of the other quoted cases.

From practice the authors are aware of cases when medical facilities require patients to sign up forms where they agree with the electronic way of keeping their health records, while they would not require any signature from these patients for the same processing of the data in case of paper form of the health records. This, however, has no legal basis in Czech law. To the authors it also does not seem to be a meritorious effort to respect the rights of the data subject but instead, in a better case, an unnecessary bureaucracy, or, in worse case, an attempt to dodge responsibility in case of a problem. It is up to the health care providers to guarantee the proper functioning and security of the health records and in case of for example data leak, it cannot exculpate itself by an argument that the patient agreed with the form in which the record will be kept.

2.3 The Controller, the Processor and the Obligations

The subject that determines the purpose and means of personal data processing carries out such processing and is responsible for such processing is called the controller. The controller may through agreement charge another subject to process personal data - the processor.

Such an agreement must be made in writing and shall explicitly stipulate the scope, purpose and period of time for which it is concluded and must contain guarantees by the processor related to technical and organisational securing of the protection of personal data (§ 6 of the Act[1]). The contractor may in this way charge multiple processors. The processors themselves, however, cannot charge another subject with the processing. These are the agreements that are being made between medical facilities and subjects that are providing for them the data storage and other services. Thus although this legal regulation is only generic and brief, it can be stressed out, that it applies to lot of projects in the field of e-health[2].

[1] An implementation of the article 17 subarticle 3 of the Directive n. 95/46/ES.

[2] Such as the e-Health project MeDiMed (http://www.medimed.cz).

For the data subject, that is, the natural person to whom the personal data pertain, the Act does not prescribe any obligations. It does prescribe many of them though to the controllers and processors. We will discuss them here later, though not all of them will be of deeper interest to us because of the existence of the regulation in lex specialis or specific provisions in the Act itself.

Foremost, the Personal Data Protection Act prescribes to the controllers an obligation to specify the purpose for which personal data are to be processed and the means and manner of such processing; these obligations are however in the case of health records already in essence fulfilled by their legal regulation in lex specialis [8]. What the controller always must have in mind is the obligation to process only accurate personal data and, if necessary, to take adequate measures to block the processing and to correct or supplement the personal data. Further it is allowed to collect only personal data corresponding exclusively to the specified purpose and in an extent that is necessary for the fulfilment of such purpose. Therefore it is not possible to collect personal data in health records that are not related to the purpose of these records. The personal data can be processed only in accordance with the purpose for which they were collected. Furthermore it is forbidden to group personal data that were obtained for different purposes. The obligations we described in this paragraph are binding likewise for the processor.

The controller and the processor are also (§ 13 of the Act) *"obliged to adopt measures preventing unauthorised or accidental access to personal data, their alteration, destruction or loss, unauthorised transmission, other unauthorised processing, as well as other misuse of personal data. This obligation shall remain valid after terminating personal data processing."*

The Act does not define in detail which measures these are supposed to be. It is not even possible as the security risks are always developing. It states, however, that the measures must be a result of an assessment of risks from both the persons with immediate access to the personal data, and the persons attempting an unauthorized access, concerning prevention of unauthorized reading, creating, copying, transferring, modifying or deleting of records containing personal data and measures enabling to determine and verify to whom the personal data were transferred. In the area of automatic processing of personal data, the controller or processor is also obliged to

1. *ensure that the systems for automatic processing of personal data are used only by authorized persons,*
2. *ensure that the natural persons authorized to use systems for automatic processing of personal data have access only to the personal data corresponding to their authorization, and this on the basis of specific user authorizations established exclusively for these persons,*
3. *make electronic records enabling to identify and verify when, by whom and for what reason the personal data were recorded or otherwise processed, and*
4. *prevent any unauthorized access to data carriers.*

It is also worth to mention that the Office for Personal Data Protection is asking in its forms about the existence of locks, bars, central security desk, electronic security, security directive, and, in the case of automatic processing, also about access rights, security backups, anti-virus and encryption. In this area, also the technical norms can be of use[3]. The controller or the processor is obliged to document the technical-organisational measures adopted and implemented.

A special obligation of the processor is, if he finds out that the controller breaches the obligations provided by the Act, to notify the controller of this fact without delay and to terminate personal data processing.

The Personal Data Protection Act also prescribes obligations for the employees of the controllers and processors or other natural persons who process personal data on the basis of an agreement concluded with the controller or processor and other persons who, in the scope of fulfilling rights and obligations provided by law, come into contact with personal data at the premises of the controller or processor. These persons are obliged to maintain confidentiality of personal data and security measures whose publishing would endanger the security of personal data. This obligation is binding for them even after the termination of their employment or the relevant work. The obligation to maintain confidentiality, however, does not apply in cases where some other act would prescribe information obligation (such as the obligation to report crime).

3 Intellectual Property Rights

3.1 Author's Work

The authors consider clear that the individual health records created by doctors cannot be considered author's work in the sense of § 2 article 1 of the Act n. 121/2000 Sb. on Copyright and Rights Related to Copyright (hereinafter referred only as "Copyright Act" or "Act") [9], as they do not fulfil the necessary criterion of uniqueness. They do not represent an original exceptional outcome of the creative activity of the author.

We consider it necessary though to discuss more the copyright to databases. It is because the Copyright Act also states that the quoted criterion of uniqueness does

[3]Such as ISO/IEC 17799:2000 Information technology – Code of Practice of Information Security Management or Czech ČSN ISO/IEC TR 13335 Informační technologie – Směrnice pro řízení bezpečnosti IT 1 - 3

not apply to computer programs and databases (§ 2 article 2). For them it is sufficient if they are original in the sense that they are by the way of the selection or arrangement of their content the author's own intellectual creation. For databases it is also required that their individual parts are arranged in a systematic or methodical way and are individually accessible.

Database by the definition in the Act is a collection of independent works, data, or other items arranged in a systematic or methodical manner and individually accessible by electronic or other means, irrespective of the form of the expression thereof (§ 88 of the Act). The collections of health records can be in the light of this definition considered databases in the context of the Act. Before we discuss the relevance of this fact it is necessary though to mention another legal regulation in the same Act which is the regulation of the right of a database maker to his database.

3.2 The Right of a Database Maker to His Database

The Copyright Act in its § 88 and following regulates the right of a database maker to his database. This is an implementation of the Directive n. 96/9/EC on the legal protection of databases [10]. It is a type of protection sui generis which is in its nature closer to the protection against unfair competition than copyright protection [11]. The protection of the right of a database maker to his database is not a protection of a right of the author to his work, nor of a right related to copyright, but a special protection regulated in the Copyright Act existing outside these categories.

The maker of the database is the natural or legal person who, on his own responsibility, has compiled the database, or on whose impulse is the database compiled by another person (§ 89 of the Act). The maker of the database may transfer his right.

The right of a database maker to his database arises only when there is a contribution in the form of formation, verification or presentation of the content of the database, which is substantial in terms of quality or quantity.

The protection covers databases in any form, that is both electronic and non-electronic. Protected is the content of the database and also the elements necessary for the operation and searching in databases such as thesaurus and indexing system. On the other hand this protection does not include computer programs used for creating and running the database [11].

The content of the right of a database maker to his database is the right to extraction or re-utilisation of the content of the database and the right to grant to another person the authorisation to execute such a right. Extraction means a transfer of the database (all or a substantial part thereof) to another medium, re-utilisation means making it available to the public. Lending of the original or a copy of a database is not considered extraction or re-utilization.

The Copyright Act also states that the right in question is not infringed by the lawful user who extracts or re-utilises:

1. qualitatively or quantitatively insubstantial segments of a database that has been made available to the public as long as he is doing so in a normal and appropriate manner, not systematically or repeatedly, and without damaging the legitimate interests of the maker of the database,

2. a substantial part of the content of the database but only

   (a) for his personal use in case of non-electronic database, or

   (b) for scientific or educational purposes, if he indicates the source, or

   (c) for the purposes of public security or an administrative or judicial procedure.

The right of a database maker to his database runs for 15 years from the making of the database. If, however, the database is made available during that period, the right of the maker of the database expires 15 years from the date when the database is made available (§ 93 of the Act).

In case of a violation of the right of a database maker to his database the civil law proceeding as well as the public law sanctions (§105a article 1 letter a) of the Copyright Act, § 270 of the Penal Code [n. 40/2009 Sb.] [12]) can be used.

3.3 Usability of Intellectual Property Law Regulation

Such is the protection of databases in the Copyright Act. The question remains, to what extent is it possible to use the copyright to the database and the right of a database maker to his database for protection of medical databases.

We believe that the principles of these rights are not very much in line with the needs of legal protection of medical databases containing health records of individual patients. The character of such data and the requirements for their keeping are completely different from those which are typically in the scope of the intellectual property law. The very roots of databases with health records differ from others by strong public law elements compared to private law elements of the other databases.

Opposite is also the logic of the compared acts. While the author's works are typically distributed commercially, the exchange of information contained in health records should be burdened by financial questions as little as possible. While the author's works can be in certain cases accessed freely (see free uses in § 30 of the Act), in the latter this is out of the question. Also the length of the legal protection set in the Copyright Act (15 years for the right of a database maker to his database) is not corresponding. And it can be stated that the legal regulation of health records, medical confidentiality and personal data protection is so complex that the protection by means of intellectual property law would be even superfluous.

From there reasons we believe that the copyright to databases in case of medical databases with health records of patients does not exist as they must be considered official works within the meaning of § 3 letter a) of the Act. This provision defines an exemption according to which the copyright protection does not apply to official works.

In the case of a right of a database maker to his database it was before possible to get to the same conclusions in exactly the same way. However, since May 22, 2006 by the changes introduced by an amendment n. 216/2006 the § 3 letter a) of the Act does not apply to the right of a database maker to his database anymore (except databases which are part of statutes, which is not our case) [13]. Therefor it can be argued that this right exists even for databases with health records. The existence of such right in our opinion though has little practical impact, as the rules for who can and who cannot access the data in such database are strictly set in the regulation lex specialis [8]. This right thus seems to have a bare character.

What we said above does not necessarily mean that the described intellectual property rights are irrelevant in the field of healthcare. It medicine, there are other databases than those with personal data of patients. For example if the personal data from health records get anonymised (by which they are stopping to be personal in the sense of the Personal Data Protection Act) and transformed into a database designed for educational purposes, both the copyright and the right of a database maker to his database could apply. These rights thus can be used for protection of various databases with medical knowledge stored for educational and scientific purposes.

4 Conclusion

Above we dealt with medical data in the light of legislation for protection of personal data and for protection of intellectual property. As we can see from the analysis, because of the character of health data the personal data regulation appears to be more important, however, the protection of the intellectual property rights also has its place and both somewhat complement each other. The health databases with personal data of patients are regulated by legislation for personal data protection. Then, in case of their anonymization for usage for educational purposes, these databases fall into the scope of intellectual property legislation.

Unlike the regulation lex specialis analysed in previous article of the authors [7] both legal frameworks we discussed here can be judged as developed and relatively mature. This appears to be the result of the harmonisation by EU directives[4] and other supranational legislation. The part of legal regulation discussed in this article seems to be generally ready for development and deployment of e-health services.

By the previous paragraph the authors did not want to say though, that the described regulation should not be a major concern of health care providers. Quite the opposite. The Data Protection Act prescribes critical obligations, such the adoption of measures preventing unauthorised access to personal data. Also for certain types of databases the intellectual property rights cannot be ignored.

The discussed legal framework represents obligations and certain limitations but these are necessary for building trust in the environment. Lack of trust makes people hesitate to adopt new services. This risks slowing down the development of innovative uses of new technologies. We should thus think about legal framework as an important part of a foundation of every e-health project.

Acknowledgements

The paper has been supported by the SVV-2012-264 513 project of Charles University in Prague.

References

[1] Act n. 101/2000 Sb. on the Protection of Personal Data and on Amendment to Some Acts, as amended by 227/2000 Sb., 177/2001 Sb., 450/2001 Sb., 107/2002 Sb., 310/2002 Sb., 517/2002 Sb., 439/2004 Sb., 480/2004 Sb., 626/2004 Sb., 413/2005 Sb., 444/2005 Sb., 342/2006 Sb., 109/2006 Sb., 170/2007 Sb., 52/2009 Sb., 41/2009 Sb., 227/2009 Sb., 281/2009 Sb., 468/2011 Sb., 375/2011 Sb. In Czech.

[2] Charter of Fundamental Rights and Basic Freedoms n. 2/1993 Sb. In Czech.

[3] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. Available from: http://conventions.coe.int/Treaty/en/Treaties/Html/108.htm

[4] Mates P. Ochrana soukromí ve správním právu. Praha: Linde, 2004. In Czech.

[4] The Directive n. 95/46/ES might get replaced with a regulation in the future though [14].

[5] Matoušová M, Hejlík L: Osobní údaje a jejich ochrana; Praha, ASPI Publishing, 2003. In Czech.

[6] Directive n. 95/46/ES on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995 P. 31–50.

[7] Dostál O, Šárek M. Support for Electronic Health Records in Czech Law. EJBI. 2012 Jun 15;8(2):29-33.

[8] Healthcare services and conditions for their providing Act n. 372/2011 Sb. § 52-69. In Czech.

[9] Act n. 121/2000 Sb. on Copyright and Rights Related to Copyright and on Amendment to Certain Acts, as amended by 81/2005 Sb., 61/2006 Sb., 216/2006 Sb., 186/2006 Sb., 168/2008 Sb., 41/2009 Sb., 227/2009 Sb., 153/2010 Sb., 424/2010 Sb., 420/2011 Sb., 375/2011 Sb. In Czech.

[10] Directive n. 96/9/EC on the legal protection of databases. Official Journal L77, 1996/03/27, pp. 20–28.

[11] Kříž J, Holčová I, Korda J: Autorský zákon a předpisy související – komentář. Praha: Linde, 2001. In Czech.

[12] Penal Code n. 40/2009 Sb. as amended by n. 306/2009 Sb., 181/2011 Sb., 330/2011 Sb., 357/2011 Sb., 375/2011 Sb., 420/2011 Sb. In Czech.

[13] Act n. 216/2006 Sb. amending Act. n. 121/2000 Sb. on Copyright and Rights Related to Copyright and on Amendment to Certain Acts as subsequently amended, and some other Acts. In Czech.

[14] Proposal for a Regulation on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). COM/2012/011 final - 2012/0011 (COD).

How to Design an Integration Platform for Interoperable EHR?

Daniel Krsička[1], **Milan Šárek**[2]

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

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

Abstract

Background: Integration platform is a basic technical tool realizing an interoperable Electronic Health Record (EHR).
Objectives: Our goal is an analysis of the integration platform functional structure and its relations to defined interoperability levels.
Methods: The existence possibility of a simple dependency between EHR use cases and integration platform technical functions will be tested on the models.
Results: The experiments will result into a proof of existence of this dependency and into a possibility to work with it.
Conclusions: The results will be discussed according to opportunity to generalize this method, to use it practically and develop further research in this domain.

Keywords

Interoperability, electronic health record, healthcare information system, integration platform, integration pattern

Correspondence to:

Daniel Krsička
First Faculty of Medicine, Charles University in Prague
Address: Kateřinská 32, 121 08 Prague 2, CR
E–mail: dkrsicka@gmail.com

EJBI 2012; 8(5):9–18
recieved: August 15, 2012
accepted: October 15, 2012
published: November 22, 2012

1 Introduction

Massive penetration of the Healthcare Information Systems (HIS) and eHealth resources in general potentiate the significance of Electronic Health Record (EHR) interoperability as an ability of two or more subjects to achieve a common goal or mutually support each other to achieve the individual goals respectively (synergic effect). To describe this effect better, we can use the Metcalf's Law, postulated originally for telecommunication networks and Ethernet. This law introduces a network value quantity described as the number of all possible connections among subscribers (HIS in our case). So value of the whole interoperable EHR system should be dependent on the number of systems (HISs) integrated and asymptotically approximated by the quadratic polynomial of n^2.

Nevertheless it is becoming apparent [6] that the value of integrated HISs as a whole is not growing quadratic and that the Metcalf's Law is not applicable as a sufficient model. The reason is simple - Metcalf's law omits these parts of reality, essential for EHR interoperability analysis, primarily facts regarding EHR messages content and its usage in work (business) processes. The HIS integration is not a mere communication interconnection, so it is not about connection establishment only. It is necessary to pinpoint and follow many protocols enabling an information interchange for particular HIS components and layers. That implies the definition of interoperability level.

Table 1: Interoperability level definitions comparison.

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

We can use the existing definition after Gibbons [14] postulated in scope of HL7 EHR Interoperability Workgroup, defining 3 levels, or we can use the definition after Bloebel [1] setting up 5 levels of interoperability. Researching aforementioned resources we have defined relations among these 2 definitions, figured out in Table 1. For our purposes we will use the interoperability levels definition after Bloebel onwards, who demonstrates insufficiency of the traditional interoperability perception in technological degree and emphasizes the higher interoperability levels including the semantic. The classification

after Gibbons is not suitable for our work due to our focus on logical integration platform design which is in Gibbon's definition plainly abstracted into just one technical interoperability level.

Our motivation is based on lessons learned about the technological interoperability insufficiency as a means of massive dissemination of interoperable EHR including all needed attributes defined e.g. in ISO/EN:13606 [10]. This statement is supported by the professional publications focusing mainly on EHR system content and semantics. Oneself, we have published the technological interoperability view inadequacy in [15] and [16]. We have demonstrated that the higher interoperability levels cannot be assured by and based on accepted and broadly used classification into technical layers according to ISO/OSI model in ISO/IEC:7498 [17]. The process and partly the semantic interoperability has not any technical equivalent in ISO/OSI model, so these interoperability levels cannot be procured by technical resources only.

The present professional publications aiming interoperability are concentrating primarily on issues of EHR standardization, its structuring, content and usage by the end users including the semantic interoperability support in the form of data standard definitions, common vocabularies and ontologies. An EHR functional model is published in ISO/HL7:10781 [11] defining a basic set of EHR use cases. Unfortunately, the functional view research combining the EHR requirements with the technical realization of EHR integration platform in considerably underestimated in the professional society. Basic architectures of some national EHR projects, systems or efforts can be found. There are some groups like HSSP [8] engaged in EHR integration platform definitions, nevertheless a generally usable, comprehensive, logical design of the integration platform internal mechanisms as a functional composite of more than one HIS, aggregating the substantial functions centrally is not published yet.

The HIS semantic interoperability can be significantly supported by usage of EHR standards like HL7 [9] or DASTA [13] in the Czech environment. Each standard proceeds from its basic metamodel serving for deriving all the other parts of the standard. This metamodel also restricts the area and intent of standard application. This can be demonstrated on comparison between the standards HL7v3 and DASTAv3. HL7 is based on its Reference Information Model (RIM) establishing a basic "skeleton" for all the HL7 models as a relation among the subject, role, activity and object. Using this paradigm, all the relations of this type can be sufficiently described by the HL7v3 in the same way. On the other hand, the Czech national standard DASTAv3 bases its structure on information descriptive view only. It does not cover the interaction among various EHR roles, so it is good usable for data description, but unusable for the semantic expressions or managing work (business) processes. This difference has been well described and practically demonstrated by examples in [5].

