EJBI 2007 ISSN 1801 - 5603

European Journal for Biomedical Informatics

Volume 3 (2007)

www.ejbi.eu

EuroMISE s.r.o.

Aims and Scope

The European Journal for Biomedical Informatics reacts on the great European need to share the information in the multilingual and multicultural European area. The journal publishes peer-reviewed papers in English and other European languages simultaneously. This opens new possibilities for faster transfer of scientific-research pieces of knowledge to large international community of biomedical researchers, physicians, other health personnel and citizens.

The generally accepted translations of the English version of the paper to the official European languages and other 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

Cooperating journals (language)

Methods of Information in Medicine	(English)
Lékař a technika	(Czech)

Editors and Management

Editorial Board

Publisher

EuroMISE s.r.o.
Paprskova 330/15
CZ-14000 Praha 4
Czech Republic
EU VAT ID: Cz25666011

Office

EuroMISE s.r.o.
Paprskova 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

MANUSCRIPT SUBMISSIONS

European Journal of Biomedical Informatics (EJBI) is an international, peer-reviewed journal that publishes papers in the broad field of biomedical informatics. Manuscripts accepted for the electronic publication in EJBI are original contributions, reviews, brief reports, special communications, commentaries, and many other categories of papers. Due to special multilingual features of EJBI, these types of papers are published in English, but national language versions can be published simultaneously (see below).

Manuscripts should be sent electronically to the e-mail address: manuscripts@ejbi.org. Each manuscript should be submitted in MS Word, plain-test, HTML, TeX or LaTeX format and accompanied by image attachments and

- Identification Form: the name of the corresponding author with his contact address, phone number, fax number and e-mail address.
- Submission Requirement: the statement that the submitted paper has not been published in, nor has been submitted to, any other journal.
- Authorship Form: the covering letter signed by corresponding author, by which the author approves publication of the paper in the European Journal for Biomedical Informatics on behalf of all authors.
- Transfer of Copyright: All authors of the manuscript must have agreed to its publication and are responsible for its content and must also have agreed that the corresponding author has the authority to act on their behalf in all matters pertaining to publication of the manuscript. The corresponding author is responsible for informing the co-authors of the manuscript status throughout the submission, review, and publication process.

Submission Requirement, Authorship Form and Transfer of Copyright should be sent by fax +420 241 712 990 and by the surface mail to EuroMISE s.r.o., EJBI Editorial Office, Paprskova 330/15, 140 00 Prague 4, Czech Republic simultaneously.

Authors may supply a PDF file of the manuscript additionally. Submission of the PDF file only is not sufficient.

PUBLISHER

The European Journal of Biomedical Informatics is published by EuroMISE s.r.o., Prague, Czech Republic (tax identification number: CZ25666011). Papers published in the Internet version of EJBI may be republished in printed digest of EJBI.

LANGUAGE

All manuscripts should be written in English in an easily readable style. English version of the manuscript will be peer-reviewed. However, it is possible to submit another language version of the same paper additionally. The Internet version of the EJBI makes available all language versions submitted. It is also possible to publish the English version of the paper that has been already published in the national journal, when the agreement of nationa l publisher is obtained (copyright agreement). The English version follows a standard peer-review procedure.

ELECTRONIC FORM OF A MANUSCRIPT

Recommended formats of text files are: .DOC (Word version 6.0, 7.0, 97, 2000 or 2003), .RTF (Rich Text Format). The name of the text file should be identical with the first author's surname, e.g. NOVAK.DOC. Names of attached images, drawings or tables (i.e. figures and tables of the manuscript) should be composed of author's surname and figure or table number, e.g. NOVAKTAB3.BMP. Manuscripts along with attached files should be sent by e-mail to address manuscripts@ejbi.org.

ELECTRONIC IMAGE DOCUMENTATION

Figures and tables (bitmaps) should be submitted in BMP, GIF, TIF, PNG, EPS or JPEG format. JPEG files should have best compressive rate (10-20) for a good quality. All images (although included in the manuscript file) must be also submitted as separate files because the resolution of embedded images is often insufficient. Scanned images should have the resolution of at least 600x600dpi. Submissions that do not meet the Instructions for Authors will be returned.

ABBREVIATIOBS AND NOMENCLATURE

Generally known abbreviations do not need to be explained. Abbreviations for symbols and expressions for terms should be spelled and they should be comprehensibly explained in brackets. Every such abbreviation should be explained only after its first occurrence in the text.

Measurements of length, height, weight and volume should be reported in metric units or their decimal multiples. All haematological and clinical chemistry measurements should be reported in the metric system in term of International System of Units (SI). Chemical substances should be described by their own systematic name or expression, medicines by general names. Commercial names of chemicals, medicines or technical innovations may be used after they have been defined by their scientific names. Radionuclides (radioisotopes) will be symbolised by the atomic number.

REVISED MANUSCRIPT SUBMISSION

When revision of a manuscript is requested, authors should return the revised version of their manuscript as soon as possible. The prompt action may ensure fast publication if a paper is finally accepted for publication in EJBI.

FINAL PROOFREADING

The Publisher will send the accepted paper to its author for final proofreading in the PDF format. The author may then correct printing errors only. No other changes or additions will be accepted. Author should send corrected and signed paper back to the Publisher by fax to +420 241 712 990 or by surface mail to: EuroMISE s.r.o., EJBI Editorial Office, Paprskova 330/15, 140 00 Prague, Czech Republic.

ORGANIZATION OF THE MANUSCRIPT

Title page. The first (title) page should contain the title of the paper, names and workplaces of all authors. Individual workplaces are necessary to be graphically differentiated (preferably by numeral as the upper index).

Abstracts and keywords. At the beginning the author puts an abstract and keywords. The abstract should be in the extent of 250-300 words. There should be 4 to 7 keywords, according to author's consideration, preferably from MeSH index.

Main text of the paper. General rules for writing manuscripts recommend use of simple and declarative sentences; avoid long sentences, in which meaning may be lost by complicated construction. All acronyms and abbreviations should be explained when they first appear in the text. The main text of the paper should follow the style of selected type of paper.

Acknowledgement. Acknowledgements, if any, should be given at the end of the paper, before bibliographic references.

References. References should be cited in the text by their index number according to the order of appearance in the manuscript. Each reference should be marked by its index number in square bracket corresponding to bibliography section. It is possible to include references to dissertation works and technical reports. It is obligatory to include information sufficient to look up referenced text.

Examples of references in bibliography section:

[1] Knaup P., Ammenwerth E., Brandner R., Brigl B., Fischer G., Garde S., Lang E., PilgramR., Ruderich F., Singer R., Wolff A. C., Haux R., Kulikowski C.: Towards Clinical Bioinformatics: Advancing Genomic Medicine with Informatics Methods and Tools. Methods Inf Med 2004; 43, pp. 302-307

[2] Blobel B., Pharow P.: A Model-Driven Approach for the German Health Telematics Architectural Framework and the Related Security Infrastructure. In: Connecting Medical Informatics and Bio-Informatics. Proceedings of MIE2005 (Eds. R. Engelbrecht, A. Geissbuhler, C. Lovis, G. Mihalas), Vol. 116, Amsterdam, IOS Press, 2005, pp. 391-396

[3] http://www.infobiomed.org/

Tables and Figures. Authors should use tables only to achieve concise presentation, or where the information cannot be given satisfactory in another way. Tables should be numbered consecutively using Arabic numerals and should be referred to in the text by numbers. Each table should have an explanatory caption that should be as concise as possible. Figures should be clear, easy to read and of a good quality. Styles and fonts should match those in the main body of the paper. All figures must be mentioned in the text in consecutive order and should be numbered with Arabic numerals. Authors should indicate precisely in the main text where tables and figures should be inserted, if these elements are given only separately or at the end in the original version of the manuscript.

Content

English version

Croatian version

Polish version

EJBI 2007 ISSN 1801 - 5603

European Journal for Biomedical Informatics

Volume 3 (2007), Issue 1

English version

Cooperating journal - Methods of Information in Medicine
Editor-in-Chief: R. Haux (Germany)
Publisher: Schattauer
www.methods-online.com

EuroMISE s.r.o.

Content

English version

Editorial

Jana Zvárová
Editor-in-Chief

Biomedical informatics is a burgeoning field, with important applications and implications throughout the biomedical world and healthcare delivery. The European Journal of Biomedical Informatics (EJBI) is reacting on the great European need to share the information in the multilingual and multicultural European area.

EJBI opens for the field of biomedical informatics a new model of electronic publishing. EJBI is publishing accepted peer-reviewed papers in English and other languages simultaneously. This opens new possibilities for faster transfer of scientific-research pieces of knowledge of many European countries to a large international community of biomedical researchers, physicians, other health personnel and citizens. Moreover, the journal now enables to make results of scientific-research work and practical experiences of foreign specialists accessible to wider health public in a more comprehensible way in each European country.

The aim of the editorial board is to reach the highest scientific level of the journal and show the best practices of biomedical informatics applications to wide readership. The European editorial board is composed from outstanding specialists in the field of biomedical informatics. We believe that their activities for EJBI will contribute to propagation of journal's good credit. The editorial board also presumes that presentation of English versions of scientific papers with their professional translations to other languages will significantly contribute to unification of applied scientific terminology.

EJBI is now a Schattauer related journal. You may find more information in Editorial of Methods of Information in Medicine, Issue 2 2008.

Information and Communication Technology in Family Practice in Croatia

Josipa Kern[1], **Ozren Polašek**[1]
[1]Andrija Stampar School of Public Health, Zagreb University Medical School, Croatia

Summary: Family practice was the first privatized part of the health care system in Croatia. Recently, 84 % of family practices are privatized. Other 16% of them left within health centers. In order to assess current status of use of the information and communication technology (ICT) in family practice, physicians specializing in the family medicine, being also the postgraduate students of the University of Zagreb, School of Medicine in the year 2004/05, have been surveyed. The sample consisted of 159 physicians. Some kind of information system (IS) exists in 62 % of family practices under lease, 42 % within health centers, and 91 % of completely private offices. Having IS in their offices, physicians use it primarily for administration: reporting, prescriptions, sick leaves, referrals and for billing. Usage of ICT in medical work, electronic medical records, was found at about 50 % of physicians. Doing research based on information stored in e-format, is conducting by 31 % of physicians. Based on e-information 35 % of physicians evaluate their work. About using e-sources of medical knowledge (bibliographic data bases, e-journals): 86 % physicians from private offices, 79% physicians working within health centers, 60 % physicians from offices under lease. Satisfaction with current information system is small (19 %). Their information needs are covered partially. 32 % of physicians said they can get enough information by using their information system. 40 % of physicians feel that ICT gave more efficiency to their work, and 25 % of physicians feel their patients are more satisfied since the IS was put into function. Considering different types of practices – young physicians working in health centers are less satisfied with their ISs than the other two groups of physicians coming from private practice. Data security was practiced by using a password, physical protection, and by daily archiving of data.

Keywords: information system, family practice, electronic medical record, Internet, computer security, bibliographic databases, consumer satisfaction

1. Introduction

Family medicine is one of the main pillars of population health care in Croatia. Its role is very important in both, prevention and cure, for a predefined population. Willing to achieve, maintain and improve the quality of health care, family physicians need to record patients' data, do analyses, and have access to relevant information and knowledge relating to a problem. Considering the fact that data operability can be fully achieved only by use of information and communication technology (ICT), it is time indeed for such a technology to domesticate appropriately in practice of family medicine. Appropriateness of ICT in medicine and health means the application capable to give information needed to a physician or a nurse, but also to a wider community (if a family physician is obliged to give such an information, e.g. to public health, health insurance, secondary and tertiary health care, Ministry of Health etc.). Of course, all measures for data security and confidentiality should be applied. A family physician should expect to have a useful and user-friendly ICT application enabling to give an information promptly. The ICT application should be modular and scalable, able to be upgraded and improved in accordance with needs of users. It should include international standards, and be evaluated by both, health and ICT professionals [1].

Health professionals should be educated in using the ICT application, they should be aware of possibilities and limitations of such a technology and motivated for further development and improvements of their ICT-based system.

As a transitional country, Croatia started with privatization in the health care system in 1994. Family medicine practice was one of the first privatized parts of the health care system in Croatia. Recently, 84 % of family medicine practices are privatized. Other 16% left within health centers: 95 % of privatized family practices are under lease (in offices and equipment owned by health centers; family practice pays for leasing) and 5 % as completely private (in their own office, with their own equipment). Family practices have contracts with the Croatian Institute for Health Insurance (CIHI). According to the contract, the CIHI covers health care services by paying for patients pre-registered in the family practice (by payment "per capita"). There are some differences between different family practices in their obligations and their autonomy:

- practices within a health center (offices and equipment owned by a health center, the contract between the CIHI and health centers, health care providers are paid by health centers, some additionally paid health services the family practice can provide are not permitted; no autonomy at all),
- practices under lease (offices and equipment owned by a health center are leased by a family physician, the contract between the CIHI and a family physician, all financials are up to him/her, some additionally paid health services the family practice can provide are not permitted; limited autonomy),
- completely private family practice (offices and equipment owned by a family physician, the contract between CIHI and a family physician, all financials are up to him/her, some additionally paid health services the family practice can provide are permitted; autonomy).

Having privatized their practice, many family physicians started with bringing ICT-solutions to their offices.

It is much less the case of family practices left within health centers, where family physicians have no opportunities for deciding what kind of ICT-solutions to buy. The questions are: what kind of ICT-solutions they have, are they satisfied with them, are they skilful enough in using ICT, and are there any differences between different status of family medical practices considering the kind of privatization and those not being privatized.

2. ICT in the Croatian health care system

The Croatian health care system started with the ICT application in late 60's [2]. Of course, applications were appropriate to the level of technological development and possibilities the Croatia has had then. Primary health care units organized in health centers, and hospitals, started in 1970 with IT-based administration in these institutions [3], [4] . The latest efforts in that direction within the health care system reform are attributed to the Ministry of Health and its document "Strategy and plan of the reform of the health care system and health insurance of the Republic Croatia" [5]. Considering the role of primary health care in the health care system, a special treatment has been given to development of ICT based solution for primary health care [6].