As described further, to reach the highest interoperability level is not necessary and should not be an automatic goal for each HIS, because not each interoperable EHR system has to implement all the interoperability functions defined. The driving factor form the specific HIS use cases resulting in requirements on interoperability of particular level.

Our goal is to point to the importance of functional approach to the EHR communication, to verify possibilities of present semantic interoperability knowledge utilization for an integration platform design methods simplifying and formalization.

1.1 Integration Platform

The integration platform is a basic technical means for integration of information systems, the HISs inclusive. It consists of hardware and software components and also data models, structures processing rules and security mechanisms. For our purposes we will focus on the software part of integration platform logical structure only. It is composed of a logical functions basic set, cooperatively realizing the EHR messages transport and processing. Further we will define the integration platform using these functions and their aggregations in relations to the particular interoperability levels.

1.2 Integration Pattern

Integration patterns are partial functional concepts, from whose realization the integration platform consists of. Each integration pattern [7] is a generalization of a verified method (best practice) in the area of information system integration. It is a special case of design pattern [18], typically defined by an unique syntactical graphical model and informal semantic description describing the case of pattern usage. For our work we have used probably the most comprehensive set of integration patterns published by Hohpe and Gregor [7]. The alphabetical order is depicted in the Figure 1.

Each integration pattern solves a particular typical situation in data communication and processing through the integration platform. Particular EHR use cases mark off each other and each use case or even each EHR message transported among HISs can be processed in a different way, so different integration platform components can be used, thus different integration patterns and their combinations apply.

Our goal is to structure the mentioned ambiguity and define rules applicable in early EHR implementation project phases and simplifying the logical design significantly. This logical design has to be pure platform (technological) independent. The facilitation lies in the target interoperability level definition belonging to the specific organization (defined by its EHR use cases) and the proposal of basic integration platform logical structure (expressed in the sets of integration patterns to implement).

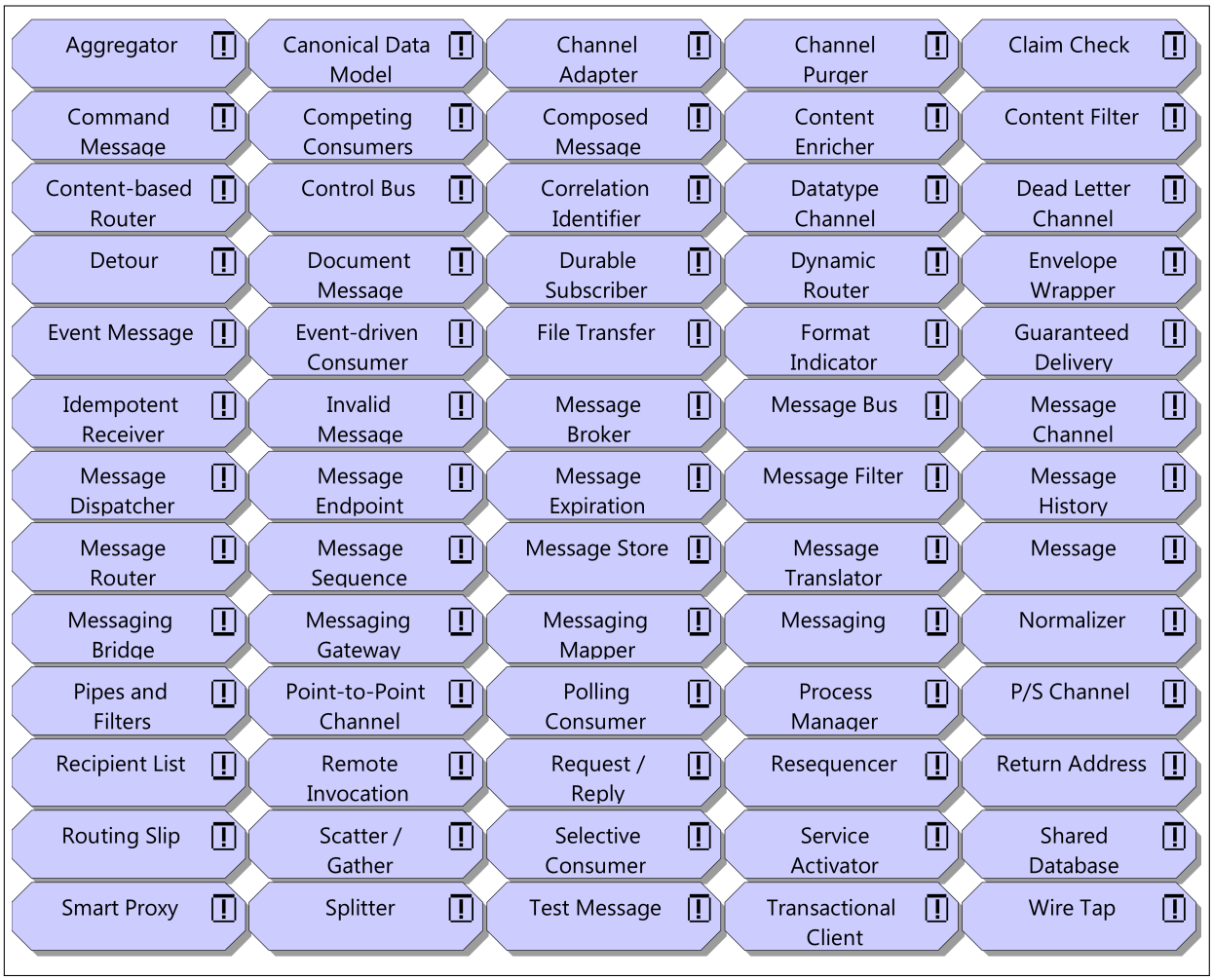

Figure 1: Alphabetical list of the integration patterns - the atomic EHR integration platform funcionalities.

Of course it will be a generic basis only which should be analyzed and customized in deeper detail during the EHR implementation project. Anyway a data analysis should be established on the advanced standards like HL7 and the project should follow a wide accepted methods like Rational Unified Process (RUP) [20].

We expect that some integration patters are already whole or partially included in existing standards like IHE [19] or that existing standards are tight coupled to them. More information can be found in the section Discussion of this article.

1.3 EHR Use Cases

The use case forms the usage specification of particular HIS function by an external role (outside the system) like user, other system etc. The typical EHR use cases can be found in [11], [6] or in [19]. The test cases of EHR use cases can be found in section Experiments of this article.

2 Goals and Hypotheses

We focus on the functional view analysis of EHR integration platform as a technical means of interoperable EHR realization. EHR integration between 2 HISs is supported by an integration platform. Its structure and behaviour has to include all the functions necessary for reaching the target EHR interoperability level. Therefore we need to find dependencies among EHR requirements, interoperability level requirements and structure of the integration

We would like to elaborate a formal method supporting the EHR use case analysis which would simplify and speed up an integration platform design. This way an interoperable EHR implementation would be supported. Aforementioned method benefits lie in analysis and design acceleration, implementation shortening, support of early prototype creation and anticipated decreasing the number of change request, so in reduction of total solution costs.

Hypothesis

Let us suppose that there is a mapping, assigning for each EHR use case a set of integration patterns. These patterns ensure the EHR integration platform functionalities required for the use case realization and will correspond to the necessary interoperability level. We propose that by a sequential aggregation of these mappings it will be possible to prepare a basic functional structure for the whole EHR integration platform required for particular set of EHR use cases (business requirements).

Let us structure the functionality sets of EHR integration platform according to the interoperability level needed and try to find a mapping from the set of use cases to this structure. It should result into a definition of assignment from set of EHR use cases into a necessary interoperability level.

The benefit is a software analysis simplification and EHR integration platform design optimization.

3 Methods

Interoperability Related Classification of Integration Patterns

To enable an assignment of each integration pattern to the typical HIS interoperability level, it has been inevitable to establish a hierarchical model of integration patterns. This hierarchy follows the interoperability levels and also the typical structure of integration platform, i.e. transport and processing parts, but without generally accessible business services, which can be established with data semantic usage only. This structure introduces a basic technical means for EHR integration among systems (HISs interoperable integration). Descriptions of individual integration platform layers follow:

- Access Layer: forms a place, where all the integrated systems connect to, to establish a suitable communication. It contains algorithms and structures enabling technical resources compatibility. From the ISO/OSI perspective it is a solution on layers 1 to 5.
- Transport Layer: ensures a basic user data transmission up to the ISO/OSI layer 6. Data is encapsulated into messages. During the analysis, it is necessary to define the technical metadata determining communication endpoints and data structures. Transport layer takes care about all the transmission mechanisms including failover, high-availability, reliability or idempotence.
- Transformation and Routing Layer: manipulates with data transmitted within the meaning of format and structure change on the ISO/OSI layer 7. There is a necessary condition of existence and compliance with common registers, vocabularies and rules. The layer also routes the messages, their parts or aggregations to the right recipients.
- Semantic Layer: Works with the meaning of transmitted information. Components of this layer has to be able to ensure communication among mutually heterogeneous business (or information) domains within the meaning of Generic Component Model [3]. The semantic layer algorithms focus on the data meaning, nor on the data structure or information syntax. In contrast to the well known accord [9], we suppose that this layer has not an equivalent layer in the ISO/OSI model, because this does not solve the data transport, but presentation and sense only.
- Business Processes Layer: concentrates on processes executed by given roles. These processes can have a known structure or be dynamic and to progress according to the actual system state, environmental (contextual) information and to the data processed by the layer. It also includes processes solving a feedback-based process / system optimization. It has not an equivalent in ISO/OSI model.

Above mentioned integration platform structure enables an assignment of corresponding interoperability levels after Bloebel et alli [1][2][4]. Here we are looking for and testing a relation (dependency) between EHR / HIS interoperability levels and integration platform layers.

By a combination of interoperability levels and semantics of particular integration patterns, we obtained an integration patterns set structuring into the 5 subsets. For more information see the Figure 2. We propose that the EHR use cases majority will be resolvable by some of these patterns of particular subsets. But to do this, we have to evaluate the EHR use cases and assign a necessary interoperability level to each use case according to specific method. Thus we need a classification or evaluation system for the EHR use cases. This system is suggested in the next chapter.

EHR Use Cases Classification

To prepare an EHR use case structuring it is appropriate to define them the classification criterions with following features:

- applicable universally to any EHR use case,
- with trivial semantics excluding misunderstanding and facilitating the utilization,
- moderate number of possible values.

Inspired by the HL7v3 RIM [9] and the law of 5W [21] we have proposed following classification criterions for the EHR use cases:

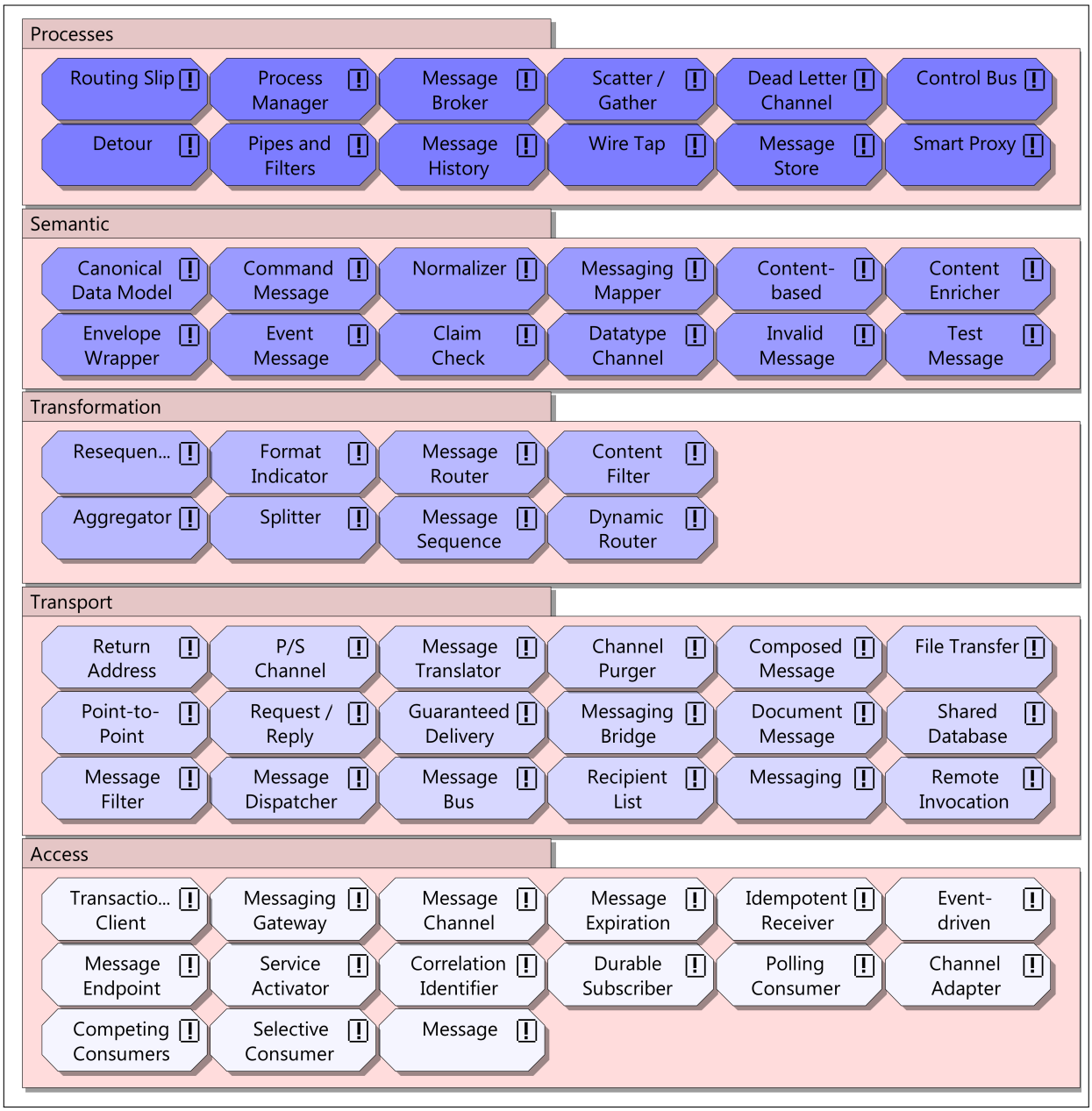

Figure 2: Integration patterns divided into the groups each supporting a particular level of interoperability.

- Space - reflecting the perspective given by questions: "Where the information communication takes place? How distant the points of presence are?"
- Time - reflecting the perspective given by questions: "When the communication takes place? How fast and often it runs?"
- Subject - reflecting the perspective given by questions: "Who is communicating? What is his skills?
- Object - reflecting the perspective given by questions: "What is communicated? Why the communication runs?"

For our experimental purposes we draft weighted values of these classification criterions:

The Dimension of Space

Considering the interoperability perspective, the physical distance of communicating roles is not so important in comparison with the logical distance emerging from the mutual conversance of communicating roles. It can be distinguished in 2 groups. The first one forms persons, i. e. there is difference between information sharing e. g. the physician and nurse in one hospital department or whether communicate a GP with a specialized detached laboratory. Due to we are modelling with point to optimize logical design of technological components, we omit the cultural and social specifics. The second group incorporates the organizations and we can scale, as in the first group, from private praxis, particular hospital de-

partments, clinics, hospitals to insurance companies and national healthcare-related institutions.

For EHR use case ration we will apply 1 from 3 following values possible and the corresponding score.

- Communication in a work team (0 points) - The communicating know each other in person. Communication runs in real time and brings a lower formalization level.
- Communication in an organization (1 point) - The particular communicating are motivated by the same goals and common working methods in outline.
- Communication between organizations (2 points) - Strictly formal communication way with necessity to establish a contract for all the services provided or consumed between organizations.

The Dimension of Time

The time dimension impacts the EHR integration mostly in the requirement specification (business processes or use cases) and in the technical realization. On the other hand, the application of data standards is less affected. The necessary interoperability level is not influenced by the time dimension directly, but it is a suitable additional information to the use case specification and it will be used for the analysis and particular implementation design. It is important to see that it characterizes the data access frequency and so amount of the formalization required (data not red or changed become obsolete and unreadable). We propose the following weights and score.

- Real time communication (0 points) - Information interchanged immediately after creation and often also immediately utilized. Typical examples are daily records, statim indications etc.
- Daily communication (1 point) - Information interchange once or more times a day, Mostly it is regarding to the primary (business) processes like a care provision.
- Monthly communication (2 points) - Communication of often aggregated data. The indication arises from lower use case criticality or from necessity to process data in the batch transactional way (e.g. reporting for payments or perhaps data mining for statistical studies with need to lock a large data set for a while to ensure consistency).

The Dimension of Subject

For our experiment a small set of role is enough. For the comprehensive set definition a concept from ISO/TS:22600 [16] can be used. For consideration of necessary interoperability level it is much more important to evaluate the differences among communicating roles regarding specialization and education of communications. For definition of the subject dimension meaning we use the Generic Component Model [3], its Domain Perspective dimension respectively.

- Roles with the same knowledge (0 points) - Roles in the communication have approximately the same education and specialization. They work in the same or similar processes, activities and their aspects. They understand the same terminology and paradigms. E.g. physicians in the same department
- Roles with a similar knowledge (1 point) - Communicating roles works in the same discipline (domain), but they do not have the same education and knowledge. In this domain they perform different activities. They understand a certain common language and terminology, but each of them maintain its own specializations. Examples can be physician and nurse, physicians of different specializations, scientist in primary research and clinical doctor etc.
- Roles with completely different knowledge (2 points) - The roles have completely different education and knowledge. They a priori do not understand the opposite role principles and means of expression. Confronted with a particular problem or question they focus on different aspects and apply different approaches to the solution. Typical comparison can be physician and patient, administrative worker and manager, ...

The Dimension of Object

At first sight, the communication object classification is quite complex due to its diversity and set cardinality. Nevertheless with regards to the classification model intent an analysis of particular attributes is enough and so we do not need to know the complete messages content. Our goal is to design a logical structure of technical resources (components), not their content like rules, algorithms, registers or vocabularies. So we focus on syntax and semantics expression in the transferred messages. With regard to possible interpretation after [3] we define the following criterion values:

- Usage of syntax (0 points) - The information shared is written in a formalized way. Data is readable by machines in platform independent way, the data structures are defined with use of EDI, XSD, ... and also shared registers.
- Usage of semantics (1 point) - Includes the Syntactic group attributes and also use metadata defining the meaning and sense (for the end user or for processing engines) of transmitted information. This enables a sharing among different roles thanks the information unambiguity.

- Usage for deterministic action (2 points) - The transmitted information is structurally and semantically deterministic enough to execute and automatic processing in HIS or to propose a working method /process for a role a priori unskilled in the domain / profession. For example the advanced systems for decision support or automatic business process management such as optimization and planning processes.

In this article we disregard other partial classification, namely the questions of technological data records and their structuring. These attributes influences the data modelling which is out of scope of this article.

EHR Use Case Classification and the Interoperability Levels

The basic classification challenge in the proposed method is a derivation of target interoperability level from the values of aforementioned classification criterions. Each EHR use case can get from 0 to 8 points in total (4 criterions, 0 - 2 points in each criterion). After more detailed consideration we conclude that the summation is not the primary but much more important is the combination of criterion values. For assessment we specify rules in Table 2.