3. Current status of IS in family medicine practice

3.1. The basis for analysis of current status

In the analysis of the current status of informatization of family medicine practice (ICT skillfulness and using ICT, some medical informatics knowledge and readiness of medical doctors to accept and improve information systems in their offices), physicians specializing in family medicine, being also the postgraduate students at the University of Zagreb, Medical School, in the school-year 2004/05, have been surveyed. The survey was anonymous and it was taken after the first, introductory medical informatics class.

The survey consisted of three modules: (a) questions about the physician (age, years of working in practice, sort of practice – office under lease, office in a health centre, completely private office), (b) questions on using of ICT, and, finally (c) on the information system installed in their office. This last module was taken only for physicians with computers in their offices and a certain ICT-based application installed on it.

3.2. Results

Table 1 shows a sample of primary health care physicians, postgraduate students in the school year 2004/2005 during the specialization in family medicine. All of them were working for several years as primary health care physicians. Most of the 159 physicians participating in the survey had their office under lease (117 of them), 28 of them had their offices within a health centre, and 14 had the completely private office. It can be perceived that practices within health centers are less computerized. Medical doctors working there are younger. They are (on average) 8 years younger than others, with 10 years shorter working period in practice. Completely private offices have been computerized in most cases (91 %).

Patterns of use of ICT are equal for all three groups of medical doctors: using e-mail was on the top, keeping on reading health portals was on the second place. Then searching medical literature through PubMed, Current Contents etc., and electronic journals like Croatian Medical Journal, British Medical Journal and alike follows. Differences in use of electronic sources of medical knowledge within physicians from the three types of offices are statistically significant (Chi-square test p=0.043).

The term "using a source of medical knowledge" means that physicians state they use at least one of the following sources: read e-journals, search through PubMed or similar publications, and keep on reading health portals. According to this definition the first place in using e-sources of medical knowledge belongs to physicians from private offices, 86 %. The second place belongs to physicians working within health centers, 79 %, while physicians coming from offices under lease use most rarely such sources of medical knowledge, 60 %. Figure 1 shows frequency of use of different electronic sources of medical knowledge based on types of practices.

Patterns of use of ICT are equal for all three groups of medical doctors: using e-mail was on the top, keeping on reading health portals was on the second place. Then searching medical literature through PubMed, Current Contents etc., and electronic journals like Croatian Medical Journal, British Medical Journal and alike follows. Differences in use of electronic sources of medical knowledge within physicians from the three types of offices are statistically significant (Chi-square test p=0.043).

Table 1. Medical doctors specializing in family medicine.

	Age median (interquartile range)	Years of working median (interquartile range)	Having a computer in his/her office (%)
Office under lease	43 (41-46)	15 (12-19)	62
Office within a health centre	35 (32-37)	5 (2-7)	42
Completely private office	43 (41-44)	15 (7-19)	91

The term "using a source of medical knowledge" means that physicians state they use at least one of the following sources: read e-journals, search through PubMed or similar publications, and keep on reading health portals.

According to this definition the first place in using e-sources of medical knowledge belongs to physicians from private offices, 86 %. The second place belongs to physicians working within health centers, 79 %, while physicians coming from offices under lease use most rarely such sources of medical knowledge, 60 %. Figure 1 shows frequency of use of different electronic sources of medical knowledge based on types of practices.

Having some kind of an information system installed in computers in their offices, physicians use it primarily for administration: reporting, prescriptions, sick leaves, referrals and for billing (Figure 2). Usage of ICT in medical work (electronic medical records) was found at just above average 50 % of physicians. But, as a rule, they have also "papers". That is doubling documentation as a result of a current legislative in the Republic of Croatia. Doing research based on information stored in e-format in their information system, is conducting just by 32 % of physicians having an electronic medical record in their practice. In spite of prevailing of physicians from private offices, there are no statistically significant differences. Using an information system in their practices, 35 % of 97 physicians evaluate their work (whatever it assumes). Also, there is no statistically significant difference between types of practices.

Satisfaction with the current status of computerization of primary health care is more than modest, only 19%, considering the whole sample of physicians having some ICT application in their practice. In the same time, information needs are covered just partially. 32 % of physicians said they can get enough information by using their information system. 40% of physicians feel that ICT gave more efficiency to their work, and 25 % of physicians feel their patients are more satisfied since the information system was put to function.

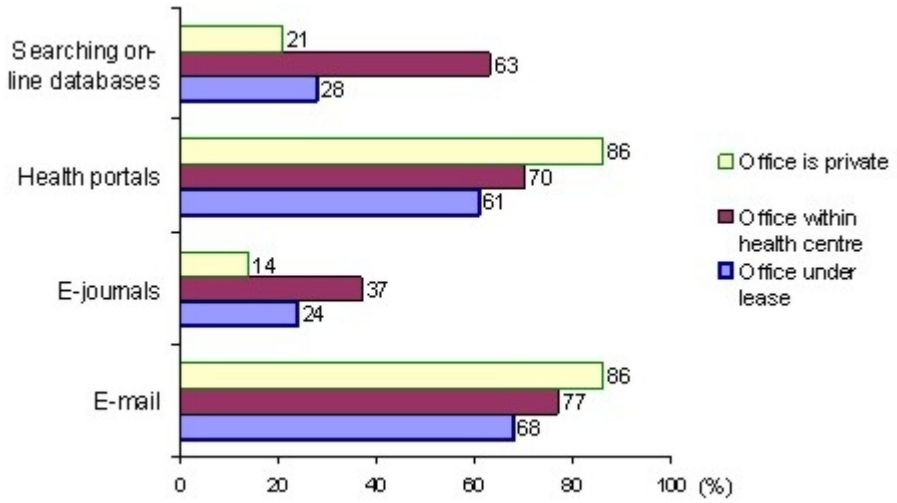


Figure 1. Frequency of using ICT by physicians specializing in family medicine.

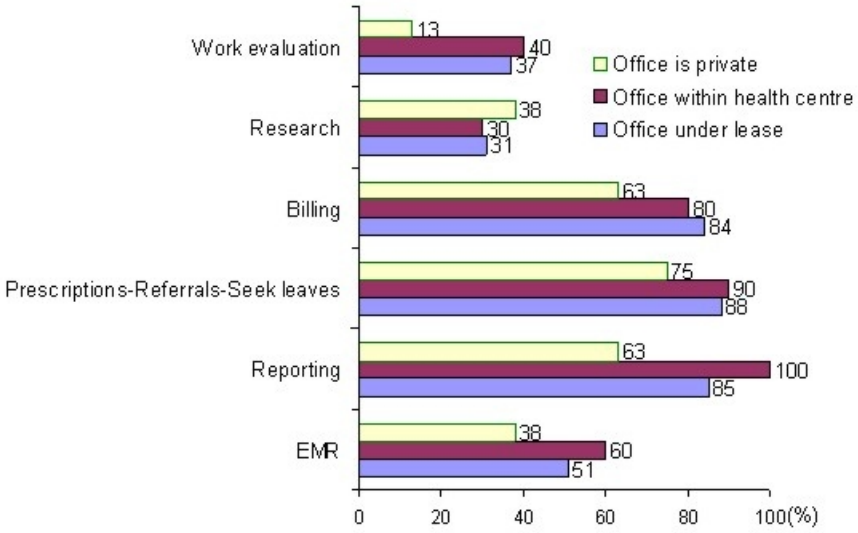


Figure 2. What for the ICT applications in family medicine practice are used.

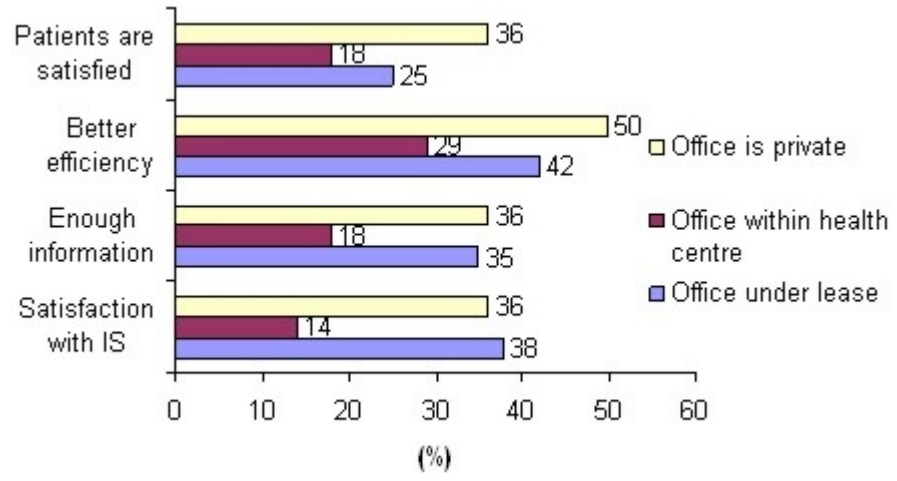


Figure 3. Satisfaction with ICT-based appliction – physicians' estimate.

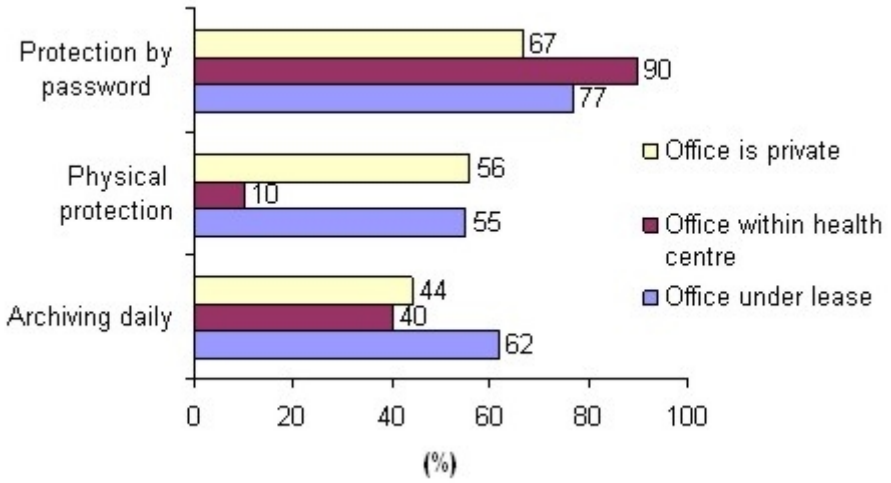


Figure 4. Data security in the information systems.

Considering different types of practices (Figure 3) there are some differences. Young physicians working in their offices within health centers are globally less satisfied with their information system then the other two groups of physicians coming from private practice (under lease or completely private offices).

Quality of data security in the information system was estimated through a) using a password, b) physical protection, and c) by archiving of data daily. There are statistically significant differences between types of practice considering physical protection (Chi-square test p=0.028). Daily archiving and protection by a password is nearly the same.

4. Discussion

Croatian physicians in family medicine practice use ICT less than those in developed countries [7]. Practicing family medicine physicians use ICT applications primarily for reporting and administration, and it is the same as in developed countries [7]. Information needed for health care and curing of a particular patient, do not have priority in informatization of the family practice. It is the same with information being supposed to improve their work through evaluation and research. For all this the basis is, of course, electronic medical record. Nowadays, the Croatian law requires the classical paper record. As a consequence, with electronic medical records, physicians must have also corresponding conventional paper records. It means double documentation. Doubling the work does not encourage physicians to use and improve electronic records. Therefore these records are mostly unstructured, mostly free texts what makes difficulties in any analysis.

Satisfaction with the current information system is pretty small. Only physicians in privately own practices are satisfied with their efficiency with using the information system. In the same time they positively evaluate patients' satisfaction with practices using some kind of an information system.

Security and safety of information systems can be discussed from different aspects. Avery and co-authors are saying on coordinating the most important elements of system safety from the point of view of clinical work (warnings whenever it needs, safety in repeated prescribing, etc.) [8]. Honyeman and co-authors discuss potential access to personal electronic medical records by patients themselves [9]. In age of new technologies, like the Grid-technology enabling to build and access distributed electronic records, it is necessary to consider more than ever security, safety and ethics [10], [11]. Safety of information in the family medicine practices information systems in Croatia is not good enough. Based on physicians' evaluation, less than 50% of physicians protect their e-information to a satisfactory level. However, it is a very rough estimate.

Considering the results of this survey, physicians working in family medicine practices are not skilful enough (or engaged) in using the information and communication technology in their professional work. Although they found Internet very useful (like in developed countries) and use e-mail and "surf" the Internet, they do not read enough professional electronic journals and secondary publications either. However, it is similar as in developed countries [12]. There is a little increase among younger physicians employed in (yet not-privatized) practices within health centers. It can be recognized as a hope that ICT using is getting better.

On the other side some physicians invest a lot in improving their work and relationship to the patients using the contemporary technology. There are several positive examples of taking information technology courses. These courses are organized by Medical Schools, Croatian Chamber of Physicians, but also by certain companies acting as vendors for such a technology. The Web technology is used for advertising practices. In the same time there are family physicians trying to appear "friendly" to patients, informing them on health and teaching them how to be healthy [13].

Acknowledgement

The authors would like to thank to Professor Milica Katić, professor of family medicine at the School of Medicine, University of Zagreb, who helped us to explain specificity of family practice organization in the Croatian health care system.

References

[1] Končar M., Gvozdanović D.: Primary Healthcare Information System – the Cornerstone for the Next Generation Healthcare Sector in Republic of Croatia. Int J Med Inform 2006; 75, pp.306-314.