Table 2: Classification criterion values evaluation.

1 earned	*2 earned*	*Target Interoperability Level*
-	≥ 2	Process
-	≥ 1	Semantic
≥ 1	-	Syntactic
0	0	Structural, Technical

Aggregation Results in EHR Integration Platform Design

In the following experiments we are going to classify each model EHR use cases according to the criterions. We will get a set of pairs [use case; interoperability level]. Based on the highest interoperability level required in this set and with regard to the distribution of their relative frequencies we suppose to design an initial EHR integration platform layers. These layers are defined by sets of integration patterns as the basic functionalities of each layer. Analysis in a specific implementation project should focus just on these layers. From the relative frequencies distribution we can expect the majority of analytical work in the project. Let us show on 2 small examples

Model Situation Nr. 1:

A small purpose-built application for one clinical department, 25 use cases in total. Distribution of interoperability levels required is in Table 3.

Table 3: Interoperability level required by use cases in example Nr. 1.

Interoperability level	*Number of cases*
Technical	24
Structural	20
Syntactic	18
Semantic	2
Process	0

Conclusion: The initial integration platform design has to be focuses on technological compatibility, transport protocols and messages format standardization.

Model Situation Nr. 2:

2 HISs integration between 2 independent hospitals, 250 use cases in total. Distribution of interoperability levels required is in Table 4.

Table 4: Interoperability level required by use cases in example Nr. 2.

Interoperability level	*Number of cases*
Technical	250
Structural	230
Syntactic	200
Semantic	180
Process	40

Conclusion: The initial integration platform design has to encompass the support of access, transport, transformation and routing of data based on technical and also user metadata. Processes (workflow) are defined within the services between hospitals and a request for orchestration emerges. This can be realized by specializes process interoperability integration patterns and components (broadly by an orchestration engine).

4 Experiments - Model EHR Use Cases and Interoperability

We have applied the aforementioned method on 6 following model EHR use cases. Each use case has been defined by its initial (business) description. Usually the description is supplemented during the analysis phase with the customer (e.g. physicians). In our experiments we have used our own information and knowledge for the simulation.

The overall use case semantics has been evaluated under given classification criterions and we obtained the required combinations of weighted values. Based on these combinations we set the required interoperability level for each EHR use case.

Aggregating all the experiments together we gained the relative distribution of interoperability level frequencies as a basis for an initial EHR integration platform design.

4.1 Experiment Nr. 1

Use case description: Management of daily records in one clinical department.

Analysis: The roles work in a compact team, the co-worker know each other and all belongs to one professional domain.

Classification: can be found in Table 5.

Table 5: Use cases evaluation in experiment Nr. 1.

Criterion	*Valuee / Score*
Space	in team / 0
Time	real time / 0
Subject	similar / 1
Object	syntactic / 0

Conclusion: Interoperability level required for use case Nr. 1 is: **Syntactic**.

4.2 Experiment Nr. 2

Use case description: Access to the patients radiological data for other physicians.

Analysis: The co-workers do not need to know each other and their specialization can (and probably will) differ, even if we suppose a quite good knowledge and experience with reading the results from visualization methods (here RTG).

Classification: can be found in Table 6.

Table 6: Use cases evaluation in experiment Nr. 2.

Criterion	*Value / Score*
Space	in organization / 1
Time	real time / 0
Subject	similar / 1
Object	semantic / 1

Conclusion: Interoperability level required for use case Nr. 2 is: **Syntactic**.

4.3 Experiment Nr. 3

Use case description: Patient's laboratory test results access for a GP, processed by an external testing laboratory.

Analysis: Cooperating roles do not know each other. There is no need for real time communication. The specialization and knowledge can differ but the most common tests have to be able to read all the physicians. We do not consider the special laboratory tests (like CVS, cancer marks, detailed haematology or immunology) which are not commonly indicated by GPs. The functionality can be offered as a service so the contract definition is necessary (SLA - Service Level Agreement).

Classification: can be found in Table 7.

Table 7: Use cases evaluation in experiment Nr. 3.

Criterion	*Value / Score*
Space	between orgs. / 2
Time	daily / 1
Subject	similar / 1
Object	semantics / 1

Conclusion: Interoperability level required for use case Nr. 3 is: **Semantic**.

4.4 Experiment Nr. 4

Use case description: Access to the anonymized patient data in an university hospital from an university research centre for the purpose of a statistical longitudinal study.

Analysis: It is necessary to define not only content and semantics of the data but also the way and purpose of its processing. We have to respect the regulatory law and also must not omit some information relevant for the study (false positive/negative results risk).

Classification: can be found in Table 8.

Table 8: Use cases evaluation in experiment Nr. 4.

Criterion	*Value / Score*
Space	in organization / 1
Time	monthly / 2
Subject	similar / 1
Object	deterministic action / 2

Conclusion: Interoperability level required for use case Nr. 4 is: **Process**.

4.5 Experiment Nr. 5

Use case description: Reporting of provided healthcare from the provider to the payer.

Analysis: A periodical rigid communication in the form of a service provided and consumed among organizations (more service consumers / healthcare providers). The contract (SLA) definition is absolutely inevitable.

Classification: can be found in Table 9.

Table 9: Use cases evaluation in experiment Nr. 5.

Criterion	*Value / Score*
Space	between organizations / 2
Time	monthly / 2
Subject	different / 2
Object	semantic / 1

Conclusion: Interoperability level required for use case Nr. 5 is: **Process**.

4.6 Experiment Nr. 6

Use case description: On-line access for the patient to his/her EHR.

Analysis: Ad hoc access which realization request emerges from the valid Czech law. The patient (user) stays out of the organization, its motivation, knowledge and experience is completely different in comparison with healthcare professionals. The accessible EHR must include also additional information enabling the patient's understanding.

Classification: can be found in Table 10.

Table 10: Use cases evaluation in experiment Nr. 6.

Criterion	*Value / Score*
Space	between organizations / 2
Time	real time / 0
Subject	different / 2
Object	semantic / 1

Conclusion: Interoperability level required for use case Nr. 6 is: **Semantics**.

5 Results

Based on the knowledge about particular interoperability levels and with use of classification rules mentioned above we have evaluated a required interoperability level in each model EHR use case. Thus we have demonstrated that a mapping required in our hypothesis really exists and that for the level definition we can use quite simple classification criterions, understandable also for persons not skilled in computer science. We have demonstrated that required mapping can be found for aforementioned the EHR use cases, because of classification according to generic criterions.

It is clear form experiment's results Nr. 1 - 6 how the model integration platform design looks like. It is determined by the highest interoperability level found in the use cases and by the relative distributions of levels found. Let us summarize this data in Table 11.

Table 11: Aggregation of experiments results.

Required Interoperability Level	*Total incidences*
Process	2
Semantic	4
Syntactic	6
Structural	6
Technical	6

Looking on the table it is evident that this model situation has to base the initial integration platform design on common access, transport and transformation & routing layer as an inevitable basis. Also an essential functional support for semantic interoperability is necessary. The dedicated process engine realizing the integration patterns from the highest level should be considered, because its commonly a little bit expensive, so it could not be justified just for 2 use cases. But in the real project, if the Process interoperability forms more than 30% of total requirements, a standalone orchestration engine is absolutely needed.

6 Discussion

The classification rules for EHR use cases mentioned in this article can be apparently applied on any EHR use case and so it should be possible to evaluate any of them. The understanding of these rules is quite simple so the use cases can be evaluated also by a person without a specialized training in computer science and software engineering (physician, manager ...). This way a mapping between different GCM domains [3] is enabled in the integration platform development process. The definition of target interoperability results from the method stated in this article.

The method implication lies in the possibility to structured view to the often heterogeneous set of (business) requirements. For optimal method set up it is necessary to execute more experiments and tests on model and also real EHR use cases. It has to be tested whether the method can really simplify the analysis project phase and enable the development of an early integration platform prototype. The benefit of early prototype is the possibility to test soon after the requirement specification, to decrease the number of change requests, to speed up the project and to lower the costs in total.

According to our present research, it seems that some of presented integration patterns forming the range of values of our mapping already exist or are partly included in existing standards like the IHE profiles [19]. These standards define the specific EHR use cases with some realization specifications inclusive. In the further research it will be appropriate to focus also on relations among these standards and logical functionality view represented by the integration patterns and their classification mentioned here.

7 Conclusion

With regard to the cost cutting need and the EHR implementation projects acceleration we have defined a supporting method for the EHR use case analysis. By application of this method we have obtained an information set for a logical, platform independent design of an EHR integration platform. The testing on model situations was successful and we are motivated for the further experiments including the real use cases in the healthcare provider environment. We expect that these tests together with further method advancement will be executed in the environment of Krajska zdravotni, the major healthcare provider in district of Ustecky kraj, incorporating 5 hos-

pitals and cooperating on science and research. A part of this research should be also a comprehensive analysis of relations among various integration patterns and existing IHE profiles.

Building up the dedicated integration platforms is a natural evolutionary result of ICT penetration not only into the healthcare and its related to quadratic growth of communications among HISs. Crossing a particular limit complexity indicated a need to formalize these communications in objective and also functional manner. So we expect the further development not only in the field of data standards but also in the functional perspective of healthcare integration platforms and EHR.

Acknowledgements

The paper has been supported by the SVV-2012-264 513 project of Charles University in Prague.

References

[1] Bloebel, B. Architectural Approach to eHealth for Enabling Paradigm Changes in Health. Methods of Information in Medicine. 2010, 49, s.123-134.

[2] Bloebel, B., Gonzalez, C., Oemig, F., Lopez, D., Nykanen, P., Ruotsalainen, P. The Role of Architecture and Ontology for Interoperability. Stud Health Technol Inform. 2010, 155, s.33-39.

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

[4] Bloebel, B., Oemig, F., Gonzales, C., Lopez, D.: What is Missing in Health Informatics. Medical and care compunetics, 2010, 156, s.3-12.

[5] Nagy, M., HanzlicekA, P., Preckova, P., Riha, A., Dioszegi M., Seidl, L., Zvarova, J. Semantic Interoperability in Czech Healthcare Environment Supported by HL7 Version 3. Methods of information in medicine. 2010, 49, s.186-195.

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

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

[8] Healthcare Services Specification Project:HSSP [online]. 2012 [cit. 2012-06-19]. Via: http://hssp.wikispaces.com

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

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

[11] ISO/HL7 10781:2009 Electronic Health Record-System Functional Model, Release 1. Geneva, Switzerland: International Organization for Standardization, 2006-2009.

[12] ISO/TS 22600 Health informatics – Privilege management and access control. Geneva, Switzerland: International Organization for Standardization, 2006-2009.

[13] Datovy standard Ministerstva zdravotnictvy CR:DASTA [online]. 2012 [cit. 2012-06-19]. Via: http://dastacr.cz

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

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

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

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

[18] Fowler, M. Patterns of enterprise application architecture. Boston: Addison-Wesley, c2003, xxiv, 533 p. ISBN 03-211-2742-0.

[19] Integrating the Healthcare Enterprise:IHE [online]. 2012 [cit. 2012-06-19]. Via: http://www.ihe.net

[20] Jacobson, I., Fowler, M., Rumbaugh J. Unified software development process. Boston: Addison-Wesley, 1999, 463 s. ISBN 02-015-7169-2.

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

Behavioural Biometrics for Multi-Factor Authentication in Biomedicine

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

[1] EuroMISE Centre, 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., Prague, Czech Republic

Abstract

Objectives: The goal of this work is to suggest an improved authentication method for biomedicine based on analysis of currently used behavioural biometric methods.
Methods: A brief definition of identification, authentication and biometric characteristics is provided. The main part of the work focuses on keystroke dynamics, its advantages, disadvantages and applications in biomedicine. Keystroke dynamics is then proposed as an interesting behavioural biometric characteristic for use in computer security not being widely used so far.
Results: The result of the work will be a new set of methods, which allows optimal multi-factor authentication method regarding its comfort, cost and reliability.
Conclusions: The purpose of this paper is to focus on the available information about keystroke dynamics.

Keywords

Biometrics, anatomical-physiological biometrics, behavioural biometrics, multi-factor authentication, keystroke dynamics, mouse dynamics

Correspondence to:

Anna Schlenker
EuroMISE Centre, Institute of Computer Science, AS CR, v.v.i.
Address: Pod Vodarenskou vezi 2, Prague 8, Czech Republic
E–mail: schlenker.anna@gmail.com

EJBI 2012; 8(5):19–24
recieved: August 15, 2012
accepted: September 20, 2012
published: November 22, 2012

1 Introduction

A wide range of authentication methods have accompanied us through during the whole existence of human society. One group of these methods is directly associated with human physiognomy. This corresponds to the initial recognition of persons by body, face, eyes or voice. It was a system that allowed identification of people in a relatively narrow group, where everyone knows each other. This method obviously has its weaknesses, one can for example temporarily change his/her physical appearance (wigs, fake beards, haircut, glasses etc.) or similar-looking individuals (doubles) may be contained in the group. When comparing only one physiological characteristic, a mistake may occur in simple characteristics such as face shape. In the case of scanning more than one characteristic or complex characteristics (iris or retina), the processing may be slow and uncomfortable for users.

On the other hand, we can use some external attributes, whether it is formal clothing (uniforms), seal rings or passwords. One major weakness of this system is that the external attribute may by stolen by unauthorized person. And it is no matter whether it is a seal ring or token[1].

Based on the shortcomings of single-factor authentication methods presented above, only multi-factor authentication seems adequately reliable to securely eliminate unauthorized access. It can be for example combination of anatomical or behavioural features with an external attribute or password.

2 Identification and Authentication

In biomedicine there is a need to protect information and data. There are two necessary conditions to assure

[1] A security token may be a physical device that an authorized user of computer services is given to ease authentication [18].

that only the authorised person can access or modify the data [4]:

1. identification and
2. personal authentication,

which both together assure the control of the access to the information.

The process of *identification* establishes who the person is. It happens during the initial login to the system, while the *authentication* confirms or denies the personal identity. It also demands a proof of identity to obtain the certainty that the person is really who is affirming to be [4].

Basically, there are three ways in which a person can be authenticated to the system [11, 13]:

1. The first method of authentication is based on something that the person knows, e.g. password or Personal Identification Number (PIN), called a *knowledge factor*.
2. The second method of authentication is based on something that the person has, e.g. a magnetic strip card or a secret key stored on a smart card, called a *possession factor*.
3. The third method of authentication is based on the fact that the person itself has a unique set of measurable characteristics that can be used to verify or recognize the identity of the person. This is called a *biometric factor*.

Security measures belonging to the first two categories are inadequate because possession or knowledge may be compromised without discovery – the information or article may be retrieved from its rightful owner. Therefore, attention is being shifted to reliable identification by biometric techniques that encompass the third class of identification (i.e. biometrics) as a solution for more foolproof methods of identification. For the foreseeable future, these biometric solutions will not eliminate the need for I.D. cards, passwords and PINs. The use of biometric technologies will rather provide a significantly higher level of identification than passwords and cards alone, especially in situations where security is paramount [13].

2.1 Multi-Factor Authentication

Multi-factor authentication is a security system in which more than one form of verification is used in order to prove the identity and allow access to the system. In contrast, single factor authentication involves only one form of verification, most frequently a combination of user ID and password [17].

Additional authentication methods that can be used in multi-factor authentication include biometric verification such as fingerprinting, iris recognition, facial recognition and voice verification. In addition to these methods, smart cards and other electronic devices can be used along with the traditional user ID and password [17].

3 Biometric Characteristics

In the context of authentication, biometrics have several advantages over traditional authentication techniques that verify identity based on something one knows (e.g. a password) or something one has (e.g. a hardware token). In particular, biometric characteristics cannot be forgotten, stolen, or misplaced [9].

Biometric systems recognize a living person (see [19]) and encompass both physiological and behavioural characteristics. Physiological characteristics such as fingerprints are relatively stable physical features that are unalterable without causing trauma to the individual (see [19]). Behavioural traits, on the other hand, have some physiological basis, but also reflect a person's psychological qualities. Unique behavioural characteristics such as the pitch and the amplitude of one's voice, the way of signing names, and even the way of typing, form the basis of non-static biometric systems [13].

Biometric technologies are defined as "automated methods of verifying or recognizing the identity of a living person based on a physiological or behavioural characteristic" [12]. Biometric technologies are gaining popularity because when used in combination with traditional methods for authentication they provide an extra level of security.

3.1 Anatomical-Physiological Biometric Characteristics

Some examples of biometric features used in identification systems include include [19, 5]:

- fingerprints – patterns found on the fingertip, including the location and direction of ridge endings and bifurcations,
- palm prints – a larger-scale version of the fingerprint biometrics,
- hand geometry – shape of the hand including height and width of bones and joints in the palm and fingers,
- blood vessel patterns in the hand – vein and capillary patterns on the palm or the back of the hand,
- patterns in the face – facial characteristics such as position and shape of nose and position of cheekbones, eye sockets and mouth (but not hairline area, which is prone to change),
- patterns in the retina – layer of blood vessels in the back of the eye,

- patterns in the iris – inherent radial pattern and visible characteristics (e.g., freckles, rings, furrows, corona) of the iris.

Today, a few devices based on these biometric techniques are commercially available. However, some of the currently deployed techniques are easy to fool, while others (like iris pattern recognition) are too expensive and uncomfortable for users [19].

3.2 Behavioural Biometric Characteristics

Behavioural biometric characteristics have the advantage of being less obtrusive than other biometric characteristics and do not require special hardware in order to capture necessary biometric data [9]. They are also cheaper and easier to use.

The most known examples of behavioural biometrics are [15]:

- signature dynamics – measurement of combination of appearance, shape, timing and pressure during the writing of user's signature,
- voice verification – tone, pitch and cadence of voice,
- mouse dynamics – measurement of mouse movement distance, speed and angle during the work,
- keystroke dynamics – the duration of each key-press and the time between keystrokes.

4 Keystroke Dynamics

Keystroke dynamics analysis utilizes the way a user types at a terminal to identify users. The identification is based on habitual typing rhythm patterns [13] and realized by constant monitoring the keyboard inputs. It has already been shown that keystroke rhythm is a good sign of identity [10].

Moreover, unlike other biometric systems which may be expensive to implement, keystroke dynamics is almost free – the only hardware required is a keyboard [13, 8].

The application of keystroke rhythm to computer access security is relatively new, but there has been some sporadic work done in this area. Joyce and Gupta [10] present a comprehensive review on the progress in this field prior to 1990. The brief summary of these efforts and examination of the research, that has been undertaken since then, can be found in [13].

Keystroke verification techniques can be classified as either *static* or *continuous* [13].

- *Static verification* approaches analyse keystroke verification characteristics only at specific times, for example, during the login sequence. Static approaches provide more robust user verification than simple passwords, but do not provide continuous security – they cannot detect a change of the user after the initial verification.
- *Continuous verification*, on the contrary, monitors the user's typing behaviour throughout the course of the whole interaction.

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 [3]. This offers a possibility of constant monitoring 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 certain period of time [6].