[2] Kern J., Strnad M.: Informatics in the Croatian Health Care System. Acta Med Croatica 2005; 59, pp. 161-168. [in Croatian]

[3] Golec B, Krajačić S.: Project of Automated Data Processing for Outpatient Health Organization. Health Centre Remetinec-Zagreb. Zagreb, Centre for economic development of the city of Zagreb, 1980. [in Croatian].
[4] Rosandić D.: Report on Introducing Computerization in the General Hospital "Dr. J.Kajfeš" – Zagreb, Zagreb, Commision for computerization of health care system 1975. [in Croatian].
[5] Reform of Healthcare System: Strategy and Plan of Reform of Health and Health Insurance System in Croatia. Zagreb, Ministry of Health 2002. [in Croatian]
[6] Stevanović R., Stanić A., Varga S.: Information System in Primary Health Care, Acta Med Croatica 2005; 59, pp. 209-212. [in Croatian].
[7] Western M. C., Dwan K. M., Western J. S., Makkai T., Del Mar C.: Computerisation in Australian General Practice. Aust Fam Physician 2003;32, pp. 180-5.
[8] Avery A.J.,.Savelyich B.S, Sheikh A., Cantrill J., Morris C.J., Fernando B., Bainbridge M., Horsfield P., Teasdale S.: Identifying and Establishing Consensus on the Most Important Safety Features of GP Computer Systems: e-Delphi Study, Inform Prim Care 13 (2005) 3-12.
[9] Honeyman A., Cox B., Fisher B.:.Potential Impacts of Patient Access to Their Electronic Care Records, Inform Prim Care 2005; 13, pp. 55-60.
[10] Kalra D., Singleton P., Milan J., Mackay J., Detmer D., Rector A., Ingram D.: Security and Confidentiality Approach for the Clinical E-Science Framework (CLEF), Methods Inf Med 2005; 44, pp. 193-197.
[11] Claerhout B., De Moor G. J.: Privacy Protection for HealthGrid Applications. Methods Inf Med 2005; 44, pp. 140-143.
[12] Bennett N. L., Casebeer L. L., Kristofco R., Collins B. C.: Family Physicians' Information Seeking Behaviors: a Survey Comparison with Other Specialties. BMC Med Inform Decis Mak 2005; 5, p. 9.
[13] http://www.ordinacije-lazic.hr/. Accessed on August 14, 2006.

Classification of Body Surface Potential Maps: A Comparison of Isointegral Measurements in the Diagnosis of Old Myocardial Infarction

Dewar D. Finlay[1], **Chris D. Nugent**[1], **Haiying Wang**[1], **Huiru Zheng**[1], **Mark P. Donnelly**[1], **Paul J. McCullagh**[1]
[1]University of Ulster, School of Computing & Mathematics, Northern Ireland, UK

Summary: The electrocardiogram (ECG) is one of the most common ways to record, in an non-invasive manner, a patient's cardiac activity. Once recorded the information can be pre-processed and subsequently analyzed to assess if the patient is suffering from any forms of cardiac abnormality which may require clinical intervention. In the current study we investigate ways in which more can be obtained from the ECG through analysis of the diagnostic properties of body surface potential maps (BSPM). A set of 192 lead BSPMs recorded from a mixture of 116 normal and abnormal subjects (59 normal vs 57 old myocardial infarction) were analyzed. For each patient, diagnostic features were obtained by calculating isointegral measurements from the QRS, STT, and entire QRST segments. These isointegrals provide a measure of the mean distribution of potential during ventricular depolarization, repolarisation, and a combination of both, respectively. For each isointegral type, 192 discrete measurements, and hence 192 features, were obtained; these correspond with the 192 leads recorded. Subsequent to this a signal-to-noise ratio-based feature ranking methodology was applied to select subsets of the best three, six and ten measurements (features) from the 192 available for each isointegral. These subsets of features were then applied to four different classifiers Naïve Bayes (NB), support vector machine (SVM), multi-layer perceptron (MLP) and random forest (RF) and in each application ten-fold cross validation was employed. It was found that when using the subsets of features obtained from the STT or QRST isointegrals, classification results in excess of 80% were attainable. This was in contrast to the results obtained using the QRS isointegral features where poorer performance (between 62.9% and 74.1%) was observed. The results from this study have illustrated that, for the studied dataset, the mean distribution of potentials during ventricular depolarization, and during ventricular repolarization and depolarization combined possessed greater diagnostic information. Overall it was concluded that this approach to BSPM analysis does provide a useful means for illustrating the usefulness of various features in diagnostic classification.

Keywords: electrocardiogram, body surface potential map, myocardial infarction, feature selection

1. Introduction

The 12-lead ECG is used to detect many cardiac abnormalities which include electrical conduction defects and myocardial infarction (MI) [1]. The accuracy of the 12 lead ECG has, however, been called into question [2], and this is based largely on the appreciation that the necessary diagnostic information may not be captured by the recording sites that make up this format. To counter this, investigators have looked to alternative recording techniques to capture more useful information. The most extreme example of this is the BSPM. In this approach ECG information is recorded from as many as 200 sites on the torso [3], [4]. This level of spatial sampling provides a much more comprehensive picture of cardiac activity as effectively all ECG information, as projected onto the body's surface, is captured. Although superior in terms of their diagnostic yield [5], BSPMs are not widely used in clinical practice. This is because the large number of recording channels makes the acquisition process more cumbersome and BSPMs have not experienced widespread utility outside of the research laboratory. Despite this there is much to be gained from the study of BSPM data, as in effect a more comprehensive picture of cardiac activity is being studied. In this paper we detail the investigation of the use of BSPM data in the classification of old MI. In particular we focus on locating the most useful diagnostic information in BSPMs and we use this to address the classification problem.

2. Methods

In this study we analyze a set of 192 lead BSPMs that were recorded from a mixture of patients that were previously diagnosed as being normal or having old MI. The clinical data and experimental procedures are described as follows.

Clinical Data

The 192 lead BSPMs were recorded from a group of 116 subjects. This was made up of 57 subjects with MI at various locations and 59 normal subjects. The breakdown of this dataset is listed in Table 1. The recording procedure has previously been described in [6], [7], [8] and is summarized as follows. On each subject the electrodes were positioned by placing 16 columns of 12 electrodes on the torso. These columns were equi-spaced around the thoracic circumference and a schematic of this electrode array is illustrated in Figure 1. For each subject the 192 channels of information were sampled simultaneously for a number of seconds. Subsequent to recording the data were averaged to represent one cardiac cycle. Beat markers were then inserted on this averaged beat by a human expert.

Isointegral Measurements

Due to the abundance of data that is recorded using BSPMs, techniques have been developed that allow the effective reduction of this data prior to interpretation. A technique that has been widely adopted in BSPM representation is the use of 'isointegral' or 'isoarea' maps.

Symbol		
Normal Patients		**59**
MI Patients		**57**
Inferior	30	
Anterior	14	
Posterior	2	
Aterolateral	8	
Inferolateral	2	
Inferior-posterior	1	
Total		**116**

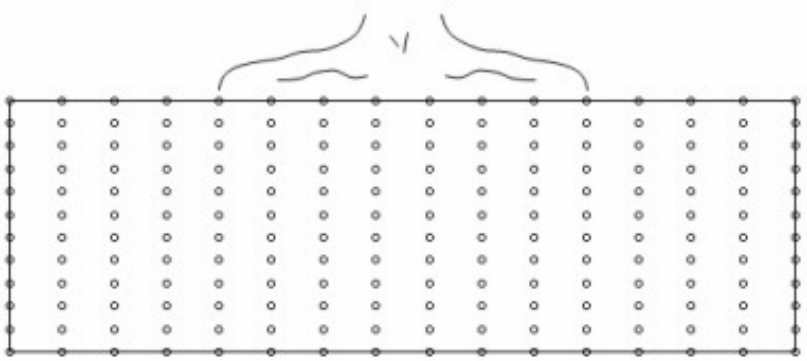


Figure 1. Schematic of electrode array employed to record BSPM data. This illustration depicts the array as an unrolled cylindrical matrix of 16 x 12 = 192 recording sites. The top row corresponds with a horizontal line running around the circumference at the level of the suprasternal notch. The bottom row corresponds to a horizontal line at the level of the umbilicus.

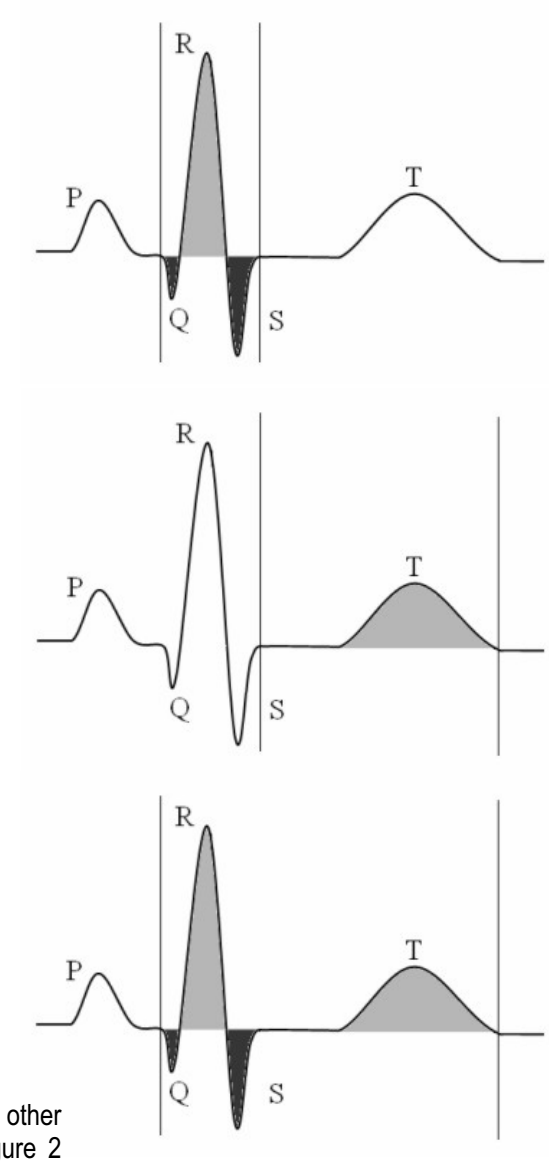


Figure 2. Illustration of area incorporated by (a) QRS, (b) STT and (c) QRST isointegrals.

In this approach the area under a specific portion of the ECG wave is calculated for each recorded lead; the resulting value for each individual lead is then used to generate a contour map. This technique summarizes the information contained in dozens of instantaneous maps into one picture [9], [10].

Although some information is lost, isointegral maps are useful as they provide an indication of the mean distribution of potentials over the selected interval. Two such maps that are commonly studied are the QRS isointegral and the STT isointegral [9]. These maps provide an indication of the mean distributions during ventricular activation (depolarization) and recovery (repolarization) respectively. QRST isointegral maps have also been studied as they provide an indication of the 'ventricular gradient', a measure of how much the processes of depolarization and repolarisation do not cancel one and other out in any particular lead [10]. Figure 2 illustrates the regions of a representative cardiac cycle incorporated in each isointegral.

In the current study QRS, STT and QRST isointegrals were calculated. Each of these isointegrals consisted of 192 values which are used to generate a contour map. As the patterns of extrema, maxima and minima of such a map are studied by the clinician in order to provide diagnosis and because these patterns are characterized by the 192 calculated values, these values can be considered as features in the context of computerized classification. For the studies presented in this paper, three such maps were calculated; this effectively resulted in 576 features for each subject (3 x 192). This also translates to having three features per recording site per patient, e.g. for each recording site we have one QRS, one STT, and one QRST value.

Feature selection

After calculation of the isointegral values further reduction in dimensionality was achieved by employing a signal-to-noise ratio-based feature ranking procedure. This approach is similar to the 'filter' method proposed in [11] where each individual isointegral feature was ranked based on its utility when considered as an input to a single variable classifier (SVC). In the current study each variable is ranked using a signal-to-noise ratio-based feature ranking criterion [12], [13], [14].

Let $\mu_1(f_i)$ and $\mu_2(f_i)$ be the mean values of feature f_i for the classes 1 and 2; $\sigma_1(f_i)$ and $\sigma_2(f_i)$ be the respective standard deviation values of i^{th} feature f_i for the same classes, hence S_i is determined as:

$$S_i = \frac{\mu_1(f_i) - \mu_2(f_i)}{\sigma_1(f_i) + \sigma_2(f_i)} \quad (1)$$

A higher value of $|S_i|$ indicates a stronger correlation between the feature value and the class distinction and hence infers that such a feature is useful in discriminating between classes.

Classification

Following the signal-to-noise ratio-based feature ranking the best subsets of three, six and ten measurements (features) from the 192 available for each isointegral were used as inputs to four classification models (NB, SVM, MLP and RF). A brief description of these common classifiers is given as follows:

NB is a simple probabilistic classifier. It is based on the Bayes rule of conditional probability and it naively assumes independence between features. It uses the normal distribution to model numeric attributes by calculating the mean standard deviation for each class [15].

SVM is a kernel based classifier. The basic training for SVMs involves finding a function which optimizes a bound on the generalization capability, i.e., performance on unseen data. By using the kernel trick technique, SVM can apply linear classification techniques to non-linear classification problems [16].

A MLP is a non-linear classification approach that may be trained using the back propagation algorithm. A MLP consists of multiple layers of computational units (an input layer, one or more hidden layer and one output layer) [17].

A RF classifier constructs a number of decision trees. Each tree is grown from a different set of training data which are randomly selected with replacement. At each decision node the RF determines the best splitting feature from a randomly selected subspace of features. The final classification is based on the majority votes among instances decided by the forest of trees [18].