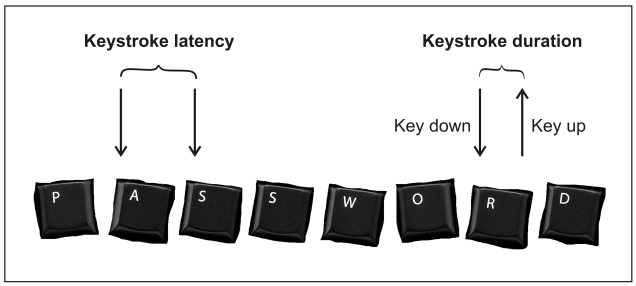

Figure 1: Keystroke duration and keystroke latency.

Some features can be extracted from the keystroke rhythm, for example [4, 19]:

- the period time a key is held for (keystroke duration) – see figure 1,
- the time between individual keystrokes (keystroke latency) – see figure 1,
- frequency of errors,
- style of writing of capital letters,
- speed of the keystroke,
- placement of the fingers and
- pressure that the person applies when pressing a key (pressure keystroke).

The latter three types 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 [13, 8].

The history of keystroke dynamics can be found in [13, 10] or in [4].

We must also mention that there might be a large difference in typing characteristics depending on the current type of user's activity, for example when chatting with

friends compared to writing a program in Java [2]. You need to think more, to analyse and then to type when you are writing a Java program. The set of frequently used characters may also differ (you use more special characters when programming, for example). For more details about this problem, see [2].

4.1 Advantages of Keystroke Dynamics

1. The ultimate goal is ability to continually check the identity of a person as they type at a keyboard [13, 3].
2. Neither login nor verification affect the regular work flow because the user would be typing the needed text anyway. Easy to use for example with login and password during a logon process [21].
3. Unlike other biometric systems, keystroke dynamics is almost free. The only hardware required is the keyboard [13, 8].
4. Time to train the users is minimal and ease of use is very high [21].
5. Public acceptability is very high. There are no prejudices such in a case of fingerprint verification or discomfort such as retina pattern scanning [19].
6. Keystroke dynamics is ideal also for remote users.

4.2 Disadvantages of Keystroke Dynamics

1. Keystroke dynamics is a non-static biometrics like for example voice. This can change quite fast during time, also one-hand typing (due to injury), etc. can influence typing rhythm [13].
2. Low accuracy – keystroke dynamics one of the less unique biometric characteristics [21].
3. Small commercial widespread of technology [21].
4. Dependency on keyboard characteristics, for example layout of keys. Some users may be used to a full-sized keyboard, while the others may prefer to use a laptop, where the typing behaviour will probably be very different [20].
5. Typing style usually differs depending on the language (native vs. foreign) [2].

5 Mouse Dynamics

While authentication with keystroke dynamics has been studied extensively over the past three decades, mouse dynamics has just recently begun to gain interest over the last decade [9]. The idea behind this biometric is to monitor all mouse actions generated as a result of user interaction with a graphical user interface, and then process the data obtained from these actions in order to analyse the behaviour of the user [1].

Mouse dynamics describes an individual's behaviour with a pointing device, such as a mouse or a touch-pad [9]. Similar to keystroke dynamics, mouse dynamics does not require a special device for data collection [16].

Mouse actions can be classified under the following four different categories [14]:

- mouse movement – corresponds to general movement,
- drag and drop – the action starts with mouse button down, movement, then mouse button up,
- point and click – mouse movement followed by a click or double click, and
- silence – no movement.

Same as in other fields of behavioural analysis, mouse dynamics utilizes neural networks and statistical approaches to generate a number of factors from the captured set of actions; these factors are used to construct what is called a Mouse Dynamics Signature or MDS, a unique set of values characterizing user's behaviour over the monitoring period. Some of the factors consist of the calculated average speed against the travelled distance, or the average speed against the movement direction. In [1] up to seven factors that exhibit strong stability and uniqueness capability are reported.

When collecting the actions, several factors have to be taken into account because they can affect the accuracy of the analysis of the mouse biometric samples. These factors are listed below [14]:

1. Desktop Resolution: If the samples are collected with a different screen resolution than assumed, it will affect the results by changing the range of the collected data.
2. Mouse Cursor Speed Setting: This is the speed and acceleration setting of the cursor set by the operating system. Any changes done to those settings can affect the calculated figures, and also affect the user behaviour itself in dealing with the mouse input device.
3. Mouse Button Configuration: In order to achieve reproducible results, the mouse button configuration should be fixed for each user on a specific workstation.
4. Hardware Characteristics: Factors such as the workstation speed, and the pointing device type and properties can also impact the data collection process.

6 Applications in Biomedicine

Keystroke dynamics can be used very well in cooperation with other authentication methods, especially with login and password (structured text), which gain good security results [21]. Now only one company, Net Nanny, works on commercial release of their product BioPassword [7].

There are many potential areas of application for this technology, especially for its low cost and feature of continuous checking. Limitations are mainly non-consistent typists [21].

Monrose [13] also believes that keystroke dynamics can be theoretically used as possible attack to PGP[2], because random seed collected during key generation is calculated from user's typing. This can be weakness, if users typing characteristics are known [21].

Monrose [13] also reports, that there can be some differences between left-handed and right-handed users, but he does not have enough left-handed users to give some useful results [21].

Alternatively, dynamic or continuous monitoring of the interaction of users while accessing highly restricted documents or executing tasks in environments where the user must be alert at all times (for example air traffic control), is an ideal scenario for the application of a keystroke dynamics authentication system. In such case, keystroke dynamics may be used to detect uncharacteristic typing rhythm (brought on by drowsiness, fatigue etc.) and notify third parties [13].

7 Conclusion

For centuries handwritten signature is maintained as an important identification datum. This is a unique expression of human brain. The signature is formed already at school and influenced further by personality and health of individual. We have to accept that a new generation of students is gradually replacing handwriting by typing on a keyboard. So it is appropriate to deal with this new way of human signing. This paper summarizes the available information about this new phenomenon. We can assume that typing has its own specifics, which can be used similarly to the case of handwritten text.

Acknowledgements

This work has been supported by "Projects of Large Infrastructure for Research, Development, and Innovations (LM2010005)" and by the specific research project no. SVV-2012-264513 "Semantic Interoperability in Biomedicine and Health Care", Charles University in Prague.

[2]Pretty Good Privacy (PGP) is a computer program that provides cryptographic privacy and authentication. PGP is often used for signing, encrypting and decrypting electronic mails (e-mails) to increase the security of e-mail communications (see [22]).

References

[1] Ahmed AAE, Traore I. A New Biometrics Technology based on Mouse Dynamics. IEEE Transactions on Dependable and Secure Computing. 2007;4(3):165-179.

[2] Barghouthi H. Keystroke Dynamics. How typing characteristics differ from one application to another. [Master's thesis]. Gjovik, Norway: Gjovik University College; 2009.

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

[4] Boechat GC, Ferreira JC, Carvalho ECB. Using the Keystrokes Dynamic for Systems of Personal Security. Proceedings Of World Academy Of Science, Engineering And Technology. 2006;24(18):61-66.

[5] Coventry L. Usable Biometrics. In: Cranor LF, Garfinkel S, editors. Security and Usability. Sebastopol, CA. O'Reilly Media, Inc.; 2005.

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

[7] Identity Assurance as a Service: AdmitOne Security [Internet] 2010 [cited 2012 Aug 4] Available from: http://www.biopassword.com/

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

[9] Jorgensen Z, Yu T. On Mouse Dynamics as a Behavioral Biometric for Authentication. Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security. 2011;476-482.

[10] Joyce R, Gupta G. Identity authorization based on keystroke latencies. Communications of the ACM. 1990 Feb;33(2):168-176.

[11] Matyas SM, Stapleton J. A Biometric Standard for Information Management and Security. Computers & Security. 2000;19(2):428-441.

[12] Miller B. Vital sings of identity. IEEE Spectrum. 1994;31(2):20-30.

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

[14] Nazar A, Traore I, Ahmed AAE. Inverse Biometrics for Mouse Dynamics. International Journal of Pattern Recognition and Artificial Intelligence. 2008;22(3):461-495.

[15] Olzak T. Reduce multi-factor authentication costs with behavioral biometrics. TechRepublic. [Internet]. 2007 [cited 2012 Aug 5] Available from: http://www.techrepublic.com/article/reduce-multi-factor-authentication-costs-with-behavioral-biometrics/6150761

[16] Raj SBE, Santhosh AT. A Behavioral Biometric Approach Based on Standardized Resolution in Mouse Dynamics. International Journal of Computer Science and Network Security. 2009;9(4):370-377.

[17] Rouse M. Multifactor authentication (MFA) [Internet] 2007 [cited 2012 Aug 10] Available from: http://searchsecurity.techtarget.com/definition/multifactor-authentication-MFA

[18] RSA SecurID [Internet] 2012 [cited 2012 Sep 15]. Available from: http://www.rsa.com/node.aspx?id=1159

[19] Schlenker A, Sarek M. Biometric Methods for Applications in Biomedicine. EJBI. 2011;7(1):37–43.

[20] Senathipathi K, Batri K. Keystroke Dynamics Based Human Authentication System using Genetic Algorithm. European Journal of Scientific Research. 2012;28(3):446-459.

[21] Svenda P. Keystroke Dynamics. [Internet] 2001. [cited 2012 Jul 28] Available from: http://www.svenda.com/petr/docs/KeystrokeDynamics2001.pdf

[22] Zimmermann P. PGP Source Code and Internals. MIT Press; 1995.

Stochastic Models for Low Level DNA Mixtures

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

[1] EuroMISE Centre, 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 increasing sensitivity of forensic analysis methods allows to investigate less and less amount of biological samples. For samples of low quality or quantity, there are stochastic events that require intensive statistical analysis.
Methods: There are several models how to calculate the probability of a given set of alleles. We have described three of them and compared them to verify their accuracy.
Results: The two models proposed in [1] extend so far the most widely used model by the possibility of dropout and peak areas of individual alleles. The first one is incorrect, while the second model highly improves the possibility of DNA mixture analysis.
Conclusions: We have shown the inaccuracy of one of the recently proposed models. We have added the possibility of determining the dropout probability into the second model, otherwise this model overestimates the probabilities calculated.

Keywords

Forensic DNA interpretation, low level samples, allele peak areas, dropout probability

Correspondence to:

Dalibor Slovák
EuroMISE Centre, Institute of Computer Science, AS CR, v.v.i.
Address: Pod Vodarenskou vezi 2, Prague 8, Czech Republic
E–mail: slovak@cs.cas.cz

EJBI 2012; 8(5):25–30
recieved: August 16, 2012
accepted: September 4, 2012
published: November 22, 2012

1 Introduction

With the increasing sensitivity of methods used for forensic DNA analysis, collection of forensic traces can be accomplished from a very small amount of biological material. Therefore, the increasing number of incomplete or contaminated profiles and profiles originating from more contributors are appearing. The samples containing only a small amount of DNA (approximately up to 100 pg / ml) are called low level samples and various stochastic effects occur increasingly for these samples.

Some laboratories perform the analysis of samples twice or more. Curran et al. [2] introduced the set theory in order to enable the calculations to be made in these cases. However, we do not attempt to explain their theory in this paper.

The result of laboratory processing of DNA samples is electropherogram (epg), which displays the alleles present at particular loci and peak heights measured in relative fluorescence units (RFU). Currently the most common laboratory sets process sixteen loci.

Two main approaches to DNA mixture interpretation are currently discussed in forensic practice. The Random Man Not Excluded method (RMNE) calculates the probability of observing the DNA profile needed for evidence, given that the DNA profile comes from a random individual, unrelated to the suspect. In other words, it is the probability that the DNA profile from a random person is the same as the evidence DNA profile, and that this person therefore, due to the evidence, cannot be excluded from suspicion.

The Likelihood Ratio approach (LR) compares the probabilities of observing the evidence under two rival hypotheses: typically the prosecution hypothesis H_p, the probability that the suspect is one of the contributors to the mixture, and the defense hypothesis H_d, the probability that the suspect does not contribute to the mixture.

The advantage of the LR framework is that dropout can be assessed probabilistically and it is the only way to provide a meaningful calculation based on the probability of the evidence under H_p and H_d. A likelihood ratio approach is therefore preferred [3]. For a more detailed comparison of both methods, see [4].

If the allele which is present in the sample is not displayed on the epg we call such an event an (allelic) dropout. If no allele is displayed at the locus, we talk about locus dropout.

If n persons is assumed to contribute to the mixture, maximum of $2n$ alleles can appear at the locus. However, some alleles may be represented several times, others may be missing due to the dropout. The observed mixed profile is therefore usually made up of fewer alleles. Under such conditions, there are more possibilities how to reconstruct individual DNA profiles from observed mixed profile.

Kelly et al. [1] suggested two stochastic models to compute the probability of observing the mixed profile. They compare them with most commonly used model, designated there as the unconstrained combinatorial (UC) method. In this article, the comparison of the three models will be discussed.

Although this theory is easily extended to multiple loci, in the present article, we consider only one locus in the profile and some realities are omitted for simplification, e.g. contamination and drop-in possibility or population structure. The number of contributors to the mixture will be assumed to be known.

2 Methods

From the epg, not only alleles present may be found out but also the peak heights. This information can help us to distinguish e.g. component belonging to the dominant contributor, but even if it is not possible to divide precisely individual components of the mixture, peak heights can inform us about the presence of multiple copies of several alleles. However, the decision on whether the allele is present in multiple copies strongly depends on the assessment of forensic expert and his experience.

The calculation of a LR may proceed by either a binary, a semi-continuous, or a fully continuous method. The binary and semi-continuous methods treat alleles as present or absent, moreover the semi-continuous method assigns a probability to the events of dropout or non-dropout. Fully continuous method deals with the probability of drop-out and other stochastic events based on the heights of the peaks visualised at a locus. Only binary methods are compared here.

Software processing epgs usually shows two thresholds for more simple interpretation. If the signal is below the limit of detection (LOD), we consider it as a noise. The detection limit is usually determined as 25 or 50 RFU or is calculated as the average noise signal plus three its standard deviations.

The stochastic threshold T is a value above which the dropout is excluded. In case that there is only one signal above the stochastic threshold, it may be assumed that it is a homozygous profile [5]. T is usually in the range of 150-300 RFU or may be calculated as the average noise signal plus ten its standard deviations.

Now let us consider two examples with the limit of detection $LOD = 50$ RFU and the stochastic threshold $T = 300$ RFU. The observed profile will be denoted by X and the set of all occurring alleles (allelic vector) will be denoted by A.

Example 1

The alleles 13, 14 and 15 with values of 180, 195, and 212 RFU, respectively, are observed at the locus. The mixture is assumed to originate from two contributors. Thus it is the profile $X = [13, 14, 15]$ for which the peak heights on the epg are approximately the same for all alleles. Under these assumptions, one allele is missing in the allelic vector A - either there was a dropout, or some of the contributors is homozygote, or both contributors have an allele of the same type.

Example 2

The alleles 13, 14 and 15 with values of 150, 470 and 420 RFU, respectively, are observed at the locus. From the analysis of other loci in the same sample, the mixture is assumed to originate from three contributors. Thus it is the profile $X = [13, 14, 15]$ again but now there are three missing alleles to complete the allelic vector. The observed alleles also have quite different peak heights which encourage to the inclusion of multiple copies of some alleles into the allelic vector, but for now we let this opportunity unused. We will return to it later in the section 3.

Now we describe proposed models and show their application to both the examples mentioned above.

2.1 UC Model

The unconstrained combinatorial method does not allow for possibility of dropout nor include peak heights to the calculation. The allelic vector can be completed only by copies of alleles observed.

Example 1:

$$\begin{aligned}
&\mathsf{P}(X = [13, 14, 15]) = \\
&= \mathsf{P}\left(A \in \left\{[13^2, 14, 15], [13, 14^2, 15], [13, 14, 15^2]\right\}\right) = \\
&= \frac{4!}{2!1!1!}p_{13}^2 p_{14} p_{15} + \frac{4!}{1!2!1!}p_{13} p_{14}^2 p_{15} + \frac{4!}{1!1!2!}p_{13} p_{14} p_{15}^2 = \\
&= 12 p_{13} p_{14} p_{15}(p_{13} + p_{14} + p_{15}). \qquad (1)
\end{aligned}$$

Example 2:

$$\begin{aligned}
&\mathsf{P}(X = [13, 14, 15]) = \\
&= \mathsf{P}\left(A \in \left\{[13^4, 14, 15], [13^3, 14^2, 15], [13^3, 14, 15^2],\right.\right. \\
&\quad \left.\left.[13^2, 14^3, 15], [13^2, 14^2, 15^2], [13^2, 14, 15^3], [13, 14^4, 15],\right.\right. \\
&\quad \left.\left.[13, 14^3, 15^2], [13, 14^2, 15^3], [13, 14, 15^4]\right\}\right) = \\
&= \frac{6!}{4!1!1!}p_{13}^4 p_{14} p_{15} + \frac{6!}{3!2!1!}p_{13}^3 p_{14}^2 p_{15} + \frac{6!}{3!1!2!}p_{13}^3 p_{14} p_{15}^2 + \\
&+ \frac{6!}{2!3!1!}p_{13}^2 p_{14}^3 p_{15} + \frac{6!}{2!2!2!}p_{13}^2 p_{14}^2 p_{15}^2 + \frac{6!}{2!1!3!}p_{13}^2 p_{14} p_{15}^3 +
\end{aligned}$$

$$+\frac{6!}{1!4!1!}p_{13}p_{14}^4p_{15}+\frac{6!}{1!3!2!}p_{13}p_{14}^3p_{15}^2+\frac{6!}{1!2!3!}p_{13}p_{14}^2p_{15}^3+$$
$$+\frac{6!}{1!1!4!}p_{13}p_{14}p_{15}^4=$$
$$=30p_{13}p_{14}p_{15}\left(p_{13}^3+2p_{13}^2p_{14}+2p_{13}^2p_{15}+2p_{13}p_{14}^2+\right.$$
$$\left.+3p_{13}p_{14}p_{15}+2p_{13}p_{15}^2+p_{14}^3+2p_{14}^2p_{15}+\right.$$
$$\left.+2p_{14}p_{15}^2+p_{15}^3\right). \quad (2)$$

2.2 F and Q Models

F and Q models were suggested by Kelly et al. [1] as an extension of UC model. Compared to this model, they allow to calculate with the possibility of dropout and to use the information about peak heights.

In F model, any allele completing the observed profile to the allelic vector is denoted by F. For example, under conditions of Example 1 Kelly et al. state

$$\mathsf{P}(X=[13,14,15])=\mathsf{P}(A=[13,14,15,F])=$$
$$=\frac{4!}{1!1!1!1!}p_{13}p_{14}p_{15}=24p_{13}p_{14}p_{15}. \quad (3)$$

However, F model is incorrect due to the non-differentiation between observed and unobserved alleles. If the allele designated as F is of the same type as an allele already observed, the number of possible combinations is less than if we assume that all alleles are different. Thus, F model overestimates computed probabilities. In the case of Example 2, we get $120p_{13}p_{14}p_{15}$ which gives the senseless probability 1.875 for values $p_{13}=p_{14}=p_{15}=0.25$. Therefore, we will continue to consider only model Q.

In Q model, any allele which does not appear on the epg (e.g. due to the dropout) is denoted by Q. The probability of allele marked Q is equal to one minus the sum of the probabilities of observed alleles.

Example 1:

$$\mathsf{P}(X=[13,14,15])=\mathsf{P}\left(A\in\left\{[13^2,14,15],\right.\right.$$
$$\left.\left.[13,14^2,15],[13,14,15^2],[13,14,15,Q]\right\}\right)=$$
$$=\frac{4!}{2!1!1!}p_{13}^2p_{14}p_{15}+\frac{4!}{1!2!1!}p_{13}p_{14}^2p_{15}+\frac{4!}{1!1!2!}p_{13}p_{14}p_{15}^2+$$
$$+\frac{4!}{1!1!1!1!}p_{13}p_{14}p_{15}(1-p_{13}-p_{14}-p_{15})=$$
$$=12p_{13}p_{14}p_{15}(2-p_{13}-p_{14}-p_{15}). \quad (4)$$

Example 2:

$$\mathsf{P}(X=[13,14,15])=$$
$$=\mathsf{P}\left(A\in\left\{[13^4,14,15],[13^3,14^2,15],[13^3,14,15^2],\right.\right.$$
$$[13^3,14,15,Q],[13^2,14^3,15],[13^2,14,15^3],$$
$$[13^2,14,15,Q^2],[13^2,14^2,15^2],[13^2,14^2,15,Q],$$
$$[13^2,14,15^2,Q],[13,14^4,15],[13,14^3,15^2],$$
$$[13,14^3,15,Q],[13,14^2,15^3],[13,14^2,15^2,Q],$$
$$[13,14^2,15,Q^2],[13,14,15^4],[13,14,15^3,Q],$$
$$\left.\left.[13,14,15^2,Q^2],[13,14,15,Q^3]\right\}\right)=\ldots=$$
$$=30p_{13}p_{14}p_{15}\left(2-p_{13}-p_{14}-p_{15}\right)\times$$
$$\times\left(p_{13}^2+p_{14}^2+p_{15}^2+p_{13}p_{14}+p_{13}p_{15}+\right.$$
$$\left.+p_{14}p_{15}-2p_{13}-2p_{14}-2p_{15}\right). \quad (5)$$

3 Inclusion of Peak Heights

As can be seen in equation (5), the number of possible allelic vectors and the complexity of their quantification increases very markedly with a higher number of unknown alleles. In fact, the possibility of the peak height inclusion was not employed to the calculation.

Since the peaks of alleles 14 and 15 (470 and 420 RFU) in Example 2 are above the stochastic threshold (300 RFU) and are significantly higher than the third observed value (150 RFU), alleles 14 and 15 can be assumed to be present in two copies. Taking the peak height into account, observed profile X may be adjusted to $X^*=[13,14^2,15^2]$. Quantification is thus considerably simplified:

$$\mathsf{P}(X^*=[13,14^2,15^2])=$$
$$=\mathsf{P}\left(A\in\left\{[13^2,14^2,15^2],[13,14^3,15^2],[13,14^2,15^3],\right.\right.$$
$$\left.\left.[13,14^2,15^2,Q]\right\}\right)=\frac{6!}{2!2!2!}p_{13}^2p_{14}^2p_{15}^2+$$
$$+\frac{6!}{1!3!2!}p_{13}p_{14}^3p_{15}^2+\frac{6!}{1!2!3!}p_{13}p_{14}^2p_{15}^3+$$
$$+\frac{6!}{1!2!2!1!}p_{13}p_{14}^2p_{15}^2(1-p_{13}-p_{14}-p_{15})=$$
$$=30p_{13}p_{14}^2p_{15}^2(6-3p_{13}-4p_{14}-4p_{15}). \quad (6)$$

The model Q is in this part an appropriate extension of the UC model.

4 Probability of Dropout

As was mentioned, the model Q enables to calculate also with possibility of dropout. Due to the small amount of DNA, allelic dropout of one or more alleles is very common in low level samples. Ignoring the possibility of dropout tends to the disfavour of defense [6] so there are some methods to inform about probabilities of dropout ([7], [8]).

However, the model Q includes dropout to the calculation without considering of its probability. We think that this approach is as incorrect as the exclusion of dropout itself and may results in a strong overestimation of calculated probabilities.

Let us suppose that the dropout probability is determined as $d \in (0,1)$. If the probability of allelic vector is calculated considering allelic dropout, this probability should be multiplied by d. For example, the fourth summand in equation (6) must be multiplied by a value of d:

$$\begin{aligned}\mathsf{P}(X^* = [13, 14^2, 15^2]) &= \frac{6!}{2!2!2!}p_{13}^2p_{14}^2p_{15}^2 + \\ &+ \frac{6!}{1!3!2!}p_{13}p_{14}^3p_{15}^2 + \frac{6!}{1!2!3!}p_{13}p_{14}^2p_{15}^3 + \\ &+ d\frac{6!}{1!2!2!1!}p_{13}p_{14}^2p_{15}^2(1-p_{13}-p_{14}-p_{15}) = \\ &= 30p_{13}p_{14}^2p_{15}^2 \times \\ &\times [6d + 3p_{13}(1-2d) + 2(p_{14}+p_{15})(1-3d)]. \quad (7)\end{aligned}$$

The original formula may be obtained by putting the value of $d = 1$ which means that the dropout occurred with the probability equal to 1. However, it would exclude the possibility that the allele is a copy of some of the observed alleles.

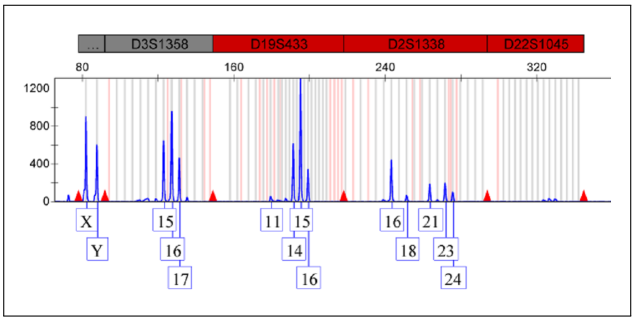

Figure 1: Part of the mixed profile.

If the possibility of two dropouts is assumed, the parameter d must also be considered in the square; if three dropouts are assumed, third power of d is necessary etc. In equation (5), the parameter d should appear in the first, second, and third power. In practice, summands with second and third power have usually an order of magnitude too small to affect the overall probability and could be neglected. See [9] for more complex discussion.

5 Comparison of Models

Figure 1 shows epg of DNA mixture for which three persons are assumed to be contributors. At locus D19S433, four peaks are displayed. Table 1 shows peak heights and allele frequencies in Czech population [10]. There are two suspects with alleles $14, 15$ and $15, 16$. Both calculations are performed independently.

Table 1: Locus D19S433: present alleles and their frequencies in the Czech population.

Allele	Value (RFU)	Frequency
11	55	0.0035
14	610	0.3617
15	1385	0.172
16	391	0.0408

The likelihood ratio is equal to the proportion of probabilities of evidence under prosecution and defense hypotheses:

$$LR = \frac{\mathsf{P}(E|H_p)}{\mathsf{P}(E|H_d)},$$

where H_p means "suspect and two unknown persons contributed to the mixture" and H_d means "three unknown persons contributed to the mixture".

In the following examples we calculate LRs first for the suspect's profile $S_1 = [14, 15]$. Since peak of allele 11 is small, it will be considered later.

5.1 UC Model

Let us evaluate UC model with crime scene profile $X = [14, 15, 16]$ and suspect's profile $S_1 = [14, 15]$.

Hypothesis H_p assumes two persons having together at least one allele 16 and no other than 14, 15 and 16.

$$\begin{aligned}\mathsf{P}(E|H_p) &= \mathsf{P}(X = [14,15,16]|S_1 = [14,15]) = \\ &= 12p_{14}^2p_{15}p_{16} + 12p_{14}p_{15}^2p_{16} + 12p_{14}p_{15}p_{16}^2 + 6p_{14}^2p_{16}^2 + \\ &+ 6p_{15}^2p_{16}^2 + 4p_{14}p_{16}^3 + 4p_{15}p_{16}^3 + 4p_{14}^3p_{16} + \\ &+ 4p_{15}^3p_{16} + p_{16}^4 = 0.0278018\end{aligned}$$

Hypothesis H_d assumes three persons having together alleles 14, 15 and 16 only.

$$\begin{aligned}\mathsf{P}(E|H_d) &= \mathsf{P}(X = [14,15,16]) = \\ &= 30p_{14}p_{15}p_{16}\left(p_{14}^3 + 2p_{14}^2p_{15} + 2p_{14}^2p_{16} + 2p_{14}p_{15}^2 + \right. \\ &+ 3p_{14}p_{15}p_{16} + 2p_{14}p_{16}^2 + p_{15}^3 + 2p_{15}^2p_{16} + \\ &+ \left. 2p_{15}p_{16}^2 + p_{16}^3\right) = 0.01076452\end{aligned}$$

Thus LR for UC model is

$$LR_1 = \frac{\mathsf{P}(E|H_p)}{\mathsf{P}(E|H_d)} = 2.582726. \quad (8)$$

5.2 Original Q Model

If Q model is considered, it may be assumed from analysis of peak heights that allele 15 occurs twice at least. Then the crime scene profile is $X = [14, 15^2, 16]$. The possibility of dropout may be included and let put $p_Q = 1 - p_{14} - p_{15} - p_{16}$.

Hypothesis H_p assumes two persons having together alleles 15 and 16.

$$\begin{aligned}\mathsf{P}(E|H_p) &= \mathsf{P}\left(X=[14,15^2,16]|S_1=[14,15]\right)= \\ &= p_{15}p_{16}\left(4p_{16}^2+6p_{15}p_{16}+12p_{14}p_{16}+12p_{16}p_Q+\right. \\ &\quad +4p_{15}^2+12p_{14}^2+12p_{14}p_{15}+12p_{15}p_Q+ \\ &\quad \left.+\ 12p_Q^2+24p_{14}p_Q\right)=0.0674637\end{aligned}$$

Hypothesis H_d assumes three persons with alleles 14, 15, 15 a 16.

$$\begin{aligned}\mathsf{P}(E|H_d) &= \mathsf{P}\left(X=[14,15^2,16]\right)= \\ &= 30p_{14}p_{15}^2p_{16}\left(2p_{14}^2+2p_{14}p_{15}+3p_{14}p_{16}+6p_{14}p_Q+\right. \\ &\quad +p_{15}^2+2p_{15}p_{16}+4p_{15}p_Q+2p_{16}^2+ \\ &\quad \left.+\ 6p_{16}p_Q+6p_Q^2\right)=0.0377721\end{aligned}$$

LR for original Q model is

$$LR_2=\frac{\mathsf{P}(E|H_p)}{\mathsf{P}(E|H_d)}=1.786072. \tag{9}$$

5.3 Modified Q Model

The process from section 4 is applied. The crime scene profile is $X=[14,15^2,16]$ again and dropout probability is $d=0.45$.

Hypotheses H_p and H_d are the same as in the original Q model, the only change is inclusion of parameter d.

$$\begin{aligned}\mathsf{P}(E|H_p) &= \mathsf{P}\left(X=[14,15^2,16]|S_1=[14,15]\right)= \\ &= p_{15}p_{16}\left(4p_{16}^2+6p_{15}p_{16}+12p_{14}p_{16}+12dp_{16}p_Q+\right. \\ &\quad +4p_{15}^2+12p_{14}^2+12p_{14}p_{15}+12dp_{15}p_Q+ \\ &\quad \left.+\ 12d^2p_Q^2+24dp_{14}p_Q\right)=0.03685446\end{aligned}$$

$$\begin{aligned}\mathsf{P}(E|H_d) &= \mathsf{P}\left(X=[14,15^2,16]\right)= \\ &= 30p_{14}p_{15}^2p_{16}\left(2p_{14}^2+2p_{14}p_{15}+3p_{14}p_{16}+6dp_{14}p_Q+\right. \\ &\quad +p_{15}^2+2p_{15}p_{16}+4dp_{15}p_Q+2p_{16}^2+ \\ &\quad \left.+\ 6dp_{16}p_Q+6d^2p_Q^2\right)=0.01691434\end{aligned}$$

LR for modified Q model is

$$LR_3=\frac{\mathsf{P}(E|H_p)}{\mathsf{P}(E|H_d)}=2.178889. \tag{10}$$

5.4 Modified Q Model with Allele 11

Now, allele 11 is also included to the calculation using modified Q model; it means crime scene profile $X=[11,14,15^2,16]$. Dropout probability is $d=0.45$ again.

Hypothesis H_p assumes two persons with alleles 11, 15 and 16.

$$\begin{aligned}\mathsf{P}(E|H_p) &= \mathsf{P}\left(X=[11,14,15^2,16]|S_1=[14,15]\right)= \\ &= 12p_{11}p_{15}p_{16}\left(p_{11}+2p_{14}+p_{15}+p_{16}+2dp_Q\right)= \\ &= 0.0003889084\end{aligned}$$

Hypothesis H_d assumes three persons with alleles 11, 14, 15, 15 and 16.

$$\begin{aligned}\mathsf{P}(E|H_d) &= \mathsf{P}\left(X=[11,14,15^2,16]\right)= \\ &= 180p_{11}p_{14}p_{15}^2p_{16}\left(p_{11}+p_{14}+p_{15}+p_{16}+2dp_Q\right)= \\ &= 0.0002634395\end{aligned}$$

LR for modified Q model with allele 11 is

$$LR_4=\frac{\mathsf{P}(E|H_p)}{\mathsf{P}(E|H_d)}=1.476272. \tag{11}$$

5.5 Suspect S_2

Calculations for the second suspect $S_2=[15,16]$ are similar. $\mathsf{P}(E|H_d)$ are the same as for first suspect but $\mathsf{P}(E|H_p)$ and hence LRs are much higher:

- LR = 9.929154 for UC model.
- LR = 10.88783 for original Q model.
- LR = 11.58568 for modified Q model.
- LR = 9.904598 for modified Q model with allele 11.

6 Conclusion

Suppose the number of contributors is known and let us briefly summarize the possible statistical processing of epg.

If the number of observed alleles is twice the number of contributors, then all necessary alleles are known and the probability of the profile may be directly calculated. If any alleles are missing in the allelic vector, the procedure from the section 3 is used. The stochastic threshold T is set and the alleles whose peak is above threshold are counted twice. Thereby the set of present alleles is determined more precisely.

If the allelic vector is still incomplete (i.e. the number of alleles $\neq 2n$), all the possibilities of adding any number of alleles present may be calculated. If the possibility of

dropout is also assumed, its probability is predicted and the modified Q model is used as was shown in section 4.

As shown in section 5, substantially different results can be obtained according to the used model and investigated profiles. Generally speaking, the rare alleles present in the profile of the suspect, the higher the likelihood ratio and thus the posterior probability of guilt of the suspect.

When comparing UC and Q model, higher LR was received first and then smaller. On the other hand, it appears that adding of parameter d increases LR because it reduces the denominator more than the numerator.

Acknowledgements

The paper has been supported by the SVV-2012-264 513 project of Charles University in Prague. The authors are grateful to Vlastimil Stenzl for providing the data.

References

[1] Kelly H, Bright J-A, Curran J, Buckleton J. The interpretation of low level DNA mixtures. Forensic Sci Int Genet. 2012; 6: 191–197

[2] Curran JM, Gill P, Bill MR. Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure. Forensic Sci Int. 2005; 148: 47-–53.

[3] Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz M, Schneider PM, Weir BS. DNA Commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. Forensic Sci Int. 2006; 160: 90-–101.

[4] Buckleton J, Curran J. A discussion of the merits of random man not excluded and likelihood ratios. Forensic Sci Int Genet. 2008; 2: 343–348

[5] Gill P, Puch-Solis R, Curran J. The low-template-DNA (stochastic) threshold - Its determination relative to risk analysis for national DNA databases. Forensic Sci Int Genet. 2009; 3: 104–111

[6] Balding DJ, Buckleton J. Interpreting low template DNA profiles. Forensic Sci Int Genet. 2009; 4: 1–10

[7] Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. Estimating the probability of allelic drop-out of STR alleles in forensic genetics. Forensic Sci Int Genet. 2009; 3: 222–226

[8] Haned H, Egeland T, Pontier D, Pène L, Gill P. Estimating drop-out probabilities in forensic DNA samples: A simulation approach to evaluate different models. Forensic Sci Int Genet. 2011; 5: 525–531

[9] Gill P, et al. DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. Forensic Sci Int Genet. 2012; in press

[10] Šimková H, Faltus V, Marvan R, Pexa T, Stenzl V, Brouček J, Hořínek A, Mazura I, Zvárová J. Allele frequency data for 17 short tandem repeats in a Czech population sample. Forensic Sci Int Genet. 2009; 4: e15–e17

Mutation Analysis of the COL1A1 Gene in Czech Patients Affected by Osteogenesis Imperfecta, Type I-IV

Lucie Šormová[1], **Ivan Mazura**[1], **Ivo Mařík**[2]

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

[2] Ambulant Centre for Defects of Locomotor Apparatus, Prague, Czech Republic

Abstract

Background: Osteogenesis imperfecta is a worldwide widespread disorder of connective tissue characterized by extensive clinical heterogeneity. The main clinical feature is increased bone fragility due to defective collagen type I production which is encoded by two genes – COL1A1 and COL1A2. Based on clinical, radiological and genetic features there is described 11 forms of the disease. Only the first four types result from the collagen type I mutations. Severity of the disorder ranges from mild to lethal forms.
Objectives and Methods: The aim of this study is the molecular-genetic analysis of COL1A1 gene of 25 Czech patients suffering from the disease named osteogenesis imperfecta, specifically type I-IV, and comparison of clinical pictures of individuals with the same identified mutations.
Results: COL1A1 gene mutations were identified in three of twenty-five Czech OI patients. These individuals come from unrelated families and are affected by osteogenesis imperfecta type IA, III and IVB.
Conclusion: Further molecular-genetic analyses of other patients and their relatives are important for detection of the biggest mutational spectrum necessary for determination of possible genotype phenotype relationship of affected individuals and for comparison the Czech population with others countries.

Keywords

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

Correspondence to:

Lucie Šormová
First Faculty of Medicine, Charles University in Prague
Address: Kateřinská 32, 121 08 Prague 2, CR
E–mail: black.luca@seznam.cz

EJBI 2012; 8(5):31–38
recieved: August 15, 2012
accepted: October 2, 2012
published: November 22, 2012

1 Introduction

Osteogenesis imperfecta (OI) is a heritable disorder of connective tissue. Hallmark feature of the disease is increased bone fragility with increased risk of fractures. Other associated signs are subnormal to low stature, joint hypermobility, skin hyperlaxity, blue sclera, hearing loss and dentinogenesis imperfecta. Some patients suffer from pulmonary or vascular defects.

The first classifications created in 1979 by David Sillence included four clinical different OI types. Current classification distinguishes eleven forms on the basis of clinical, radiological and genetic signs. First four types result from collagen type I genes (Collagen, type I, alpha-1 (COL1A1) and Collagen, type I, alpha-2 (COL1A2)) mutations.