In the evaluation of each classifier we used ten-fold cross validation. The quality of each classifier was assessed by the extent to which the correct class labels have been assigned. In order to appreciate the experimental outcomes it is important not only to examine how many samples have been correctly classified in relation to a particular class, but also to indicate how well a classifier can classify an unknown sample as not belonging to a particular class. Thus, this study evaluates classifiers based on three statistical measures: precision (*Ppv*) (equation 2), true positive rate (also known as sensitivity, *Se*) (equation 3) and true negative rate (also known as specificity, *Sp*) (equation 4) which can be calculated as follows:

$$Ppv(\%) = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$Se(\%) = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$Sp(\%) = \frac{TN}{FP + TN} \times 100\% \quad (4)$$

Where TP is true positive (samples correctly classified to appropriate class), FN is false negative (samples incorrectly classified as not belonging to appropriate class), FP is false positive (samples incorrectly classified as appropriate class), and TN is true negative (samples correctly classified as not belonging to appropriate class).

All four classification models were implemented within the framework provided by the Weka open-source platform [19]. The configuration of the various classification models is summarized as follows:

The SVM results were obtained by using a polynomial kernel. For the MLP model, the results were obtained using a model consisting of one hidden layer with six nodes when evaluating the top ten features, four nodes when evaluating the top six features and two nodes when considering the top three features (the choice of feature subsets is discussed later in the paper). Each MLP was trained for 500 epochs and the learning rate was set to 0.3. For the RF algorithm, ten trees were grown in each run and the minimum number of instances per leaf was equal to two. A more detailed description of the selection of learning parameters for these models can be found in [19].

3. Results

Feature selection

The feature selection approach adopted resulted in a set of scores for each of the isointegral types studied. As there are 192 feature scores, we were able to plot these as a 192 dimensional contour plot. These plots are illustrated in Figure 3. In the past this means of representation has been referred to as a lead performance map (LPM) [11], [20]. In Figure 3 we have plotted one such map for each isointegral. The LPMs effectively show the distribution of the scores for each available feature with the output as calculated using equation 1. Based on these values the features were ranked and the top three, six and ten features were selected. The locations of the recording sites from which these features are measured are illustrated in Figure 4.

Classification

The classification accuracy of the three subsets of features for each isointegral are listed in Table 2. These results illustrate the performance of the four different classifiers (NB, SVM, MLP and RF) on each feature subset. Final accuracies are based on ten-fold cross validation as previously described.

4. Discussion

We have divided this section between the discussions of the (a) selected features and associated recording sites and (b) the actual classification results.

Feature selection

Firstly, referring to the QRS isointegral based LPM depicted in Figure 3a. It can be seen that there are two areas where the correlation is greatest.

These are a region on the inferior anterior beneath the area interrogated by the standard precordial leads and a region on the superior posterior almost between the two shoulders. In the case of the STT isointegrals (Figure 3b) it can be seen that again there are two regions that correlate highly with the output. Again these are located on both the anterior and posterior surfaces, however, this time the area of high correlation on the anterior is located more laterally (towards the subject's left). The same also applies to the region on the posterior where the high, this time, is closer to the right shoulder as opposed to that in the QRS map.

The characteristics of the QRST LPM lie somewhere between that of the other two maps. This is to be expected as the QRST data is effectively a combination of that in the QRS and STT portions. Overall, these observations would indicate that, for this population, there would be benefit in locating recording sites outside the area interrogated by the standard locations in the 12 lead ECG. This consolidates the findings of similar previous studies [6], [11], [20], [21], [22], [23].

Classification

On analysis of the classification results that have been presented in Table 2, it can be seen that for each subset of features the QRS based features exhibit the poorest performance. This is observed from the fact that, regardless of the classifier or the size of the feature subset, the classification accuracy attained does not exceed 75%. In fact it is with this isointegral that lowest accuracy of all is observed, this is 62.9% using the RF classifier. The STT based features generally exhibit superior performance to the QRS based features as in most cases a classification accuracy in excess of 75% is observed. This is with the exception of the subset of three STT features in conjunction with the RF classifier which exhibits an accuracy of just under 70%.

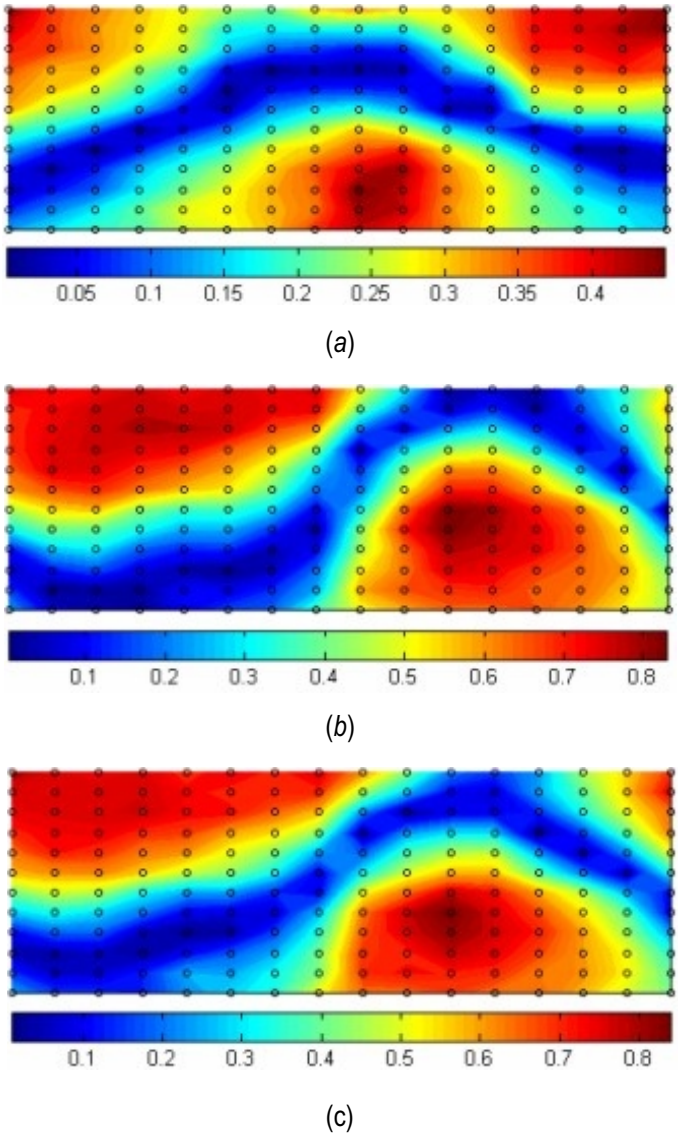


Figure 3. Lead Performance Maps showing spatial distribution of values as defined by Equation 1. Figures (a), (b) and (c) represent QRS, STT and QRST isointegrals respectively.

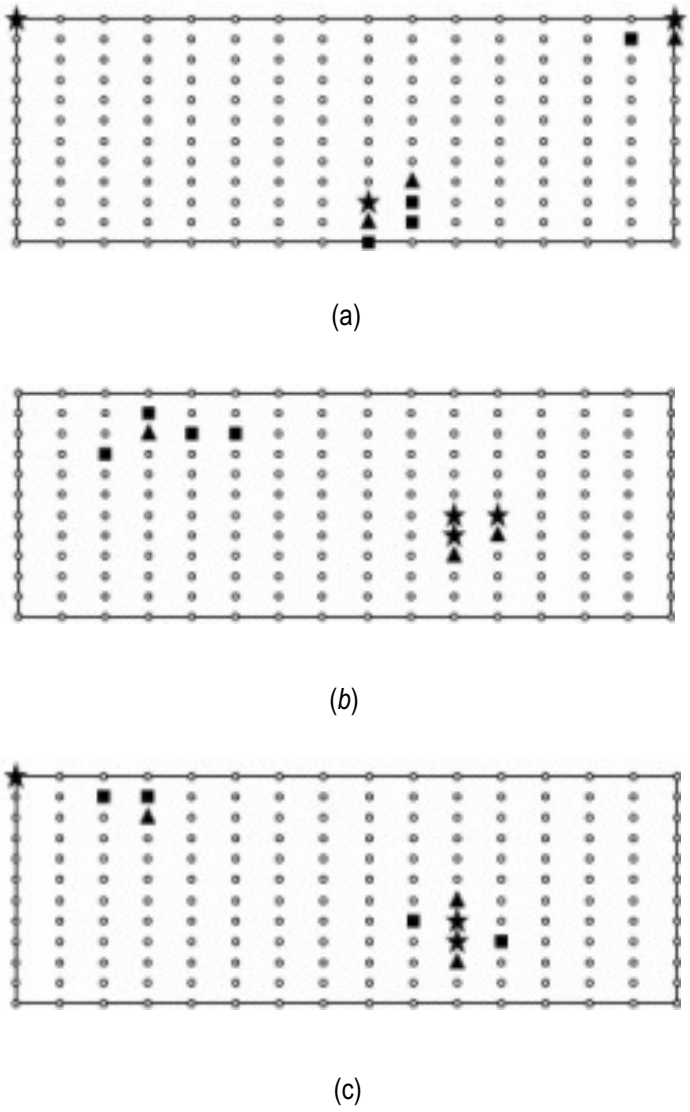


Figure 4. Positions of recording sites required to measure features selected using ranking method. Figures (a), (b) and (c) represent QRS, STT and QRST features respectively. In each case the top three features are shown as stars, the next three as triangles and the remaining four as squares.

Table 2. Performance of feature subsets for all four classifiers.

Features selected	Classification Accuracy (%)			
	NB	SVM	MLP	RF
Top 3 QRS features	71.6	74.1	71.6	63.8
Top 3 STT features	79.3	81.0	77.6	69.8
Top 3 QRST features	80.2	80.2	77.6	77.6
Top 6 QRS features	74.1	73.3	66.4	66.4
Top 6 STT features	81.0	79.3	78.4	80.2
Top 6 QRST features	78.4	80.2	76.7	81.0
Top 10 QRS features	73.3	73.3	64.7	62.9
Top 10 STT features	80.2	80.2	75.0	77.6
Top 10 QRST features	77.6	79.3	78.4	83.6

The QRST based features exhibit performance that is comparable to that obtained with the STT features. The QRST features also exhibit the highest attained accuracy which was 83.6%. This was obtained when using the RF classifier along with the subset of 10 features. The fact that similar results can be attained using the STT and QRST features may be based on the fact that the QRST data encompasses both the QRS and STT distributions.

5. Conclusion

Based on the above experiments and presented results we have illustrated how diagnostic electrocardiographic information is localized on the body surface. It can also be seen that the localities of this information may be outside the regions currently interrogated using the standard 12 lead ECG. These results have validated our initial hypothesis in that it is possible to improve the automated diagnostic process of cardiac assessment by trying to identify alternative subsets of features from BSPMs. Such findings offer the potential for future recommendations in alternative lead sets for cardiac assessment.

The current study has focused on the investigation of a generic dataset which represents both normal subjects and subjects with infarctions at various locations. In future studies we intend to extend the work by considering subgroups of patients. This includes the investigation of patients with other disease types (hypertrophy and conduction defects) and investigation of sub groups of MI patients based on infarct locations. A further issue that needs to be addressed is the impact of sample size on the prediction results [24]. In this study, we only use the BSPM recorded from 116 subjects. In order to statistically justify the results, a larger dataset is required.

Acknowledgment

The authors would like to acknowledge the support of Professor Robert L. Lux of the University of Utah, Salt Lake City in the realization of this study. In particular they would like to thank him for providing the clinical data used.

References

[1] Wagner G. S.: Marriott's Practical Electrocardiography, 10th Edition, Lippincott Williams & Wilkins, 2001.

[2] Menown I. B. A., Patterson R. S. H. W, MacKenzie G., Adgey A. A. J.: Body Surface Map Models for Early Diagnosis of Acute Myocardial Infarction. Journal of Electrocardiology 1998; 31, pp. 180-188.

[3] Sun G., Thomas C. W., Liebman J., Rudy Y., Reich Y., Stilli D., Macchi E.: Classification of Normal and Ischemia from BSPM by Neural Network Approach. In: Proceedings of the 10th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, 1988, pp. 1504-1505.

[4] Hoekema R., Uijen G. J. H., van Oosterom A.: On Selecting a Body Surface Mapping Procedure. Journal of Electrocardiology 1999; 32(2), pp. 93-101.

[5] Maynard S. J., Menown I. B., Manoharan G., Allen J., McC Anderson J., Adgey A. A.: Body Surface Mapping Improves Early Diagnosis of Acute Myocardial Infarction in Patients with chest Pain and Left Bundle Branch Block. Heart 2003; 89(9), pp. 998-1002

[6] Lux R. L., Smith C. R., Wyatt R. F., Abildskov J. A.: Limited Lead Selection for the Estimation of Body Surface Potential Maps in Electrocardiography. IEEE Transactions on Biomedical Engineering 1978; 25(3), pp. 270-276.

[7] Lux R. L., Burgess M. J., Wyatt R. F., Evans A. K., Vincent G. M., Abildskov J. A.: Clinically Practical Lead Systems for Improved Electrocardiography: Comparison with Precordial Grids and Conventional Lead Systems. Circulation 1979; 59(2), pp. 356-363.

[8] Lux R. L., Evans A. K., Burgess M. J., Wyatt R. F., Abildskov J. A.: Redundancy Reduction for Improved Display and Analysis of Body Surface Potential Maps. I. Spatial compression. Circulation Research 1981; 49, pp. 186-196.

[9] Taccardi B., Punske B. B., Lux R. L., MacLeod R. S., Ershler P. R., Dustman T. J., Vyhmeister Y.: Useful Lessons from Body Surface Mapping. Journal of Cardiovascular Electrophysiology 1998; 9, pp. 773-786.

[10] Flowers N. C., Horan L. G.: Body Surface Potential Mapping. In: Cardiac Electrophysiology: From Cell to Bedside.(Eds. D. Zipes, and J. Jalife), Saunders, 1995, pp. 1049-1067.

[11] Finlay D. D., Nugent C. D., McCullagh P. J., Black N. D.: Mining for Diagnostic Information in Body Surface Potential Maps: A Comparison of Feature Selection Techniques. Biomedical Engineering Online 2005, 4(51), 2005.

[12] Golub T. R., Slonim D. K., Tamayo P., Huard C., Gassenbeck M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., and Lander E. S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999; 286, pp. 531-537.