Origin of remaining types are mutations of the gene Serpin peptidase inhibitor, clade F, member 1 (SerpinF1) (OI type VI), genes of the 3-prolyl hydroxylation complex - Cartilage associated protein (CRTAP), Leucine- and proline-enriched proteoglycan 1 (LEPRE1) and Peptidyl-prolyl isomerase 1 (Cyclophylin B) (PPIB) (OI types VII, VIII and IX), and defects of chaperones Serpin peptidase inhibitor, clade H, member 1 (SerpinH1) and FK506-binding protein 10 (FKBP10) (OI types X and XI). Etiology of the fifth OI type is currently unknown [5, 18].

1.1 Collagenous Forms of Osteogenesis Imperfecta

OI type I is the mildest form of OI inherited by autosomal dominant manner. The risk of fractures is increased in childhood, after woman's menopause and after 60th year of life in men. Individuals are normal stature, have mild or no deformities, may have blue sclera or suffer from hearing loss. Some of them have dentinogenesis imperfecta (DI). This feature distinguishes type IA (absent DI) and type

IB (presence DI) of OI. Presence of all clinical signs is very variable [9, 17].

OI type II is a perinatal lethal form. Stillborn is common, perinatal mortality occurs in 80% of cases in the first week of the life. First fractures occur in the uterus, patients have severe deformed bones, triangular face and blue or grey sclera. There are three subtypes of this form – types IIA, IIB and IIC, differentiated according to radiographic features such as deformity of ribs and long bones and cephalometric features (macrocephaly, microcephaly), type of heredity and mutated gene (autosomal dominant types IIA and IIC result from mutations of COL1A1 and COL1A2 genes, the autosomal recessive type IIB is caused by mutation of CRTAP gene) [2].

The third type of the disease is moderately deforming form of OI with autosomal dominant or recessive inheritance. Patients achieve subnormal body height. They have short extremities, severe deformities of bones, hypermobile joints, triangular face, dark blue sclera (turn white in adulthood) and DI. Typical radiological features are wormian bones of skull and popcorn calcification of epiphyses and metaphyses of long bones. Severe scoliosis, thin diaphyses of long bones and high frequency of fractures during normal daily activities are the main reason for using of the wheelchair [13].

OI type IV is the most heterogeneous type of this disorder with autosomal dominant inheritance. Growth retardation is moderate to severe, affected individuals have bowing bones, popcorn-like structure of epiphyses is less common than in the OI type III. First fractures may occur at birth, sclera is white, blue or grey and some patients suffer from otosclerosis. Typical clinical feature is basilar impression. Based on presence of DI we distinguish the types IVA (absent DI) and IVB (presence DI) [7].

1.2 Non-Collagenous Forms of Osteogenesis Imperfecta

OI type V is the autosomal dominant osteogenesis imperfecta type with unknown genetic origin. It is moderate deforming form which presents with hypertrophic callus formation in areas of fractures and with interosseous ossification of the forearm bones [5, 12].

The sixth type of the disease is inherited by autosomal recessive manner. It is a progressive deforming disorder characterized by presence of bone lamellae like fish scale, osteopenia, long bone deformities and bulbous metaphyses [12, 13]. It is caused by SerpinF1 gene mutations [5].

Type VII OI is an autosomal recessive OI form with severe to lethal clinical manifestation. Main signs of this type are rhizomelic shortening of humerus and femur and exophtalmos. Frequency of fractures decreases throughout adulthood. It results from CRTAP gene mutations [16].

Next autosomal recessive form is OI type VIII. Phenotype of affected individuals is various from severe to lethal. The typical clinical feature is rhizomelic shortening of extremities. Other radiological features are bulbous epiphyses, osteoporosis and shortened long bones. Causative gene of this OI type is LEPRE1 [4, 5].

Osteogenesis imperfecta type IX, the moderate to lethal form, resembles with its clinical picture the III and the IV type of the disease. Familial transmission of the disorder is autosomal recessive. This type results from defects in a PPIB [9].

A severe to lethal OI type X, the autosomal recessive form of the disease, results from SerpinH1 gene mutations. Phenotype of individuals is presented by rhizomelic shortening of extremities like in types VII and VIII [5].

The last type of OI is the type XI. It is a progressive deforming form inherited in autosomal recessive manner and caused by defects of a FKBP10 gene. Typical clinical features are bone lamellae like fish scale as well as in the sixth type of the disorder [5].

1.3 Molecular-Genetic Origin of Osteogenesis Imperfecta

80-90% of OI cases are caused by mutations in one of two collagen type I genes - COL1A1 and COL1A2. The molecule of the protein is composed of two alpha1 chains encoded by the COL1A1 gene localized on chromosome 17 and one alpha2 chain encoded by the COL1A2 gene situated on chromosome 7. The unfolded chains undergo several modifications (4-prolyl hydroxylation, 3-prolyl hydroxylation, lysine hydroxylation, glycosylation) increasing stability of the molecule. Such modified alpha chains fold in the direction from the C-terminus to the N-terminus in a heterotrimer terminated by C- and N-propeptides (this is the reason of more severe disability in individuals whose collagen is mutated in the C-region of the molecule) [1, 9].

The most important amino acid in the alpha chain is glycine (Gly) that produces inter-chains links. It is contained in every third position in 338 repetitive Gly-X-Y sequences and is required for correct alpha chains folding into the triple helix formation. About 75%-80% of structural defects of collagen type I result from substitution mutations of another amino acid instead of glycine [10]. 36% of COL1A1 glycine substitutions are lethal while in COL1A2 gene 19% of mutations of this amino acid have lethal outcome [5].

The other crucial areas of the alpha chains are the transcription factors binding sites - activating proteins (so-called enhancers and silencers) binding sites whose binding to the alpha chain activates or inhibits transcription [6], CpG rich areas can undergo methylation resulting

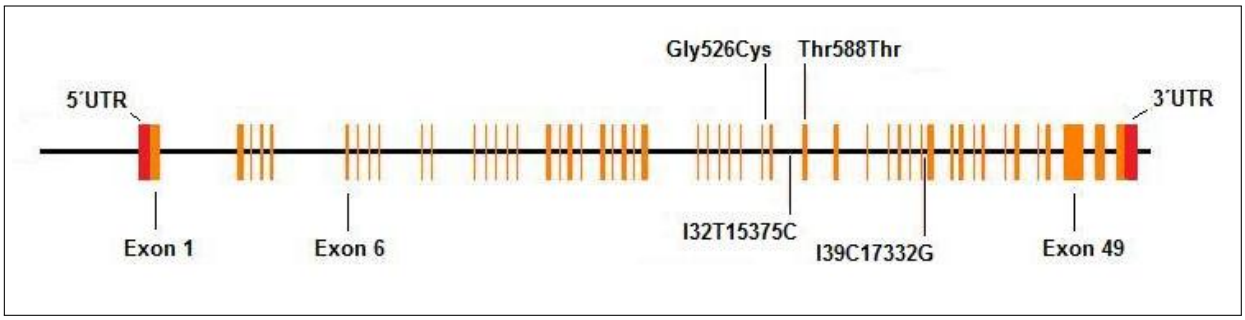

Figure 1: Overview of identified mutations/polymorphisms in the gene COL1A1.

in moderate phenotype if it occurs in promoter, exon 1 or intron 1 and in severe clinical picture if this occurs in the coding sequence of COL1A1 and COL1A2 genes. This process can happen in 26 of the 338 glycine codons of the alpha chains [8].

Finally, there are three Multi Ligand Binding Regions (MLBR1-3) producing intermolecular linkages with other molecules of the connective tissue, for example integrins, Cartilage Oligomeric Matrix Protein (COMP), SerpinH1 and other. These interactions increase strength and flexibility of bones. Mutations of MLBR2 and MLBR 3 result in most cases in lethal osteogenesis imperfecta [5, 15].

1.4 Treatment of Individuals Affected by Osteogenesis Imperfecta

Treatment of patients with osteogenesis imperfecta is different and individual based on concrete clinical, biochemical and radiological picture. Medical treatment includes calcium, vitamin D and bisphosphonates therapy. Bisphosphonates are the most commonly used medicaments for moderate and severe forms of OI. Their specific function is inhibition of osteoclasts on the surface of bones leading in increase of bone mineral density and decrease of risk of fractures [3].

Orthotic treatment is introduced for patients with scoliosis and mild deformities of extremities, while severe deformities and fractures with significant displacement are treated surgical using osteotomy and fixation with intramedullary rods, nails, pins etc. Severe scoliosis is surgically resolved by fixation with Harrington rods, however, this procedure greatly reduces subsequent range of motion of the spine.

At present, methods called cell and gene therapy are being developed. The aim of these methods is replacement of defective osteoblasts with subsequent increasing of bone mineral density (cell therapy) and deactivation of the mutated gene resulting in decreasing of OI severity (gene therapy) [9, 11].

2 Materials and Methods

We analyzed in this research gDNA samples obtained from whole blood of the 25 Czech patients (four unrelated families and seventeen sporadic cases) diagnosed with osteogenesis imperfecta type I-IV nineteen of these individuals are affected by OI type IA, five suffer from the third type of the disorder and one is diagnosed with OI type IVB. All of them signed an informed agreement permitting the molecular genetic analysis of their DNA. Blood samples were collected at several workplaces in the Czech Republic, such as Prague, Brno, Hradec Králové, Olomouc or Ostrava. Molecular-genetic analyses of the isolated gDNA were focused on the COL1A1 gene.

The gDNA was isolated by using the QIAamp DNA Blood Midi Kit (QIAGEN) and stored at -20°C. The quality of isolated samples was determined by gel electrophoresis and the quantity was detected spectrophotometrically.

Thus verified DNA samples were amplified using a polymerase chain reaction (PCR) and specially de-

Table 1: Overview of detected mutations/polymorphisms in Czech OI patients.

Patient No.	OI form	Gender	Age (years)	Nucleotide change	Mutation/ Polymorphism	COL1A1 position
1	III	Female	23	GGC/TGC	Gly526Cys	exon 31
2	IA	Male	22	ACT/ACC T/C C/G	Thr588Thr I32T15375C I39C17332G	exon 33 intron 32 intron 39
3	IVB	Female	52	ACT/ACC T/C C/G	Thr588Thr I32T15375C I39C17332G	exon 33 intron 39 intron 31

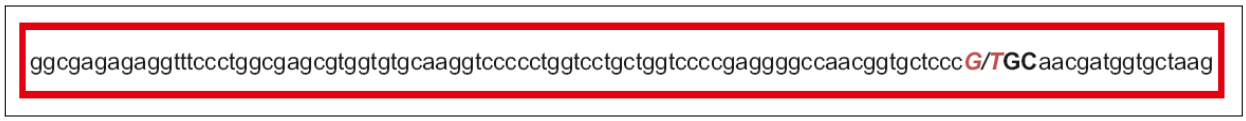

Figure 2: Sequence structure of the exon 31. Position of the Gly526Cys substitution is marked in bold, the changed nucleotide is marked in red italics.

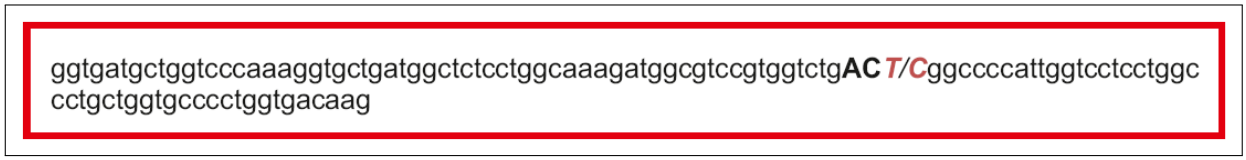

Figure 3: Sequence structure of the exon 33. Position of the Thr588Thr substitution is marked in bold, the changed nucleotide is marked in red italics.

signed 100% complementary primers focused to six regions (G1-G6) of the DNA involving exons 31 to 40. This section was chosen based on presence of the multi ligand binding region 2. The sequences of the used oligonucleotide primers are G1-1 CATCCGTCAAGGTGCGTCG and G1-2 CCTGCCCTGGTCTTTTCCC which amplify a 350bp long region including the exon 31; G2-1 CTGGAGTCTGGGCTGTGAG and G2-2 GTGTTCTGCTTGTGTCTGGG primers producing product with length of 660bp containing the exon 32; G3-1 CCAGACACAAGCAGAACACT and G3-2 CTGAGAGCAAGGGACAAGA focused on a 402bp long region including the exon 33; G4-1 TCAACCTGGGAACCTGGAG and G4-2 CAGCATCGCCTTTAGCACC that produce a 662bp long PCR product comprising exons 34 and 35; G5-1 TTCCTGCCTCCATTACTGC and G5-2 AACAGCCAACTCATCCGAC amplifying a 426bp long region with exons 36 and 37; and in conclusion primers G6-1 GGTGCTACTGGTTTCCCTGG and G6-2 TCTGTTCTCCTTGGCTCCGC defining a 645bp long DNA region containing exons 38, 39 and 40.

The polymerase chain reaction amplification was performed in 50 μl final volume, with 100 ng of genomic DNA, 25 μl Taq PCR MasterMix (1000U) (QIAGEN) (contains Taq polymerase (5 U/μl), PCR Buffer, $MgCl_2$ (1,5 mM), dNTPs (4 x 200 μM)) and 0,5 μl (50 pmol) of each of the oligonucleotide primers.

We performed 35 cycles of 0,5 min at 95°C, 0,5 min at 59°C (system G1)/ 58°C (systems G2, G4 and G6)/ 57°C (system G3) /53°C (system G5), and 1 min at 72°C. The amplified products were electrophoresed through a 2% agarose gel.

Sequencing of PCR-amplified COL1A1 gene fragments was carried out using an automatic capillary sequencing method. We used in this research BigDye® Terminator v3.1 Cycle Sequencing Kit (Life Technologies Corporation, USA) protocol.

Obtained data were compared with corresponding COL1A1 gene segments of the healthy population – for this analysis were used PC programs DNA Baser and SeqScape® Software for Mutation Profiling v2.7. Identified mutations were compared with the OMIM database.

3 Results

Molecular-genetic analyses revealed mutations in three of twenty-five analysed Czech OI patients. Changes of the DNA were found in both coding and noncoding regions of the COL1A1 gene namely in exons 31 and 33 and in introns 32 and 39 (Table 1, Figure 1).

We detected a substitution of glycine to cysteine at position 526 in the exon 31 (Figure 2) in the DNA sample from a patient affected by the third type of OI (section 3.1.1). This mutation is the most common described change within this form of the disorder. First time was this change described in 1989 by Starman et al. [14].

The second mutation of the coding sequence was identified in patients diagnosed with osteogenesis imperfecta type IA (section 3.1.2) and IVB (section 3.1.3). In both cases it is a silent mutation of threonine at position 588 in the exon 33 (Figure 3). In these patients were also detected noncoding sequences changes of introns 32 (Figure 4) and 39 (Figure 5).

3.1 Case Reports

3.1.1 Osteogenesis Imperfecta Type III

The first case is a 23 years old woman born from the second gravidity in the family with unaffected parents. Birth was performed by Caesarean section. The newborn was resuscitated. She weighed 2450 g, birth length was 45 cm. The diagnosis was confirmed immediately at the birth.

Clinical features presented in this individual are blue sclera, trigonocephaly, hyperbrachycephaly, wormian bones of the skull, moderate exophtalmos and hypermobile joints. The woman does not suffer from otosclerosis

Figure 4: Sequence structure of the intron 32. Position of the I32T15375C polymorphism is marked in red italics.

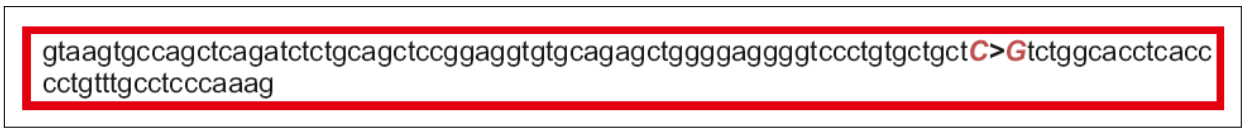

Figure 5: Sequence structure of the intron 39. Position of the I39C17332G polymorphism is marked in red italics.

and dentinogenesis imperfecta. Fractures of both femurs occurred during childbirth, X-rays showed up healed fractures of ribs and the left clavicle. The patient during childhood and adulthood suffered from multiple fractures especially of long bones of lower and upper extremities. The last fracture – fracture of the right clavicle, was described at the age of 19 years. She began to walk at six years with help of leg orthosis. She has used the wheelchair since 11 years of age. Radiological examinations of skeleton at the age of 15 and 18 year revealed suspicion on osteoporosis. Densitometric scans confirmed some decrease of bone mineral density (BMD). Anthropologic and X-ray examinations verified the presence of barrel chest with deformed ribs, pectus carinatum, severe scoliosis, platyspondyly of thoracic vertebras, higher bodies of lumbar vertebras, biconcave shape of thoracic and lumbar vertebras, deformation of skull bones, angulation of the right forearm and femurs, saber shaped deformities of humeri and tibias and shortening of femurs. Metaphyses and epiphyses of bones of the knees have popcorn-like structure, the typical radiological sign of the third type of this disorder (Figure 6).

Medical treatment namely with calcitonin has started at the age of seven years. Treatment with bisphosphonates has begun seven years later. A part of the medical treatment is supplementation with calcium and vitamin D3. The patient has undergone a lot of surgeries since 2nd year of age (corrective and multiple osteotomies with intramedullary nailing). Orthotic treatment was a part of comprehensive treatment since 6 to 16 years of life.

Molecular-genetic analyses identified the most typical mutation for this OI type – Gly526Cys. Further was at the patient identified mutation of MTHFR gene (heterozygous A1298C) increasing blood coagulation.

3.1.2 Osteogenesis Imperfecta Type IA

The second case of our report is a case of a 22 years old man affected with the 1st type of the disease. The patient was born from the third gravidity of unaffected couple. Birth weight was 2800 g, the birth length was 50 cm.

He has light blue sclera (Figure 7), suffers from hearing loss and tinnitus. On the skin of face, neck and chest are numerous lentigo. Other clinical signs include slim chest with narrow vertical ribs, high palate, weak muscles, hypermobility of joints and asymmetric shoulders. The first fracture occurred in the age of 2 years.