[13] Kornreich F., Montague T. J., Rautaharju P. M.: Identification of First Acute Q Wave and non-Q Wave Myocardial Infarction by Multivariate Analysis of Body Surface Potential Maps. Circulation 1991; 84, pp. 2422-2453.

[14] Kozmann G., Green L. S., and Lux R. L.: Nonparametric Identification of Discriminative Information in Body Surface Maps, IEEE Transactions on Biomedical Engineering, 1991; 38(11), pp. 1061-1068.

[15] Irina R.: An Empirical Study of the Naive Bayes Classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001.

[16] Burges C., A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 1998; 2, pp. 121-167.

[17] Hampshire J. B., Perlmutter B. A.: Equivalence Proofs for Multilayer Perceptron Classifiers and the Bayesian Discriminant Function. In: Proceedings of the 1990 Connectionist Models Summer School, 1990.

[18] http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm.

[19] Witten I. H. Frank E.: Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco, 2005.

[20] Lopez J. A., Nugent C. D., van Herpin G., Kors J. A., Finlay D., Black, N. D.: Visualisation of Electrocardiographic Features in Myocardial Infarction. In: Proceedings of the 29th Annual Conference of the International Society for Computerized Electrocardiology (ISCE), Journal of Electrocardiology. Vol. 37, 2004, p. 149.

[21] Barr R. C., Spach M. S., Herman-Giddens S.: Selection of the Number and Position of Measuring Locations for Electrocardiography. IEEE Transactions on Biomedical Engineering 1971; 18, pp. 125-138.

[22] Kors J. A., van Herpen G.: How Many Electrodes and Where: A "Poldermodel" for Electrocardiography. Journal of Electrocardiology 2002; 35(suppl.), pp. 7-12.

[23] Finlay D. D., Nugent C. D., Donnelly M. P., Lux R. L., McCullagh P. J., and Black N. D.: Selection of Optimal Recording Sites for Limited Lead Body Surface Potential Mapping: A sequential Selection Based Approach. BMC Medical Informatics and Decision Making; 6(9), pp. 1-9.

[24] Kozmann G., Lux R. L., and Scott M.: Sample Size and Dimensionality in Multivariate Classification: Implications for Body Surface Potential Mapping. Computers and Biomedical Research 1991; 24, pp. 170-182.

Ranked Modeling of Liver Diseases Sequence

Leon Bobrowski[1,2], Tomasz Łukaszuk[1], Hanna Wasyluk[3]
[1]Białystok Technical University, Faculty of Computer Science, Poland,
[2]Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw, Poland,
[3]Medical Center of Postgraduate Education, Warsaw, Poland

Summary: Ranked model in the form of linear transformation of multivariate feature vectors on a line can reflect a causal order between liver diseases. A priori medical knowledge about order between liver diseases and clinical data sets has been used in the definition of the convex and piecewise linear (CPL) criterion function.The linear ranked transformations have been designed here through minimization of such CPL criterion functions.

Keywords: sequential patterns, ranked linear transformations, convex and piecewise linear (CPL) criterion functions, linear separability of data sets, sequence of liver diseases

1. Introduction

Discovering regularities in multivariate data sets or databases is one of the main goals of exploratory data analysis and pattern recognition methods [1], [2]. Discovering trends in temporal databases is a particularly interesting problem with many important applications.

An adequate standardization is usually needed before applying data analysis tools. In a standardized clinical data representation, patients are represented in the form of feature vectors with the same number of numerical components (features) or as points in a multidimensional feature space. A particular pattern in clinical data is represented as a distinct set of feature vectors or as a special cloud of points in a feature space.

The regression analysis plays a prominent role in data exploration [3]. The regression model may describe a dependence of one feature on a selected set of other features. The ranked regression method can also serve a similar purpose [4], [5], [6]. The ranked models are particularly useful when values of the dependent feature cannot be measured precisely or directly and additional information about feature vectors is available only in the form of ranked relations within selected pairs of these vectors. Such ranked relations can be treated as a priori knowledge about linear sequential patterns hidden in data. In this context, inducing the linear ranked model from the ranked pairs can be treated as a pattern recognition problem. The induced ranked model can also be used for prognosis or decision support purposes.

The method of inducing linear ranked models from a set of feature vectors and ranked relations within selected pairs of these vectors was proposed in the previous papers [4], [5]. This method is based on the minimization of convex and piecewise-linear (*CPL*) criterion functions. Properties of this approach in the context of modeling a causal sequence of liver diseases are analysed in the presented paper. Feature vectors from hepatological database of the system Hepar and additional medical knowledge in the form of a causal sequence of liver diseases were used in designing ranked linear transformation [7].

2. Feature vectors and oriented dipoles

We are taking into consideration a data set C built from m feature vectors $\mathbf{x}_j=[x_{j1},...x_{jn}]^T$ which were numbered in a fixed manner

$$C = \{\mathbf{x}_j\}\ (j = 1,...,m) \qquad (1)$$

The vectors $\mathbf{x}_j$ belong to the n-dimensional feature space $F[n]$ ($x_j \in F[n]$). The component (feature) x_{ji} of the vector $\mathbf{x}_j$ is a numerical result of the i-th examination (i=1,...,n) of a given patient or event O_j (j=1,...,m). The feature vectors $\mathbf{x}_j$ are of a mixed type if they represent different types of diagnostic measurements ($x_i \in \{0,1\}$) or ($x_i \in R$)).

Let the symbol "$\prec$" means the relation "*follows*" which is fulfilled within ranked pairs $\{\mathbf{x}_j, \mathbf{X}_j'\}$ ($j<j'$) of the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_j'$ with the indices (j,j') from some set :

$$(\forall (j,j') \in J) \quad \mathbf{x}_j \prec \mathbf{x}_{j'} \Leftrightarrow (\mathbf{x}_{j'} \ \ follows \ \ \mathbf{x}_j)$$
$$or \quad \mathbf{x}_{j'} \prec \mathbf{x}_{j''} \Leftrightarrow (\mathbf{x}_j \ \ follows \ \ \mathbf{x}_{j'}) \qquad (2)$$

The relation $\mathbf{x}_j \prec \mathbf{x}_j'$ between the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_j'$ means that the vector $\mathbf{x}_j'$ follows the vector $\mathbf{x}_j$ in some sequence. This relation should be determined on the basis of some additional information about set of pairs (not necessary all) of the vectors $\mathbf{x}_j$. For example, medical doctors can compare two patients with the same disease and decide that one of them is in a more serious condition. As another example, it can be stated that one of two students is more talented than another one.

In the paper we analyse the problem of designing such transformations of the feature vectors $\mathbf{x}_j$ on the (ranked) line $y=\mathbf{w}^T\mathbf{x}$ which preserve the relation "$\prec$" (2) as precisely as possible

$$y_j = y_j(\mathbf{w}) = \mathbf{w}^T\mathbf{x}_j, \qquad (3)$$

where $\mathbf{w}=[w_1,...,w_n]^T$ is the weight vector.

The family of the relations (2) defines the sequential pattern $S(\mathbf{x})$ of the vectors x_j in the feature space $F[n]$ ($\mathbf{x}_j \in F[n]$).

Definition 1: The sequential pattern $S(\mathbf{x})$ is linear in the feature space $F[n]$ if and only if there exists such n-dimensional weight vector $\mathbf{w}$ ($\mathbf{w} \in R^n$) that the below implication holds:

$$(\forall (j,j') \in J) \quad \begin{array}{l} \mathbf{x}_j \prec \mathbf{x}_{j'} \Rightarrow \mathbf{w}^T\mathbf{x}_j < \mathbf{w}^T\mathbf{x}_{j'} \\ and \\ \mathbf{x}_{j'} \prec \mathbf{x}_j \Rightarrow \mathbf{w}^T\mathbf{x}_{j'} < \mathbf{w}^T\mathbf{x}_j \end{array} \tag{4}$$

where J is a set of indices (j, j') of ranked pairs of vectors $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$.

Definition 2: The ranked pair $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$ of the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ constitutes the *positively oriented dipole* $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $((j,j' \in J^+)$, if and only if $\mathbf{x}_j \prec \mathbf{x}_{j'}$.

$$(\forall (j,j') \in J^+) \quad \mathbf{x}_j \prec \mathbf{x}_{j'} \tag{5}$$

Definition 3: The ranked pair $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$ of the feature vectors $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ constitutes the *negatively oriented dipole* $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $((j,j' \in J)$, if and only if $\mathbf{x}_{j'} \prec \mathbf{x}_j$.

$$(\forall (j,j') \in J^-) \quad \mathbf{x}_{j'} \prec \mathbf{x}_j \tag{6}$$

Definition 4: The line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) is completely ranked in accordance with the dipoles $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$ orientations if and only if

$$(\forall (j,j') \in J^+) \quad y_j(\mathbf{w}) < y_{j'}(\mathbf{w}) \quad and \quad (\forall (j,j') \in J^-) \quad y_j(\mathbf{w}) > y_{j'}(\mathbf{w}) \tag{7}$$

where J^+ and J^- are sets of indices (j, j') of the positively and negatively oriented dipoles $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$, and $J^+ \cup J^- = J$, $J^+ \cap J^- = \emptyset$.

3. Designing ranked models through minimization of a CPL criterion function

Let us introduce the positive set R^+ and the negative set R^- of the differential vectors $\mathbf{r}_{jj'}=(\mathbf{x}_{j'}-\mathbf{x}_j)$ on the basis of the sets of indices J^+ (6) and J (7).

$$\begin{array}{l} R^+ = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j,j') \in J^+\} \\ R^- = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j,j') \in J^-\} \end{array} \tag{8}$$

We examine the possibility of separating the sets R^+ and R^- by the hyperplane $H(\mathbf{w})$ passing through the origin 0 of the feature space.

$$H(\mathbf{w}) = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = 0\} \tag{9}$$

Definition 5: The sets R^+ and R^- (8) are separable by some hyperplane $H(\mathbf{w})$ (9) if and only if the below inequalities hold

$$(\exists \mathbf{w}) \quad \begin{array}{l} (\forall (j,j') \in J^+) \quad \mathbf{w}^T\mathbf{r}_{jj'} > 0 \\ (\forall (j,j') \in J^-) \quad \mathbf{w}^T\mathbf{r}_{jj'} < 0 \end{array} \tag{10}$$

If all the above inequalities are fulfilled for some vector $\mathbf{w}$, then the hyperplane $H(\mathbf{w})$ (9) separates the sets R^+ and R^- (8).

Lemma 1: The linear transformation $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) is completely ranked (7) in accordance with dipoles' $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ orientations if and only if the hyperplane $H(\mathbf{w})$ (9) separates (10) the sets R^+ and R^- (8).

Proof: If the hyperplane $H(\mathbf{w})$ (9) separates the sets R^+ and R^- (8), then the ranked relations (7) hold on the line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3). On the other hand, the fulfilling of the ranked relations (7) guarantees that the inequalities (10) hold.

Designing a separating hyperplane $H(\mathbf{w})$ (9) could be carried out through the minimization of the convex and piecewise linear (*CPL*) criterion function similar to the perceptron criterion function $\Phi(\mathbf{w})$ [2]. For this purpose let us introduce the positive $\varphi_{jj'}^+(\mathbf{w})$ and the negative $\varphi_{jj'}^-(\mathbf{w})$ penalty functions:

$$(\forall (j,j') \in J^-) \quad \varphi_{jj'}^-(\mathbf{w}) = \begin{array}{ll} 1+\mathbf{w}^T\mathbf{r}_{jj'} & if \quad \mathbf{w}^T\mathbf{r}_{jj'} > -1 \\ 0 & if \quad \mathbf{w}^T\mathbf{r}_{jj'} \le -1 \end{array} \tag{11}$$

$$(\forall (j,j') \in J^-) \quad \varphi_{jj'}^-(\mathbf{w}) = \begin{array}{ll} 1+\mathbf{w}^T\mathbf{r}_{jj'} & if \quad \mathbf{w}^T\mathbf{r}_{jj'} > -1 \\ 0 & if \quad \mathbf{w}^T\mathbf{r}_{jj'} \le -1 \end{array} \tag{12}$$

The criterion function $\Phi(w)$ is the weighted sum of the penalty functions $\varphi_{jj'}^+(\mathbf{w})$ and $\varphi_{jj'}^-(\mathbf{w})$:

$$\Phi(\mathbf{w}) = \sum_{(j,j'') \in J^+} \gamma_{jj''} \varphi_{jj''}^+(\mathbf{w}) + \sum_{(j,j'') \in J^-} \gamma_{jj''} \varphi_{jj''}^-(\mathbf{w}) \tag{13}$$

where $\gamma_{jj'}$ $(\gamma_{jj'}>0)$ is a positive parameter (*price*) related to the dipole $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $(j<j')$.

$\Phi(\mathbf{w})$ (13) is the convex and piecewise linear (*CPL*) function as the sum of the penalty functions $\varphi_{jj'}^+(\mathbf{w})$ and $\varphi_{jj'}^-(\mathbf{w})$ of the same kind. The basis exchange algorithms, similar to the linear programming, allow to find the minimum of such function efficiently, even in the case of large, multidimensional data sets R^+ and R^- [7]:

$$\Phi^* = \Phi(\mathbf{w}^*) = \min_{\mathbf{w}} \Phi(\mathbf{w}) \ge 0 \tag{14}$$

The optimal parameter vector $\mathbf{w}^*$ and the minimal value Φ^* of the criterion function $\Phi(\mathbf{w})$ (13) can be applied to a variety of data ranking problems. In particular, the vector $\mathbf{w}^*$ defining the best ranked line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) can be found this way.

The minimal value Φ^* (14) of the criterion function $\Phi(\mathbf{w})$ (13) can be used to measure the linearity of the sequential patterns $S(\mathbf{x})$ (Def. 1) in a given feature space $F[n]$.

Lemma 2: The minimal value Φ^* (14) of the criterion function $\Phi(\mathbf{w})$ (13) is equal to zero if and only if the sequential pattern $S(\mathbf{x})$ (Def. 1) is linear.

Proof: If there exists such a vector $\mathbf{w}^*$ that the ranking of the points $y_j(\mathbf{w}^*)$ on the line (3) is fully consistent (7) with the relations "$\prec$", then the sets R^+ and R^- (8) can be separated (10) by the hyperplane $H(\mathbf{w}^*)$ (9). In this case, the minimal value Φ^* of the perceptron criterion function $\Phi(\mathbf{w})$ (13) is equal to zero, as it results from the pattern recognition theory [1]. On the other hand, if the minimal value Φ^* (14) of the criterion function $\Phi(\mathbf{w})$ (13) is equal to zero in the point $\mathbf{w}^*$, then the values $\varphi_{jj'}^+(\mathbf{w})$ and $\varphi_{jj'}^-(\mathbf{w})$ of all the penalty functions (11) and (12) have to be equal to zero. This means that the sets R^+ and R^- (8) can be separated (10) by the hyperplane $H(\mathbf{w}^*)$ (9). As a result, the ranking of the points $y_j(\mathbf{w}^*)$ on the line (3) is fully consistent (7) with the relations (4) and (5).

4. Causal sequence of learning sets

Let us assume that a clinical database contains descriptions of m patients $O_j(k)$ $(j=1,...m)$ labeled in accordance with their clinical diagnosis $\omega_k(k=1,...K)$. Each patient $O_j(k)$ is represented by n-dimensional feature vector $\mathbf{x}_j(k)$. The feature vector $\mathbf{x}_j(k)$ represents the j-th patient $O_j(k)$ linked to the k-th disease ω_k. The learning set C_k contains m_k labeled feature vectors $\mathbf{x}_j(k)$ that are linked to the k-th disease (class) ω_k.

$$C_k = \{x_j(k)\} \quad (j \in I_k) \tag{15}$$

where I_k is the set of indices j of m_k feature vectors $\mathbf{x}_j(k)$ labeled to the class ω_k.

We assume that the learning set C_k have been formed in a learning causal sequence:

$$C_1 \rightarrow C_2 \rightarrow ... \rightarrow C_{K-1} \rightarrow C_K \tag{16}$$

where symbol "$C_{k-1} \rightarrow C_k$" means that "disease ω_k appears after ω_{k-1}" or "disease ω_{k-1} is a cause of ω_k". The consistent *indexing* of the sets C_k and the diseases ω_k has been used in the sequence (16). This means that:

$$(\forall k,k' \in \{1,...,K\}) \quad (k < k') \Rightarrow (C_k \rightarrow C_{k'}) \tag{17}$$

The causal relation "$C_k \rightarrow C_{k'}$" (17) between learning sets C_k and $C_{k'}$ can be used for determining the causal ranked relation "$\prec$" (2) between feature vectors $\mathbf{x}_j(k)$ ($\mathbf{x}_j(k) \in C_k$) and $x_{j'}(k')$ ($x_{j'}(k) \in C_k$) (15):

$$(\forall k,k' \in \{1,...,K\}) \quad (C_k \rightarrow C_{k'}) \Rightarrow ((\forall \mathbf{x}_j(k) \in C_k) \; and \; (\forall \mathbf{x}_{j'}(k') \in C_{k'})) \; \mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k') \tag{18}$$

or

$$(\forall k,k' \in \{1,...,K\}) \quad (k < k') \Rightarrow (\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')) \tag{19}$$

Let us remark that there is no ranked relation "$\prec$" (2) between feature vectors $\mathbf{x}_j(k)$ and $x_{j'}(k)$ from the same set .

We can assume that the indices *j* of the feature vectors $\mathbf{x}_j(k)$ are consistent with the learning sets C_k (15). This means that the set C_1 contains m_1 first feature vectors $\mathbf{x}_j(k)$, the set C_2 contains m_2 next vectors $\mathbf{x}_j(k)$, and so on. As a consequence, the following relation of consistent indexing holds:

$$(\forall \mathbf{x}_j(k) \in C_k) and (\forall \mathbf{x}_{j'}(k') \in C_{k'}) \; (\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')) \Rightarrow (j < j') \tag{20}$$

Lemma 3: In the case of consistent indexing (20), the linear transformation $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) is completely ranked (7) if and only if the set R^+ (8) of the differential vectors $\mathbf{r}_{jj'}=(\mathbf{x}_{j'}-\mathbf{X}_j)$ is situated on the positive side of some hyperplane $H(\mathbf{w})$ (9):

$$(\exists \mathbf{w})(\forall \mathbf{r}_{jj'} \in R^+) \quad \mathbf{w}^T\mathbf{r}_{jj'} > 0 \tag{21}$$

Proof: The relations (19) and (20) guarantee that all dipoles $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ are positively oriented (*Def. 2*) and the negative set R^- (8) is empty ($R^-=\emptyset$). As a result, the vector $\mathbf{w}$ defines such hyperplane $H(\mathbf{w})$ which linearly separates (10) the sets R^+ and R^-(8). This means that the assumptions of the *Lemma 1* are fulfilled.

Let us assume that the set R^+ (8) contains all positively oriented dipoles $\{\mathbf{x}_j(1), \mathbf{x}_{j'}(2)\}(j<j')$ (5), that can be generated from two learning sets C_1 and C_2 (15) in accordance with the relation $C_1 \rightarrow C_2$ (16) and consistent indexing (20),

$$R^+ = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'}(2) - \mathbf{x}_j(1)) : \mathbf{x}_j(1) \in C_1 \; and \; \mathbf{x}_{j'}(2) \in C_2\} \tag{22}$$

The set R+ (22) is complete if it contains all positively oriented dipoles $\{\mathbf{x}_j(1), \mathbf{x}_{j'}(2)\}$(5) that can be created from two learning sets C_1 and C_2 (15) with the relation $C_1 \rightarrow C_2$(16).

Definition 6: Two learning sets C_1 and C_2 (15) are linearly separable if and only if the below inequalities hold

$$(\exists \mathbf{w},\theta) \quad \begin{matrix} (\forall \mathbf{x}_{j'} \in C_2) & \mathbf{w}^T\mathbf{x}_{j'} > \theta \\ (\forall \mathbf{x}_j \in C_1) & \mathbf{w}^T\mathbf{x}_{j'} < \theta \end{matrix} \tag{23}$$

where θ ($\theta \in R^1$) is a *threshold*.

The above parameters $(\mathbf{w},\theta)$ define the hyperplane $H(\mathbf{w},\theta)$ in the feature space, where:

$$H(\mathbf{w},\theta) = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = \theta\} \tag{24}$$

It is possible to relate the linear separability of two learning sets C_1 and C_2 (15) with the complete ranking of the linear transformation $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3).

Theorem 1: The linear transformation $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) is completely ranked (21) in accordance with the complete set R^+ (22) if and only if there exists such threshold θ' that the hyperplane $H(\mathbf{w},\theta')$ (24) separates two learning sets C_1 and C_2 (15).

Proof: If the line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) is fully ranked (7), then

$$(\exists \mathbf{w})(\forall \mathbf{x}_j \in C_1) and (\forall \mathbf{x}_{j'} \in C_2) \quad \mathbf{w}^T\mathbf{x}_{j'} > \mathbf{w}^T\mathbf{x}_j \tag{25}$$

Let us define the positive $\theta^+(\mathbf{w})$ and the negative $\theta^-(\mathbf{w})$ thresholds on the line $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$:

$$\theta^+(\mathbf{w}) = \min\{\mathbf{w}^T\mathbf{x}_{j'} : \mathbf{x}_{j'} \in C_2\} \tag{26}$$

$$\theta^-(\mathbf{w}) = \max\{\mathbf{w}^T\mathbf{x}_j : \mathbf{x}_j \in C_1\} \tag{27}$$

The below inequality results from the relation (25)

$$\theta^+(\mathbf{w}) > \theta^-(\mathbf{w}) \tag{28}$$

The threshold θ' can be defined as follows

$$\theta' = \frac{\theta^+(\mathbf{w}) + \theta^-(\mathbf{w})}{2} \tag{29}$$

It can be directly verified that the hyperplane $H(\mathbf{w},\theta')$ (24) separates the learning sets C_1 and C_2 (15).

On the other hand, the hyperplane $H(\mathbf{w},\theta')$ (24) that separates sets C_1 and C_2 (15), determines a linear transformation $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) which is completely ranked (21). As it results from the definition of the linear separabilty (23), each element $\mathbf{r}_{jj'} = \mathbf{x}_{j'}(2) - \mathbf{x}_j(1)$ of the set R^+(22) fulfills the relation (21).

5. Causal sequence of liver diseases

The database of the system *Hepar* contains descriptions of patients with variety of chronic liver diseases ω_k (k=1,...K) [7]. The feature vectors $\mathbf{x}$ in the database of *Hepar* are the mixed, qualitative-quantitative type. They contain both symptoms and signs ($x_i \in \{0,1\}$) as well as the numerical results of laboratory tests ($x_i \in R$). About 200 different features x_i describe one patient case in this system.

For the purpose of these computations, each patient has been described by the feature vector $\mathbf{x}_j(k)$ composed of 62 features x_i chosen as a standard by medical doctors. The following K=7 groups of patients C_k (15) have been extracted from the *Hepar* database:

C_1 *Non hepatitis patients* - 16 patients
C_2 *Hepatitis acuta* - 8 patients
C_3 *Hepatitis persistens* - 44 patients
C_4 *Hepatitis chronica activa* - 95 patients
C_5 *Cirrhosis hepatitis compensata* - 38 patients
C_6 *Cirrhosis decompensata* - 60 patients
C_7 *Carcinoma hepatis* - 11 patients

Total: 272 patients (30)

The data sets C_k (30) have been formed as the causal learning sequence (16) in accordance with medical knowledge. The ranked relation "$\prec$" (2) between feature vectors $\mathbf{x}_j(k)$ $(\mathbf{x}_j(k) \in C_k)$ and $x_j(k')$ $(x_j(k) \in C_k)$ (15) has been defined (18) on the basis of the causal sequence (16). This ranked relation allowed to define both oriented dipoles $\{x_j, x_j\}$(5) (6) as well as the positive set R^+ and the negative set R^- (8) of the differential vectors $\mathbf{r}_{jj'}(\mathbf{x}_{j'}-\mathbf{x}_j)$. The sets R^+ and R^- have been used in the definition of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ (13). The optimal parameter vector $\mathbf{w}^*$ (14) constituting the minimum of the function $\Phi(\mathbf{w})$ (13) defines the ranked linear model (3) that can be used for prognosis purposes:

$$y_j = y_j(\mathbf{w}^*) = (\mathbf{w}^*)^T \mathbf{x}_j = w_1^* x_{j1} + \ldots + w_n^* x_{jn} \quad (31)$$

The solution of the feature selection problem allows to determine the most important features x_i influencing the future of a given patient x_0 and to neglect the unimportant features x_i. The feature selection problem can also be based on the minimization of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ (13) [7].

The linear model (31) fulfills the ranked relation (4) for a great part of feature vectors $\mathbf{x}_j$:

$$\mathbf{x}_j \prec \mathbf{x}_{j'} \Rightarrow (\mathbf{w}^*)^T \mathbf{x}_j < (\mathbf{w}^*)^T \mathbf{x}_{j'} \quad (32)$$

As a result, the causal sequence (16) of the learning sets C_k (30) is preserved in a great part by the ranked model (31). In accordance with the equation (31), each learning set C_k (30) is transformed in the set C_k' of the points $\mathbf{x}_j(k)$ on the ranked line:

$$C_k' = \{y_j(k)\} \quad (j \in I_k) \quad (33)$$

The sets C_k' can be characterized by mean values μ_k and variances σ_k^2, where

$$\mu_k = \frac{\sum_j y_j(k)}{m_k} \quad (j \in I_k) \quad (34)$$

and

$$\sigma_k^2 = \frac{\sum_j (y_j(k) - \mu_k)}{m_k} \quad (j \in I_k) \quad (35)$$

The results of computations based on the model (31) of data sets C_k (30) are summarized in the below Table 1:

We can remark that the ranked implications (32) do not depend on the linear scaling (36) of the model. This means that

$$(\forall \alpha > 0)(\forall \beta) \quad (\mathbf{w}^*)^T \mathbf{x}_j < (\mathbf{w}^*)^T \mathbf{x}_{j'} \Rightarrow \alpha(\mathbf{w}^*)^T \mathbf{x}_j + \beta < \alpha(\mathbf{w}^*)^T \mathbf{x}_{j'} + \beta \quad (37)$$

The parameters α and β have been fixed through minimization of the sum $Q(\alpha,\beta)$ of the differences $|k-\alpha(\mathbf{w}^*)^T\mathbf{x}_j(k)+\beta|$ for all the sets C_k (30) and all the feature vector $\mathbf{x}_j(k)$.

Table 1. The mean values μ_k and variances σ_k^2 of the sets C_k' (33).

Data sets C_k' (33)	Number of patients m_k	Mean value μ_k	Variance σ_k^2 (σ_k)
C_1'	16	-1.02	0.46 (0.68)
C_2'	8	-0.58	0.57 (0.76)
C_3'	44	0.12	1.1 (1.05)
C_4'	95	0.89	1.46 (1.21)
C_5'	38	2.11	2 (1.41)
C_6'	60	3.02	2.2 (1.48)
C_7'	11	3.78	0.62 (0.79)

Let us consider an additional linear scaling $y'=\alpha y+\beta$ of the model $y=(\mathbf{w}^*)^T\mathbf{x}$ (31) in order to improve the interpretability of its prognostic applications.

$$y_j'(k) = \alpha(\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \quad (36)$$

where α and β are the scaling parameters.

$$Q(\alpha,\beta) = \sum_{k=1,\ldots,K} \sum_{j \in I_k} |k - \alpha(\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta| \quad (38)$$

where I_k is the set of indices j of the feature vectors $\mathbf{x}_j(k)$ from the set C_k (30).