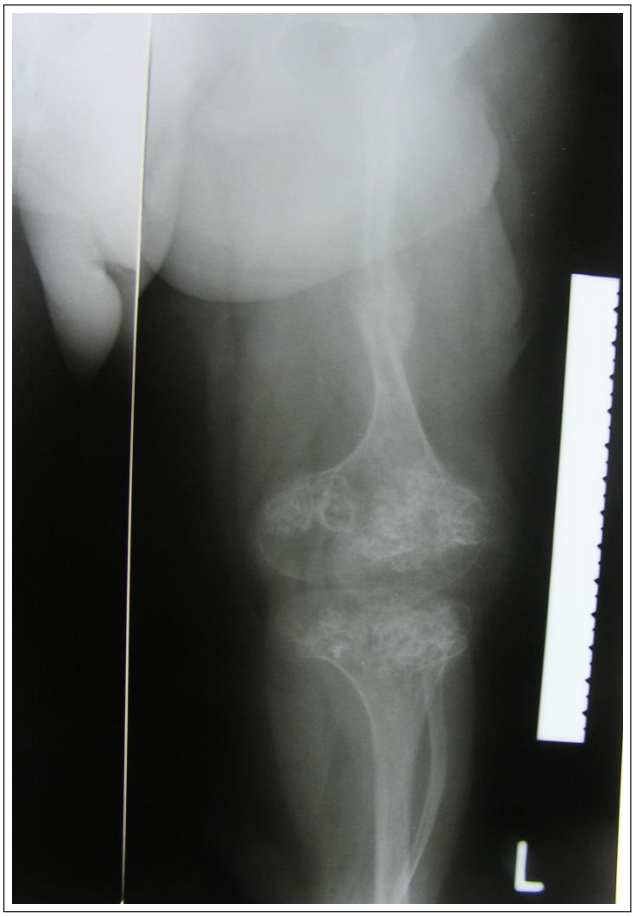

Figure 6: Popcorn-like structure of the femoral epiphysis of the patient suffering from OI type III.

Other fractures occurred at 9 years (fracture of the thoracic vertebrae), at 11 years (fractures of both ulna) at 15 years of life (fracture of the second metacarpal bone of the right hand). X-rays present deformities of the spine – straightened thoracic kyphosis, flattened lumbar lordosis, platyspondyly of thoracic vertebrae and moderate scoliosis. Long bones of the lower limbs are mild bowed and the patient has shortened fourth and fifth metatarsal and digit bones. Densitometric examination confirmed low bone mass according to chronologic age (Z-score is less than -2,0).

The patient is treated with bisphosphonates, calcium and vitamin D3. He has undergone many surgeries, such as incorporation of Kirschner´s rods and tympanostomy.

Figure 7: Light blue sclera of the patient affected by OI type IA.

3.1.3 Osteogenesis Imperfecta Type IVB

A 53 years old woman affected by the 4th type of osteogenesis imperfecta is the first child in healthy family without signs of increased bone fragility. Birth anthropometric parameters were 2840 g and 47 cm.

The patient has blue sclera, otosclerosis and dentinogenesis imperfecta - she lost her second dentice when she was 20 years old. She has generalized joint hypermobility, short body and lower limbs and suffers from back pain. The patient suffered multiple fractures especially of bones of lower limbs since she was 2 years old. The last fracture occurred at the age of 14 years. X-ray examinations demonstrate biconcave shape of thoracic and lumbar vertebrae bodies, saber deformities of tibias (Figure 8) and right femur, varus femoral necks and valgus heels. Densitometric examinations determine osteoporosis of the skeleton (T-score is less than -2,5).

Medical treatment with bisphosphonates has begun at 42nd year of age. She is further also treated with vitamin D3. The woman has undergone only one surgery when she was 47 namely of the left femur. Currently she uses wheelchair or crutches and a knee brace.

Other molecular-genetic analyses identified a heterozygous mutation of a MTHFR gene (A1298C) and homozygous mutation of a UGT 1A1 gene (7TA/7TA) that causes Gilbert syndrome.

Figure 8: Saber deformity of the left tibia of the patient diagnosed with OI type IVB.

The family anamnesis in this case is interesting because the husband and daughter of this patient are affected by the Charcot Marie Tooth syndrome in combination with the diabetic neuropathy. Her daughter also suffers from muscles atrophy, cramps and paresthesia of lower limbs.

4 Discussion

Osteogenesis imperfecta is the highly heterogeneous disorder with molecular-genetic background in mutations especially of genes coding the collagen type I. The clinical picture of affected patients differs inter- and intra-group. Currently, world literature describes some relationships between positions of mutations and resulting phenotype of individuals. Generally, lethal phenotype results from mutations situated to the C terminus of alpha chains, substitutions of the glycine and substitutions by amino acids with branched side chains. It results further from mutations resulting in skipping of exons 3´ (especially exons 14, 20, 22, 27, 30, 44 and 47) of the COL1A1 and exons 5´ to the exon 27 of the COL1A2 gene) and from mutations resulting in creation of alternative or cryptic splice sites [8]. There are also two regions named MLBR2 and MLBR3 within the alpha1 chain and eight regions of the alpha2 chain whose mutations result namely in lethal OI types II or III. On the other hand, mutations of the first 200 amino acids, the glycine substitutions at the first 85-90

amino acids, nonsense mutations resulting in production of STOP codons and changes situated in the N terminal area of the alpha chains exhibit nonlethal clinical picture [4, 8]. In conclusion we can say that in general mutations of the COL1A1 gene usually display in more severe clinical features than these of the COL1A2 gene. But we should not forget that other factors such as genetic, nutrition or environmental changes may affect expression of mutations.

In this study we analyzed 25 Czech patients suffering from collagenous forms of OI. DNA defects were detected in three of these patients. These changes are in two cases novel single point mutations or polymorphisms.

The glycine substitution for cysteine at the position 526 was determined in the case of a woman diagnosed with OI type III. Starman et al. described this mutation in 1989 in an Iraqi individual. Both of these patients showed similar clinical signs such as deformation of bones, presence of wormian bones, fractures at birth, blue sclera and defective dentin production without dentinogenesis imperfecta [14]. This mutation is situated in the integrins binding region of the alpha chains. Variations of this area affect production of intermolecular and molecule-extracellular matrix linkages and decrease strength of bones. Because it is the most common substitution identified in OI type III patients we can conclude that it results in severe bone deformity.

The Thr588Thr mutation was identified in two patients suffering from different OI types – types IA and IVB. Despite this the patients have some of the identical features – blue sclera, hearing loss, hypermobility of joints and osteoporosis. Although the silent threonine 588 substitution does not alter the reading frame it can negatively affect translation parameters and production of intermolecular linkages with the COMP which binds to the collagen type I at the site defined by codons 582 to 638. We can consider that a silent mutation may predict development of osteopenia and osteoporosis due to change of one of some nucleotides in COMP binding site. However, this is only a speculation. Currently, literature does not describe this silent mutation.

Both of identified polymorphisms (I32T15375C, I39C17332G) were detected in patients with the same substitution Thr588Thr in exon 33. Any of these changes result neither to formation of STOP codons nor to the production of an extended/shortened product due to using of cryptic splice-sites. It follows that they do not result in defective production of the collagen type I. Currently any worldwide literature does not describe these polymorphisms.

5 Conclusion

We collect currently further biologic material such as venous blood, bone grafts or tissue of aborted embryos of the Czech patients affected by osteogenesis imperfecta type I-IV for other molecular genetic analyses focused on the other coding sequences of the COL1A1 gene. For next analyses we will use methods High Resolution Melting Analysis and the Sanger sequencing technology. This will be performed in cooperation with the Centre for Medical Genetics – University of Antwerp, Edegem, Belgium.

Acknowledgements

Acknowledgements belong to Mgr. T. Pexa and to RNDr. M. Zachová, Ph.D. from the Laboratory of Forensic Genetics in Brno for the provision of laboratory facilities for molecular-genetic analyses and MUDr. Olga Hudáková, PhD. for providing useful information regarding the clinical description of the different forms the disease. The work was supported by the CBI project No. IM06014 and by the SVV-2012-264 513 project of Charles University in Prague.

References

[1] Alanay Y, Avaygan H, Camacho N, Utine GE, Boduroglu K, et al. Mutations in the Gene Encoding the RER Protein FKBP65 Cause Autosomal-Recessive Osteogenesis Imperfecta. The American Journal of Human Genetics. 2010 Apr; 86: 551–559.

[2] 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.

[3] Becker J, Semler O, Gilissen C, Li Y, Bolz HJ et al. Exome Sequencing Identifies Truncating Mutations in Human SERPINF1 in Autosomal-Recessive Osteogenesis Imperfecta. The American Journal of Human Genetics. 2011 Mar; 88: 362–371.

[4] 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.

[5] Forlino A, Cabral WA, Barnes AV, Marini JC. New perspectives on osteogenesis imperfecta. Nat. Rev. Endocrinol. 2011; 7: 540–557.

[6] Ghosh AK. Factors Involved in the Regulation of Type I Collagen Gene Expression: Implication in Fibrosis. Exp. Biol. Med. 2002; 227: 301-314.

[7] Kashyap RR, Gopakumar R, Gogineni SB, Sreejan CK. Osteogenesis imperfecta type IV. Kerala Dental Journal. 2009 Jan; 32(1): 47-49.

[8] Marini JC, et al. Consortium for Osteogenesis Imperfecta Mutations in the Helical Domain of Type I Collagen: Regions Rich in Lethal Mutations Align With Collagen Bonding Site for Integrins and Proteoglycans. Human Mutation. 2007; 28(3): 209-221.

[9] Marini JC. Osteogenesis imperfecta. 2010. Available at: http://www.endotext.org/parathyroid/parathyroid17/parathyroid17.pdf. (Revised 1 March 2010).

[10] Marini JC, Cabral WA, Barnes AM. Null mutations in LEPRE1 and CRTAP cause severe recessive osteogenesis imperfecta. Cell Tissue Res. 2010 Jan; 339(1): 59–70.

[11] Niyibizi C, Wang S, Mi Z, Robbins PD. Gene therapy approaches for osteogenesis imperfecta. Gene Therapy. 2004; 11: 408-416.

[12] Roughley PJ, Rauch F, Glorieux FH. Osteogenesis imperfecta – clinical and molecular diversity. European Cells and Materials. 2003; 5: 41-47.

[13] Sorin H, Cornel C, Cristian CG, Iuliana P. Osteogenesis imperfecta: forensic assessment of traumatic injuries. Case report and literature review. Rom J Leg Med. 2008; 16 (4): 275 – 282.

[14] Starman BJ, Eyre D, Charbonneau H, Harrylock M, Weis MA, Weiss L, Graham JM, Byers PH. The position of substitution for glycine by cysteine in the triple helical domain of the proalpha1(I) chains of type I collagen determines the clinical phenotype. J. Clin. Invest. 1989; 84:1206–1214.

[15] 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.

[16] Ward LM, Rauch F, Travers R, Chabot G, Szout EM, Lalic L, Roughley PJ, Glorieux FH. Osteogenesi imperfecta type VII: an autosomal recessive form of brittle bone disease. Bone. 2002; 31: 12-18.

[17] Wollina U, Koch A. Osteogenesis imperfecta type I and psoriasis – a report on two cases. Egyptian Dermatology Online Journale. 2006 Jun; 2(1): 15.

[18] Yang Z, Zeng C, Wang Z, Shi HJ, Wang LT. Mutation characteristics in type I collagen genes in Chinese patients with osteogenesis imperfecta. Genetics and Molecular Research. 2011; 10 (1): 177-185.

Obesity Treatment by Bariatric Surgery and Some of the Pharmacoeconomical Aspects in the Czech Republic

Zdeněk Telička[1], **Štěpán Svačina**[1], **Martin Matoulek**[1]

[1] 3rd Medical Department, 1st Faculty of Medicine, Charles University and General Faculty Hospital in Prague, Czech Republic

Abstract

Background: Obesity affects one in four people in the Czech Republic and its incidence is growing worldwide. In this article we focused on evaluation of treatment of obesity in diabetic patients by bariatric surgery and we also tried to evaluate the costs of the surgery and antidiabetics.
Methods: The total number of patients was 200 and 30 of them with type 2 diabetes mellitus. In the 1-year follow-up we evaluated remission or compensation of diabetes in patients after particular bariatric methods. We also calculated the decrease of average costs for pharmacotherapy by antidiabetics after 6 and 12 months and the costs for the bariatric surgery.
Results: We found that costs for the treatment by antidiabetics were reduced nearly 3× already in the 6th month after the surgery. Insurance companies currently do not take in consideration different costs for the partial surgery methods and the payment is in one package for approx. 60 thousands CZK.
Conclusion: The positive effect of the surgery appeared in the 6th month of the follow-up. However, to achieve more accurate results we need to evaluate the data after 3 years of the follow-up.

Keywords

Diabetes mellitus, pharmacoeconomics, body mass index, bariatric surgery

Correspondence to:

Zdeněk Telička
3rd Medical Department, 1st Faculty of Medicine,
Charles University and General Faculty Hospital in Prague, CR
Address: U Nemocnice 1, 128 08 Prague 2, Czech Republic
E–mail: zdenek@telicka.cz

EJBI 2012; 8(5):39–42
recieved: September 4, 2012
accepted: October 29, 2012
published: November 22, 2012

1 Introduction

Obesity affects one in four people in the Czech Republic and its incidence is growing worldwide. Conservative treatment does not lead to the desired effect with long-term weight reduction.

One of the most successful methods of treatment with long-lasting effect is the bariatric surgery, which is indicated in specialized centers in patients with severe obesity [1].

In this article we focused on treatment of obesity in diabetic patients by bariatric surgery and tried to evaluate the costs of the surgery and antidiabetics.

This topic is not widely elaborated in the Czech Republic probably due to partly unclear system of payment for the bariatric surgery by an insurance companies and partly due to historical reasons, when the effectiveness of the healthcare system was not so important.

2 Definition of Obesity

Obesity is defined as an excessive storage of energy as a fat. It is stored mainly under the skin which can lead to serious metabolic diseases and also in the abdominal organs it leads to the failure of their functions. From the other point of view, fat stored on the buttocks and thighs are actually protecting the internal organs of the human body and are not associated with metabolic risks. Obesity is always an imbalance in the intake and energy expenditure, which occurs due to many factors, genetic predisposition and social situation of the patient [2].

Obesity with fat stored in the abdominal cavity is called the central obesity and is characteristic in men. In contrast, excessive fat stored in the buttocks and thighs is typical in women and leads to the peripheral obesity. The type of the obesity was in the past years calculated by the ratio of waist to hip circumference and it is called waist to the hip ratio. Abdominal obesity was then defined as a value greater than 0.85 in women and greater than 0.9 in men [3]. Nowadays we use only the waist circumference.

Classification of obesity is done using variable "body mass index" (BMI), which is calculated from weight (kg) and height (m). Classification of individual values shows Table 1.

Table 1: Categories of obesity.

BMI (kg/m^2)	Category of obesity
18.5 – 24.9	normal weight
25 – 29.9	overweight
30 – 34.9	obesity – degree I
35 – 39.9	obesity – degree II
>40	obesity – degree III

Obesity leads to a higher incidence of various chronic diseases, such as:

- Diabetes.
- Gallbladder disease.
- Arthritis.
- Arthrosis.
- Cancer of the ovary, uterus, breast, or colon.

Primarily store fat in the abdominal cavity and the upper half of the chest is associated with diseases, such as:

- Hypertension.
- CHD - coronary heart disease.
- Sudden stroke.
- Insulin resistance.

3 Methods of Treatment of Obesity

Obesity is currently mostly treated conservatively, i.e. pharmacologically, changing diet and lifestyle consulted with clinicians and psychologists. In this article we focus on the treatment of obesity by bariatric surgery, which offers several options with different results. Bariatric surgery is currently one of the most effective methods to help patients reduce their high mass and thus significantly reduce the risk of death or one of the diseases listed above. We can measure success of treatment by this method by reducing patient's weight, reducing a number of antidiabetics using for treatment or disappearance of some of the following diagnoses:

- Type 2 diabetes.
- Hypertension.
- Hypertriglyceridemia.
- Low HDL cholesterol.
- Hypercholesterolemia.

Additionally, in case the patient suffers from diabetes for a long time, the success in treatment of obesity is decreased [4].

3.1 Bariatric Treatment

Today there are implemented various types of bariatric surgery which can reduce patient's stomach volume. Basically, the patient eats smaller portions of food after the intervention and feels satisfied. Malabsorption can be also performed, which reduces the absorption of nutrients and this method is mainly combined with the bariatric surgery.

Current types of bariatric interventions are as follows:

A) Restrictive:

- Adjustable bandage: Stomach is divided into upper and lower parts by a strap with a thin connecting tube which can be adjustably filled up by water and thus decrease the circumference of the strap.
- Sleeve gastrectomy: Performs resection of greater curvature of stomach. This technique is irreversible and it decreases the circumference of the stomach in its whole length.
- Gastric plication: Also called "Laparoscopic greater curvature plication". It reminds sleeve gastrectomy without resection. A portion of greater curvature is decreased by taking in tuck. This type of bariatric surgery is reversible and can be given back to original proportions also using the fibroscopy.

B) Malabsorption:

- Based on the effect of decrease of intake of nutrients. This is done by gastric bypass. The bypass prevents nutrients passing the stomach.

C) Combinations of restriction and malabsorption:

- Biliopancreatic diversion: Combination of gastric resection and creation of 3 shorter bypasses from small intestine.

Depending on the type of bariatric procedure, certain complication may be attributed to it. Most compilations are found in the adjustable bandage:

- Bandage failure.
- Ulcers.
- Narrowing or blockage of the stomach.
- Higher risk of nutritional deficiencies.

These complications may lead to new intervention and surgery corrections [5].

Table 2: Comparison of the impact of bariatric surgery on the remission of diabetes in patients with diabetes.

Type of operation	No. of patients	% Remission of diabetes (HbA1c <4.8% IFCC)	% Compensation of diabetes
Gastric bandage	17	33%	66%
Gastric bypass	6	29%	71%
Sleeve Gastrectomy	7	91%	9%

4 Results

In our research, we studied patients operated in 2007-2009. The total number was 200. Indication for bariatric surgery was as follows:

- Obesity: BMI>40kg/m^2 or obesity associated with Diabetes type 2 or hypertensy: BMI>35kg/m^2 (in rare cases connected with complications the BMI can be also lower than 35kg/m^2).
- Failure during treatment by conservative methods.
- Patient is always cooperative, suitable for long-term follow-up and not suffering from bulimia.

From this group 30 patients suffered from type 2 diabetes mellitus. In this subgroup of 30 patients we focused on monitoring the phenomenon of disappearance of diabetes or compensation. Average age of patients was 51.3 years and BMI 45.87 kg/m^2. Length of follow-up was 1 year. The parameter of diabetes disappearance was decrease of glycosylated hemoglobin (HbA1c) under 4.8% IFCC. We do not use as the parameter of successful surgery end of treatment by antidiabetics because patients are always treated by the metformin during the follow-up.

Table 2 shows that surgical methods differ significantly. These numbers correspond to foreign experience as Buchwald meta-analysis of 130,000 patients, or slightly lower [6].

4.1 Basic Pharmacoeconomical Analysis

We calculated data of 25 patients who were followed for 1 year and we were able to obtain all necessary data (i.e. number of antidiabetics used per day, etc.). Average age in this group of patients was 53.9 years and BMI 47.2 kg/m^2. Although it is known that bariatric surgery is currently the most beneficial in diabetic patients, this treatment is performed far less in the Czech Republic than in the world [2].

We calculated the decrease of average cost of pharmacotherapy by antidiabetics after 6 and 12 month after the surgery. After 6 months the decrease of cost was 3/4 and almost 2/3 after 12 months. This corresponds to the experience that diabetes remission is usually not permanent.

This decrease in daily cost of treatment was reduced nearly 3× in the 6th month of the follow-up. This may be confronted with the prices of bariatric surgery procedures in the Czech Republic. After consultation with bariatric centers we summarized the costs into several groups:

- Payment for preoperative examination before surgery, including sonography, gastrofibroscopy, ECG, spirometry and other outpatient examinations, including blood tests, etc.
- Payment of the surgery, including eventually used staplers, implants, anesthesia, etc.
- Payment of post-operative care, including hospitalization and possibly stay in the ICU.
- Outpatient care within 30 days after surgery - wound care, controls, etc.