Let us remark that $\mathcal{Q}(\alpha,\beta)$ is the convex and piecewise linear (*CPL*) function. The basis exchange algorithms also allow to find efficiently the parameters α^* and β^* constituting the minimum of the function $\mathcal{Q}(\alpha,\beta)$. Some results of the scaled model evaluation are shown in the Table 2 and on the Fig. 1.

The linear ranked model $y=\alpha^*(\mathbf{w}^*)^T\mathbf{x}+\beta^*$ can be used in the diagnosis support of a new patient $\mathbf{x}_0$. The location of the point $y_0=\alpha^*(\mathbf{w}^*)^T\mathbf{x}_0+\beta^*$ on the ranked line (30) constitutes a valuable characteristic of the patient and his perspectives. In the case the scaled model (Fig. 1), we can expect that the point y_0 representing a new patient $\mathbf{x}_0$ with the k-th disease ω_k will be situated near the index k.

6. Concluding remarks

The linear ranked models can be induced from data sets C_k (15) on the basis of additional medical knowledge in the form of the causal sequence (16) of diseases ω_k (k=1,...,K). The ranked relation "$\prec$" (2) between feature vectors $\mathbf{x}_j(k)$ from different learning sets C_k and C_k' has been defined (18) on the basis of the causal sequence (16). This ranked relation allowed to define both the oriented dipoles $\{x_j, x_j\}$ (5) (6) as well as the positive set R^+ and the negative set R^- (8) of the differential vectors .

The sets R^+ and R^- (8) have been used in the definition of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w})$ (13). The optimal parameter vector $\mathbf{w}^*$ (14), which is the minimum point of the function $\Phi(\mathbf{w})$ (13) defines the ranked linear model (31) that can be used for the purpose of prognosis. The prognostic model (31) can be improved through linear scaling (36). An example of the *CPL* criterion function (α,β) for choosing the scaling parameters α and β is provided by the equation (38).

The feature selection problem allows to determine the most important features $\mathbf{x}_i$ influencing significantly the future of a given patient and to neglect unimportant features. The feature selection problem can be solved through the minimization of a modified *CPL* criterion function $\Phi(\mathbf{w})$ (13) [6], [7].

Table 2. The mean values μ_k' and variances $\sigma_k'^2$ of the sets C_k' (33) obtained from the ranked model (31) after scaling (36) with the optimal parameters α^* and β^*.

Data sets C_k' (33)	Number of patients m_k	Mean value μ_k'	Variance σ_k^2 (σ_k)
C_1'	16	1.41	0.64 (0.8)
C_2'	8	1.93	0.79 (0.89)
C_3'	44	2.75	1.51 (1.23)
C_4'	95	3.65	1.99 (1.41)
C_5'	38	5.08	2.74 (1.65)
C_6'	60	6.14	3.02 (1.74)
C_7'	11	7.03	0.85 (0.92)

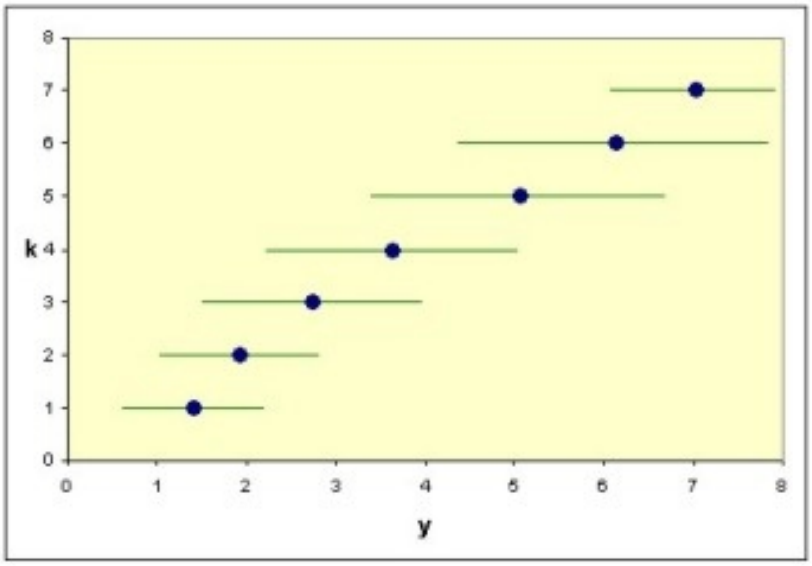


Figure 1. Graphical presentation the mean values μ_k' and variances $\sigma_k'^2$ of the sets C_k' (33) obtained from the ranked model (31) after scaling (36) with the optimal parameters α^* and β^*.

The ranked model of liver diseases (31) could be applied in screening procedures in the search for potentially ill patients eligible for further investigations and therapy. The ranked model (31) can also be specified for risk prognosis for individual patients.

This work was partially supported by the KBN grant 3T11F01130, by the grant 16/St/2007 from the Institute of Biocybernetics and Biomedical Engineering PAS, and by the grant W/II/1/2007 from the Białystok University of Technology.

References

[1] Duda O. R., Hart P. E., Stork D. G.: Pattern Classification, J. Wiley, New York, 2001.

[2] Fukunaga K.: Introduction to Statistical Pattern Recognition, Academic Press 1972.

[3] Johnson R. A., Wichern D. W.: Applied Multivariate Statistical Analysis, Prentice-Hall Inc., Englewood Cliffs, New York, 1991.

[4] Bobrowski L., Łukaszuk T.: Ranked Linear Modeling in Survival Analysis, pp. 61-67 in: Lecture Notes of the ICB Seminars: Statistics and Clinical Practice, ed. by L. Bobrowski, J. Doroszewski, N. Victor, IBIB PAN, Warsaw, 2005.

[5] Bobrowski L.: Ranked Modelling with Feature Selection Based on the CPL Criterion Functions, in: Machine Learning and Data Mining in Pattern Recognition, eds. P. Perner et al., Lecture Notes in Computer Science vol. 3587, Springer Verlag, Berlin, 2005.

[6] Bobrowski L.: Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych (Data mining based on convex and piecewise linear (CPL) criterion functions) (in Polish), Białystok Technical University, 2005.

[7] Bobrowski L., Wasyluk H.: Diagnosis Support rules of the Hepar system, pp. 1309-1313 in: MEDINFO 2001, eds: V. L. Petel, R. Rogers, R. Haux, IOS Press, Amsterdam, 2001.

[8] Bobrowski L.: Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique, Pattern Recognition, 24(9), pp. 863-870, 1991.

EJBI 2007 ISSN 1801 - 5603

European Journal for Biomedical Informatics

Volume 3 (2007), Issue 1

Croatian version

www.ejbi.eu

EuroMISE s.r.o.

Content

Croatian version

Informacijska i komunikacijska tehnologija u ordinacijama obiteljske medicine u Hrvatskoj

Josipa Kern[1], **Ozren Polašek**[1]
[1]Andrija Stampar School of Public Health, Zagreb University Medical School, Croatia

Sažetak: U Hrvatskoj je obiteljska medicina bila prvi privatizirani segment sustava zdravstvene zaštite. Zasad je privatizirano oko 84 % ordinacija obiteljske medicine dok je preostalih 16 % ostalo u sastavu domova zdravlja. Da bismo procijenili status informatizacije ordinacija obiteljske medicine, anketirali smo liječnike na specijalizaciji iz obiteljske medicine za vrijeme njihovog poslijediplomskog studija na Medicinskom fakultetu Sveučilišta u Zagrebu u školskoj godini 2004/05. Uzorak se sastojao od 159 liječnika. Pokazalo se da neki oblik informacijskog sustava postoji u 62 % ordinacija u zakupu, 42 % ordinacija u domovima zdravlja te 91 % potpuno privatnih ordinacija obiteljske medicine. Svoj postojeći informacijski sustav liječnik obiteljske medicine prvenstveno upotrebljava za administrativne svrhe: izvještavanje, preskripcije, bolovanja, uputnice i račune. Što se tiče medicinskog dijela rada, elektronički medicinski zapis koristi oko 50 % liječnika, istraživanja, koja se temelje na podacima pohranjenim u e-formatu, provodi oko 31 % liječnika, dok oko 35 % koristi podatke iz sustava za evaluaciju svog rada. Bibliografske baze podataka i elektroničke časopise čita 86 % liječnika iz privatnih ordinacija, 79 % onih koji rade u ordinacijama obiteljske medicine unutar domova zdravlja te 60 % liječnika u ordinacijama u zakupu. Zadovoljstvo s postojećim informacijskim sustavima nije veliko (19 %). Informacijski sustavi naime samo parcijalno zadovoljavaju informacijske potrebe liječnika. 32 % smatra da uz pomoć postojećeg informacijskog sustava dobiva dovoljno informacija, 40 % smatra da im je posao postao učinkovitiji, a 25 % da su i njihovi pacijenti zadovoljniji što njihov liječnika ima kompjutorizirani informacijski sustav u svojoj ordinaciji. S obzirom na tri tipa prakse može se reći da su mlađi doktori, koji rade unutar domova zdravlja, manje zadovoljni svojim informacijskim sustavom nego li doktori iz privatnih ordinacija ili ordinacija u zakupu. Zaštita podatka u sustavu se provodi pomoću zaporke, fizičkom zaštitom računala te svakodnevnim arhiviranjem podatka.

Ključne riječi: informacijski sustav, ordinacija obiteljske medicine, elektronički medicinski zapis, Internet, sigurnost računalnih sustava, bibliografske baze podataka, zadovoljstvo korisnika

1. Uvod

Obiteljska je medicina temelj zdravstvene zaštite populacije Hrvatske. Jednako je važna i u prevenciji bolesti i u liječenju unaprijed definirane populacije koja pripada određenoj ordinaciji odnosno određenom liječniku ili liječničkom timu obiteljske medicine. Sa željom da postignu, održavaju i unapređuju kvalitetu zdravstvene zaštite, liječnici bilježe podatke o pacijentima, analiziraju ih i/ili pretražuju podatke, informacije i znanja kada rješavaju konkretni problem svog pacijenta. S obzirom na činjenicu da se operabilnost podataka može postići isključivo pomoću računalne tehnologije, vrijeme je da se takva tehnologija udomaći u ordinacijama obiteljske medicine i to na primjereni način. Primjerenost tehnologije podrazumijeva dostupnost informacija sa strane liječnika i medicinske sestre, ali i sa strane šire zajednice (ako liječnik treba dati određene informacije, primjerice zavodu za javno zdravstvo, osiguranju, sekundarnoj ili tercijarnoj zdravstvenoj zaštiti, ministarstvu itd.). Jasno je da pritom moraju biti primijenjene sve potrebne mjere zaštite i povjerljivosti podatka. Liječnik očekuje da njegov informacijski sustav promptno osigurava potrebne informacije, da bude modularan i skalabilan, da se može obnavljati i razvijati odnosno unapređivati prema potrebama korisnika. Sustav treba uključivati međunarodne norme, i treba biti evaluiran i sa strane zdravstvenog osoblja i sa strane informatičara [1].

Zdravstveno osoblje se mora osposobiti da koristi informacijsku tehnologiju, mora biti svjesno mogućnosti i ograničenja takvih tehnologija te motivirano za razvijanje i unapređivanje informacijskih sustava.

Kao zemlja u tranziciji, Hrvatska je započela s privatizacijom u sustavu zdravstvene zaštite 1994. godine. Obiteljska je medicina bila prvi segment zdravstva koji je doživio privatizaciju.

Trenutno je privatizirano 84 % ordinacija obiteljske medicine. Preostalih 16 % ostalo je u sklopu domova zdravlja. Od toga je 95% ordinacija u zakupu (s opremom i prostorom u vlasništvu domova zdravlja; zakup plaćaju) i 5 % potpuno privatnih ordinacija (vlastiti prostor i oprema). Ordinacije obiteljske medicine sklapaju ugovor s Hrvatskim zavodom za javno zdravstvo (HZZO), prema kojemu HZZO pokriva troškove zdravstvenih usluga plaćajući prema broju pacijenata registriranih u pojedinoj ordinaciji odnosno timu obiteljske medicine, tzv. glavarinu. S obzirom na obaveze i autonomiju postoje razlike između triju vrsta ordinacija obiteljske medicine:

- ordinacije unutar domova zdravlja (prostor i oprema u vlasništvu domova zdravlja, ugovor o plaćanju zdravstvenih usluga između domova zdravlja i HZZO-a, dodatno plaćene zdravstvene usluge nisu dozvoljene; ordinacija nema autonomiju),
- ordinacije u zakupu (prostor i oprema u vlasništvu domova zdravlja, liječnik plaća zakup, ugovor o plaćanju zdravstvenih usluga između liječnika obiteljske medicine i HZZO-a, dodatno plaćene zdravstvene usluge su moguće i dozvoljene; ordinacija ima ograničenu autonomiju),
- privatna ordinacija (prostor i oprema u vlasništvu liječnika obiteljske medicine,

- ugovor o plaćanju zdravstvenih usluga između liječnika obiteljske medicine iHZZO-a, dodatno plaćene zdravstvene usluge su moguće i dozvoljene; ordinacija ima autonomiju).

Nakon privatizacije mnogi su obiteljski liječnici počeli uvoditi informatička rješenja u svoje ordinacije. Obiteljski liječnici čije su ordinacije ostale u domovima zdravlja, za razliku od privatiziranih i onih u zakupu, nisu imali prilike odlučivati o tome koje informatičko rješenje nabaviti, jer je o tome odlučivao dom zdravlja. Postavlja se pitanje: kakva su informatička rješenja odnosno informacijski sustavi ušli u ordinacije, jesu li liječnici s njima zadovoljni, jesu li dovoljno vješti u korištenju informacijske i komunikacijske tehnologije i da li postoji razlika u informatizaciji između triju tipova ordinacija obiteljske medicine.