However, it is very hard to calculate the cost of the bariatric surgery in the Czech Republic. Insurance companies currently do not take in consideration different costs of partial surgery methods and the payment is in one package for approx. 60 thousands CZK. This number also differs in each of the insurance companies. However, the Czech Society for the Study of Obesity and the Czech Surgical Society stated that valid cost for the particular methods including procedures listed above should be as follows:

- Gastric bandage: 60-70 thousands CZK.
- Gastric plication: 75 thousands CZK.
- Sleeve gastrectomy: 75-80 thousands CZK.
- Gastric bypass: 85-90 thousands CZK
- Biliopancreatic diversion: 110 thousands CZK.

25 CZK = approx. 1 EUR

Table 3: Number of antidiabetics and costs in CZK.

	BMI	Blood sugar (mmol/l)	HbA1c % IFCC	Number of antidiabetics	CZK / day
Before operation	47.2	8.44	6,84	1.2	19.2
6 months	41.1	7.06	5,56	0.9	5.8
12 months	40.1	6.76	5,60	1.0	7.5

25 CZK = approx. 1 EUR

Those prices would be used in near future, but it depends on the discussion between insurance companies and societies. Implementing this real cost for bariatric surgery can lead to better comparison of the particular methods together with the medical results.

5 Conclusion

In the 1-year follow-up we achieved remission of type 2 diabetes especially in patients treated by the Sleeve Gastrectomy method. The positive effect of the surgery appeared in the 6th month of the follow-up and does not changed significantly after the 12th moth of the follow-up. Our results are influenced by the fact that diabetologists do not indicate the bariatric surgery in diabetic patients as it would be optimal. If the indication is positive, it is after long prevalence of the diabetes.

We found out that the cost of the antidiabetics are decreased nearly 3× after 6 month from the surgery. Common follow-up after bariatric surgery is 29 months in the Czech Republic [7] and we are now focusing on evaluation of the cost for a longer period which would give us more accurate results.

6 Discussion

10 years ago, there were 200 to 300 bariatric procedures per year indicated in the Czech Republic. Nowadays, this number has increased to about 1650 per year [7]. Because bariatric surgery provides also excellent results in the treatment of diabetes [8], we assume that the number of diabetic patients indicated for surgery will be increased in near future. Development of this trend certainly helps diabetics prolong their life or significantly reduce their risk of complications or chronic diseases. It brings to patients also significant increase of their life's quality because a successfully treated patient can actively return to the full productive life.

Acknowledgements

The paper has been supported by the SVV-2012-264 513 project of Charles University in Prague.

References

[1] Fried M, Svačina Š, Owen K: Bariatrická chirurgie a diabetes. Trendy v diabetologii. Galén, Prague 2010. (in Czech)

[2] Svačina Š et al.: Klinická dietologie. Grada, Praha 2008; 384 pages. ISBN: 987-80-247-2256-6 (in Czech)

[3] Štejfa, M: Kardiologie; 3rd edition; Grada Prague 2007; 567 pages; ISBN: 8024713853 (in Czech)

[4] Sjostrom CD et al.: Reduction in incidence of diabetes, hypertension and lipid disturbances after intentional weight loss induced by bariatric surgery: the SOS Intervention Study. Obes. Res., 1999, 5: 477-84.

[5] Fried M.: Moderní chirurgické metody léčby obezity, Grada Publishing, Prague 2005; pages: 125. (in Czech)

[6] Buchwald H. et al.: Weight and type 2 diabetes after bariatric surgery: systematic review and meta-analysis. Am. J. Med., 2009 Mar, 122 (3): 248-56.

[7] Kasalický M: CHirurgická léčba obesity; Cited online: http://www.uvn.cz/attachments/1520_Kasalicky-Bariatrie-tiskovka-UVN_Praha.pdf. [25.10.2012] (in Czech)

[8] Sjostrom, C. D. et al.: Reduction in incidence of diabetes, hypertension and lipid disturbances after intentional weight loss induced by bariatric surgery: the SOS Intervention Study. Obes. Res., 1999; 5: 477–84.

Amenability of Czech Medical Reports to Information Extraction

Karel Zvára[1], **Vojtěch Svátek**[2]

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

[2] University of Economics, Prague, Czech Republic

Abstract

Background: Patient's history, family history, diagnoses, medications and other information concerning patient's health and possible future treatment is usually incorporated in free-form narrative reports. Extracting relevant information helps giving the information to caretakers speaking other languages, utilizing modern techniques like reminding caretakers about conflicts with medical guidelines or collecting data for scientific use.
Objectives: The aim of this paper is to summarize the field of information extraction from free-form texts and to show results the author has achieved using simple methods for information extraction.
Methods: The lexical analysis and available Czech versions of medical codebooks were used in the first experiment.
Results: We show that narrative medical reports have a form so different from general texts and cannot be treated as general texts. Additionally available Czech codebooks were found insufficient to be used directly as dictionaries for term recognition.
Conclusions: New dictionaries of Czech medical terms need to be developed. Symbolic techniques have been found effective for recognition of pattern-specific values like Czech birth number or systolic/diastolic blood pressure values.

Keywords

Information extraction from texts, Czech medical reports, lexical analysis

Correspondence to:

Karel Zvára
EuroMISE Centre, Institute of Hygiene and Epidemiology,
First Faculty of Medicine of Charles University in Prague
Address: Katerinská 32, 121 08 Prague 2, CR
E–mail: zvara@euromise.com

EJBI 2012; 8(5):43–47
recieved: September 4, 2012
accepted: October 25, 2012
published: November 22, 2012

1 Introduction

The problem of transforming the text of medical records into structured form has been addressed by medical informatics research for decades. It is well known that the parsimonious writing style of the records, with frequent acronyms and abbreviations, as well as typos caused by time pressure, causes problems to state-of-the-art methods of information extraction from text. Even if partial successes were marked for English as a language with abundance of linguistic tools, nomenclatures, training corpora and, last but not least, stable word order [1], for many other languages the task remains extremely challenging.

The presented research focuses on medical reports written in the Czech language and influenced by the local legislation. The goal was to assess how much relevant information for subsequent transformation to structured form can be revealed via automatic analysis, using simple approaches to information extraction (i.e. those not relying on labelled training corpora).

In Section 2 we provide the taxonomy of information extraction methods, as broader context of our research (including methods planned for future work). In Section 3 we briefly characterize Czech medical records. Section 4 provides an overview of target nomenclatures (i.e. classes of information) and data structures (i.e. containers for information) to which the textual medical records (with special focus on Czech ones) should be converted so as to exhibit full machine-processability.

Section 5, eventually, deals with the application of information extraction on Czech medical records proper; after a brief overview of previous research we present our own research results, divided into three areas: part-of-speech analysis, specific pattern recognition and codebook mapping. Finally, Section 6 wraps up the paper.

2 Information Extraction Methods

Methods of information extraction may be divided into groups according to their subtasks [2], for example:

- Named entity extraction methods. The task of these methods is to find (and annotate) relevant textual properties like names, codes, dates, times, e-mail addresses.
- Co-reference analysis methods. The task of these methods is to find relations among individual words according to morphology of the input text (not specific pre-defined relations).
- Template filling methods. The task of these methods is to fill values found in text into pre-defined template. These methods may be used if there is known target structure (template) to be filled in according to input text.
- Relation extraction methods. These methods are used to extract pre-defined relations among extracted entities.

According to the type of extraction algorithm, information extraction tasks methods may be divided to two groups:

- Manual techniques are based on manually set rules, usually cascaded. This group includes techniques based on regular expressions.
- Trainable techniques are able to improve their ability to extract information from input automatically or under supervision. Trainable techniques usually need some supervision at least in the form of supplying annotations of input text. Trainable techniques include the bootstrapping technique (combining extracting with training). One of bootstrapping methods is "active learning" when annotating expert working with such a system annotates the document that the extraction method is least confident with.

Trainable techniques can be further divided into three groups:

- Symbolic techniques include e.g. Top-Down Induction of Decision Trees (TDIDT) – the "divide and conquer" algorithm (top-down approach) and "separate and conquer" algorithm (bottom-up approach).
- Probabilistic techniques include Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF).
- Other symbolic techniques include e.g. neural networks and support vector machines (SVM).

In the current paper we focused on named entity extraction using manual techniques. Applicability of such methods, in small scale, are a pre-requisite for using automatic techniques and addressing more complex extraction tasks in larger scale.

3 Czech Medical Reports

Czech medical reports are usually narrative reports (free-formed texts) formatted only by spaces, tabs and new lines.

The structure and even the obligation to create and keep medical reports has been incorporated into Czech legislative in 2001 [3] and [4]. The law set requirements for medical reports concerning the content and its form, especially structure.

Czech medical reports are therefore clinical texts with standardized structure. Common form and vocabulary is also determined by common education of physicians, their membership in professional organizations and their own interest on keeping credible documentation not only to enable long term care of the patient but also to defend themselves in judicial affairs.

3.1 Creation of New Medical Reports

New medical reports are usually created from templates and by copying and modifying last report. The reason for creating new reports by copying and modifying last report is economical. Doing "cut and paste" is fast and the physician will not forget to include mandatory information that does not change much during the time - like diagnoses, family history etc. This could lead to serious problems like neglecting changes in diagnoses. Similar problems have been observed also in other countries [5].

3.2 Content from External Systems

Some information comes from external systems in a form that can be simply copied, especially laboratory results. In the case of biochemical laboratory results, rows usually represent individual measurements and columns represent various properties like name of the measured variable, measured value, lower and upper limit. Sometimes simple graphics (created using symbols) is also provided.

3.3 Other Problems

Czech medical reports contain lots of typing error and abbreviations. That is not typical only for Czech medical reports. Individual abbreviations are usually not unambiguous, context is usually needed to decode correctly to find correct meaning. This problem has been also addressed by other authors, see [6].

Table 1: Parts of speech found in narrative reports (total annotations average).

	Annotations (avg per report)	Avg annotations: tot. tokens
Noun	75	30,32 %
Adjective	23	9,3 %
Pronoun	0	0 %
Number (non-digits)	0	0 %
Verb	17	6,87 %
Adverb	3	1,21 %
Preposition	0	0 %
Conjunction	0	0 %
Particle	0	0 %
Interjection	0	0 %

4 Target Structures and Nomenclatures

4.1 Nomenclatures

Target nomenclatures need to by recognized by users and/or their tools (information systems). Internationally and nationally (in the Czech Republic) the ICD (International Classification of Diseases) nomenclature is recognized and commonly used in medical reports. Concerning laboratory reports, SI units are also widely internationally used.

Concerning other codes, international and Czech national use differ greatly. In the Czech Republic, clinical information systems widely use the National Codebook for Laboratory (NČLP, Národní číselník laboratorních položek) which is not one codebook but tens of codebooks, some of them derived or copied from other codebooks (contains e.g. Czech version of International classification of diseases - ICD10).

Internationally, there exist more or less complex nomenclatures, specifically IHTSDO's SNOMED CT (Systematized Nomenclature in Medicine Clinical Terms), Regenstrief Institute's LOINC (Logical Observation Items Names and Codes) and Health Level Seven's Vocabulary.

These internationally recognized nomenclatures are administered by some legal entity and indexed by the National Library of Medicine and its UMLS (Unified Medical Language System). UMLS indexes more than 100 code-books and maps individual coded items to its own concepts, while maintaining network of relations between individual concepts. This way more-or-less accurate mapping among different nomenclatures is made possible.

In addition to UMLS, some nomenclature maintainers are trying to further formalize their nomenclatures to specify an ontology of described field. There are also initiatives that are trying to join partially developed ontology parts into complete ontologies (like OBO Foundry that is aimed at biomedical and biochemical ontologies).

4.2 Structures

Medical reports are non-formalized status documents describing patient's current status, observations and decisions/actions made. There are several influential organizations that concern themselves with formalizing electronic clinical documents, specifically TC 251 of CEN, Health Level Seven, ASTM American Society for for Testing and Materials) and openEHR Foundation.

Health Level Seven develops the CDA (Clinical Document Architecture) specification. It is designed to formalize administrative information, to annotate medical report on the level of report parts but allows to formalize individual clinical observations. Health Level Seven standards are usually developed using top-down approach (from general to specialized), the development is slow but the result is usually robust.

ASTM developed the Continuity of Care Record (CCR) standard. It represents just current state of the patient, so it is a state-report. Being developed not from the top but according to requests from users, CCR is more practical but less robust than CDA. ASTM and Health Level Seven together developed technical implementation of CCR using CDA. The result (CDA document containing CCR) is called Continuity of Care Document (CCD).

From the European perspective, the most important standardization of formalized electronic health record came from CEN, the EU normalization institution. CEN developed EN 13606 which has been adopted also by ISO. EN 13606 is usually referred to as "EHRcom". EHRcom specifies general way to formalize information commonly found in medical reports. It uses SNOMED CT, LOINC and other internationally used classification systems.

There are also projects which aim to standardize (minimal) content of electronic health record.

The epSOS[1] projects concerns also with a kind of minimal electronic documentation needed for urgent care of the patient. epSOS has published a specification of Patient Summary (PS) which is also mapped to existing european EHR standard EHRcom (EN 13606).

[1] European Patients Smart Open Services

Table 2: Delimited numbers recognition results.

	Found	Min. found	Max. found	Average
Blood pressure: SBP/DBP	434	0	12	1,62
Personal identification number	77	0	1	0,29
Not identified	268	0	6	1

5 Automated Analysis of Czech Medical Reports

First studies on automated (information-extraction-based) analysis of Czech medical reports were published in [7] and [8].

In the study [8] the regular analysis was used for information extraction. The paper [7] concluded that lexical analysis cannot be used because Czech medical reports are usually not made from whole sentences and the punctuation is almost not used.

The study [8] continued in the study published in [7] and enhanced regular analysis with some linguistic analysis. There were not used any codebooks and slightly better results were achieved in [8] than in [7].

We have studied the possibility of lexical analysis, recognizing specific patterns (like Czech personal identifiers or systolic/diastolic blood prossure) and using available code-books before. Partial results where published in [9].

5.1 Lexical (Part-of-Speech) Analysis

In order to analyze the distribution of differents parts of speech in the records, we reused the Czech iSpell dictionary from Petr Kolář, which was originally designed for spell-checking. The original version can be used for part-of-speech (PoS) tagging with just minor additions. Further, more complicated, addition would allow detection of inflection and gender but that has not been done because of poor results achieved from PoS tagging.

The Czech iSpell dictionary contains 260.679 basic words expanded to 4.624.350 words (some with exactly same expression but with different gender or part-of-speech tag) using affix rules. High number of annotations is determined by multiple annotations of recognized words.

Processing 268 narrative reports with a total of 66.286 tokens gave the results shown in Table 1.

5.2 Recognizing Specific Patterns

A relatively easy (though not trivial) task for information extraction consists in recognition of sequences of numerals with specific meaning. We focused on two common types of information, blood pressure values and the personal identification number of the patient. Specific combined numeric patterns were recognized with symbolic rule-based methods (similar to regular expressions). Different meanings were distinguished by fixed rules. In the case of blood pressure it was meaningfulness range of values, relation between parts of a pattern. In the case of personal identification number, the test of syntax correctness (lengths of parts) and meaningfulness has been used (personal identification number contains information on date of birth, gender and office that has allocated the number).

The Table 2 shows results of delimited numbers recognition.

There were no identified recalls, mostly because the rules for recognizing blood pressure and personal identification numbers were defined as strict. Both recognizers were defined for two decimals separated by slash with these properties:

- **blood pressure:** first number is greater than second, both numbers are positive, first number is lower than 500;
- **patient identificator:** rules for validation check of Czech personal identifiers were used (valid date and sex coded in the first number, identifiers corresponding to dates newer than January 1st 1954 are also checked for checksum).

5.3 Using Available Code-Books

Results of recognizing code-book terms have been published in [9]. Recognition of SNOMED CT and ICD10 terms has been totally unsuccessful. In case of SNOMED CT it has been expected because Czech version of SNOMED CT is not available. ICD10 has been used in the Czech version (part of NČLP code-books) but has been totally unsuccessful partly because specific expressed diagnoses have already been coded with ICD10 and partly because Czech names/descriptions of ICD10 terms are log and contain a lot of abbreviations.

The only successful coding system has been MeSH[2] in the Czech version. Even in the case of MeSH, we were able to recognize less than two terms per narrative report in average.

6 Conclusions

We can briefly summarize the main findings related to the three types of text analysis employed.

The **lexical analysis** is not a solution to information extraction from narrative reports written in Czech. The

[2]Medical Subject Headings

main reason is that narrative reports written in Czech are not regular sentences. This is manifested by the distribution of parts of speech, which clearly deviates from the distribution in contiguous text.

The main lesson learned from the lexical analysis part is that attention must be paid to typing errors and abbreviations. Both tasks should be solved alongside text extraction because abbreviations and typing errors are very often ambiguous. Therefore their translation to correct form needs context from other parts of the narrative report.

Symbolic techniques like rule-based filters or recognizing agents are good tool to recognize some specific numeric values. Such techniques can be effectively used to recognize blood pressure values and patient identification.

Looking up from standard **code-books** seems inefficient since most complete clinical code-books (especially SNOMED CT) are not available in the Czech language. Therefore some other code-book must be found, created or existing code-book translated.

Acknowledgements

This work has been supported by the specific research project no. 264513 "Semantic Interoperability in Biomedicine and Health Care", Charles University in Prague.

References

[1] Garcia-Remesal M., Maojo V., Billhardt H., Crespo J., Integration of Relational and Textual Biomedical Sources, Methods Inf Med, 2010

[2] Labský M., PhD thesis: Information Extraction from Websites Using Extraction Ontologies, Vysoká škola ekonomická v Praze, Praha, 2009 (Czech)

[3] Žďárek R., Vedení zdravotnické dokumentace a její náležitosti, Zdravotnické noviny, 3.6.2009 (Czech)

[4] Dostál O., Šárek M., Support for Electronic Health Records in Czech Law, European Journal for Biomedical Informatics, 2012

[5] Hammond K., Helbig S., Benson C., Brathwaite-Sketoe B., Are Electronic Medical Records Trustworthy? Observations on Copying, Pasting and Duplication. AMIA Annual Symposium Proceedings, 2003; 269-273

[6] Tsung O. Cheng, Letters to Editor; in: Medical abbreviations in Journal of the Royal Society of Medicine, Volume 97, 2004

[7] Semecký J., Zvárová J.(školitelka), Multimediální elektronický záznam o nemocném v kardiologii, Matematicko-fyzikální fakulta UK, Praha, 2001 (Czech)

[8] Smatana P., Paralič J. (školitel), Spracovanie lekárskych správ pre účely analýzy a dolovania v textoch, Technická univerzita v Košiciach, Košice, 2005 (Czech)

[9] Zvára K., Kašpar V., Identifikace jednotek a dalších termínů v českých lékařských zprávách, European Journal for Biomedical Informatics, 2010 (Czech)