2. Informacijska i komunikacijska tehnologija u hrvatskom sustavu zdravstvene zaštite

Hrvatski je sustav zdravstvene zaštite započeo s informatizacijom još u ranim 60-tim godinama [2]. Podrazumijeva se da su tadašnje aplikacije bile primjerene tadašnjoj tehnologiji i mogućnostima koje je Hrvatska imala. Primarna zdravstvena zaštita organizirana unutar domova zdravlja kao i bolnice započeli su s informatizacijom administrativnih poslova još 1970. godine [3], [4]. Kasniji napori u tom smjeru temelje se na dokumentu Ministarstva zdravstva Republike Hrvatske "Strategija i plan reforme sustava zdravstvene zaštite i zdravstvenog osiguranja Republike Hrvatske" [5].

S obzirom na ulogu primarne zdravstvene zaštite u zdravstvenom sustavu Hrvatske, informatizacija primarne zdravstvene zaštite, posebno obiteljske medicine, dobila je posebni tretman i prvenstvo [6].

3. Sadašnje stanje informatizacije ordinacija obiteljske medicine

3.1. Osnova za analizu postojećeg stanja

U analizi stanja informatiziranosti ordinacija obiteljske medicine (vještine u korištenju informacijske tehnologije, medicinskoinformatička znanja i motiviranost doktora da prihvate, razvijaju i unapređuju informacijski sustav u svojoj ordinaciji) krenuli smo od ankete koju smo proveli među liječnicima na specijalizaciji iz obiteljske medicine za vrijeme trajanja njihovog poslijediplomskog obrazovanja na Medicinskom fakultetu Sveučilišta u Zagrebu u školskoj godini 2004/2005. Anketa je bila anonimna i provedena je nakon uvodnog predavanja iz predmeta Medicinska informatika. Anketa se sastojala iz triju modula: (a) pitanja o liječniku (dob, duljina radnog staža u ordinaciji, vrsta ordinacije - u zakupu, u domu zdravlja, privatna), (b) pitanja o uporabi informacijske i komunikacijske tehnologije, i, konačno, (c) informacijskom sustavu instaliranom u ordinaciji. Ovaj zadnji niz pitanja u anketi ispunjavali su samo oni liječnici koji imaju računalo u svojoj ordinaciji i neki oblik informacijskog sustava instaliran na računalu.

3.2. Rezultati

Tablica 1. opisuje uzorak liječnika na specijalizaciji iz obiteljske medicine tijekom poslijediplomskog studija u školskoj godini 2004/2005. Svi oni rade već niz godina kao liječnici primarne zdravstvene zaštite. Većina od 159 liječnika u uzorku, njih 117, imaju ordinaciju u zakupu, 28 rade u ordinacijama u sklopu domova zdravlja, a 14 liječnika imaju svoju privatnu ordinaciju. Primjećuje se da su ordinacije u sklopu domova zdravlja manje informatizirane. Liječnici su u takvim ordinacijama mlađi, prosječno 8 godina mlađi od ostalih u uzorku, i rade otprilike 10 godina kraće. Potpuno privatizirane ordinacije su pretežito informatizirane (91 %).

Liječnici iz svih triju grupa imaju jednak obrazac korištenja informacijske i komunikacijske tehnologije: na prvom je mjestu elektronička pošta, na drugom praćenje zdravstvenih portala. Zatim slijedi pretraživanje stručne medicinske literature (PubMed, Current Contents i sl.) i čitanje elektroničkih časopisa poput Croatian Medical Journal, British Medical Journal i sličnih. Razlike među trima grupama liječnika (iz ordinacija u zakupu, ordinacija u domovima zdravlja i privatnih ordinacija) u korištenju elektroničke pošte pokazale su se statistički značajnim ($p=0.043$). Pojam "korištenje izvora medicinskih znanja" uključuje čitanje e-časopisa, pretraživanje publikacija poput PubMed i sličnih, kao i praćenje informacija objavljivanim na zdravstvenim portalima. U skladu s tim, liječnici iz privatnih ordinacija najviše koriste e-izvore medicinskih znanja, njih 86 %. Na drugom su mjestu liječnici iz ordinacija unutar domova zdravlja, njih 79 %, dok liječnici iz ordinacija u zakupu najmanje koriste e-izvore medicinskog znanja (60 %). Slika 1 pokazuje učestalost korištenja e-izvora medicinskog znanja za svaku od triju tipova ordinacija.

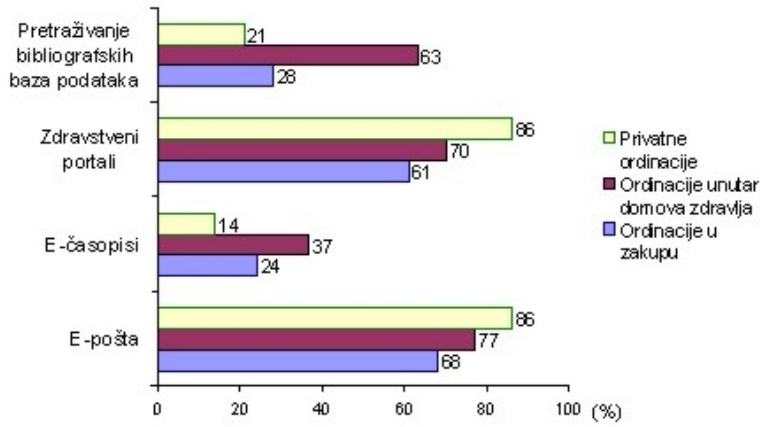
Tablica 1. Liječnici na specijalizaciji iz obiteljske medicine.

	Dob medijan (interkvartilni raspon)	Duljina staža medijan (interkvartilni raspon)	Ordinacija je informatizirana (%)
Ordinacije u zakupu	43 (41-46)	15 (12-19)	62
Ordinacije unutar domova zdravlja	35 (32-37)	5 (2-7)	42
Privatne ordinacije	43 (41-44)	15 (7-19)	91

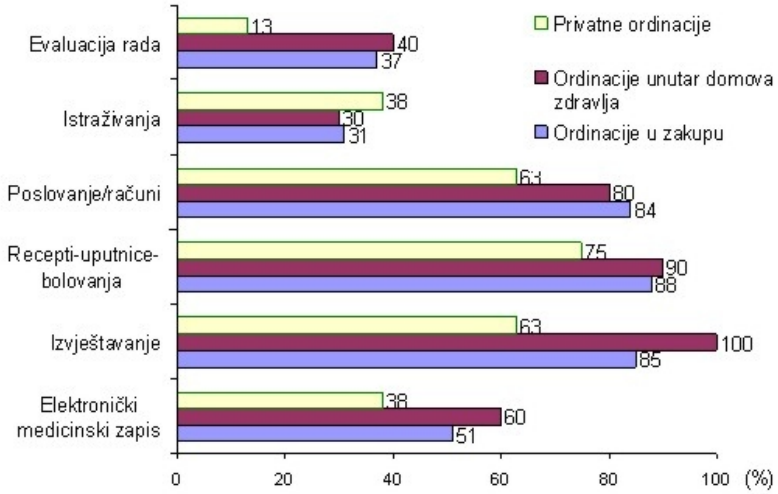
Informacijske sustave odnosno programska rješenja u ordinacijama liječnici koriste u prvom redu za administrativne potrebe: izvještavanje, propisivanje lijekova, bolovanja, upućivanja i poslovanje (Slika 2). Elektronički medicinski zapis ima oko 50% liječnika, ali u pravilu svi oni paralelno vode i papirnatu dokumentaciju. Razlog tome je i postojeće zakonodavstvo u Republici Hrvatskoj. Istraživanja koja se temelje na elektroničkom medicinskom zapisu provodi tek 32 % liječnika. Usprkos većoj proporciji liječnika iz privatnih ordinacija koji provode istraživanja utemeljena na elektroničkom obliku zapisa, nema statistički značajnih razlika između promatranih triju grupa liječnika. Isto tako nema statistički značajnih razlika niti u učestalosti korištenja e-zapisa za evaluaciju vlastitog rada (35 % od ukupno 97 liječnika izjavljuje da to čini).

Zadovoljstvo s postojećom informatizacijom je više nego skromno. Tek 19 % liječnika u informatiziranim ordinacijama izražava zadovoljstvo sa stanjem informatiziranosti. Informacijske su potrebe zadovoljene uglavnom tek parcijalno. 32 % liječnika izjavljuje da im sustav osigurava potrebne informacije, 40 % misli da uz informacijski sustav rade učinkovitije, a 25 % smatra da su i njihovi pacijenti zadovoljniji otkada je u ordinaciji instaliran informacijski sustav. Promatrajući različite tipove ordinacija (Slika 3) s obzirom na opće zadovoljstvo liječnika s instaliranim informacijskim sustavom u njegovoj ordinaciji, mogu se primijetiti razlike. Liječnici u ordinacijama unutar domova zdravlja, a to su uglavnom mlađi liječnici, manje su zadovoljni nego ostale dvije skupine liječnika.

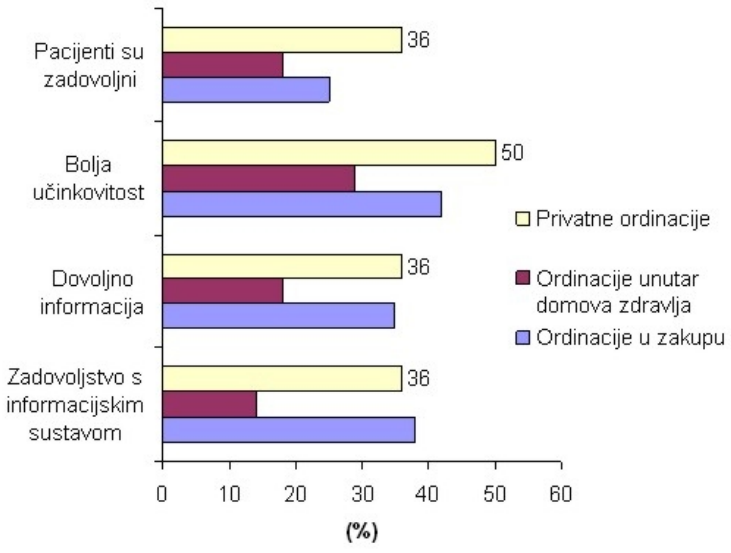
Sigurnost podatka u informacijskom sustavu procjenjivala se s obzirom na a) uporabu lozinke, b) fizičku zaštitu i c) dnevno arhiviranje podataka. Pokazale su se statistički značajne razlike u fizičkoj zaštiti podataka (p=0.028), ordinacije unutar domova zdravlja imaju najslabiju fizičku zaštitu, dok se lozinke i dnevno arhiviranje prakticira podjednako u sva tri tipa ordinacija.



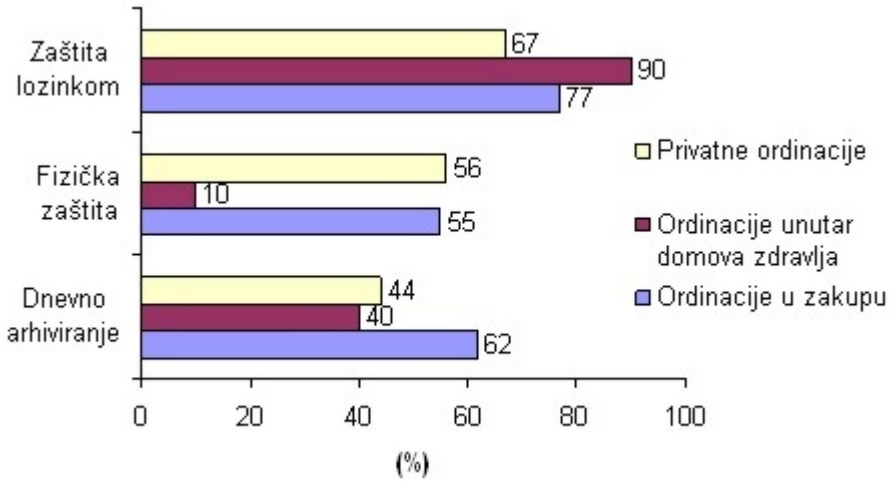
Slika 1. Uporaba informacijske i komunikacijske tehnologije od strane liječnika na specijalizaciji iz obiteljske medicine.



Slika 2. Korištenje informacijske i komunikacijske tehnologije u ordinacijama obiteljske medicine.



Slika 3. Zadovoljstvo s informacijskim sustavom - procjena liječnika.



Slika 4. Zaštita podataka unutar informacijskog sustava.

4. Diskusija

Istraživanje je pokazalo da hrvatski liječnici koriste informacijsku i komunikacijsku tehnologiju manje nego njihovi kolege iz razvijenijih zemalja [7]. Obrazac korištenja informatičkih aplikacija u obiteljskoj medicini u Hrvatskoj je vrlo sličan onom u razvijenim zemljama – koristi se u prvom redu za izvještavanje i administraciju [7]. Informacije nužne za zdravstvenu zaštitu odnosno liječenje, nemaju prioritet prilikom informatizacije. Slično je i s informacijama nužnim za istraživanja i evaluaciju rada. Za sve to, naravno, osnova je elektronički medicinski zapis. Zasad međutim hrvatski zakoni zahtijevaju klasični papirnati zapis, pa kao posljedicu imamo da liječnici vode dvostruku dokumentaciju. Dupliciranje posla ne ohrabruje liječnika da upotrebljava i unapređuje elektronički zapis. Zato su elektronički zapisi uglavnom nestrukturirani, uglavnom je to slobodni tekst što stvara probleme prilikom analize podataka.

Zadovoljstvo s postojećim informacijskim sustavima u ordinacijama je prilično malo. Samo su liječnici u privatnim ordinacijama zadovoljni sa svojom učinkovitošću pri uporabi informacijskog sustava. Istovremeno oni pozitivno ocjenjuju zadovoljstvo pacijenata što se liječe u informatiziranoj ordinaciji.

Zaštita i sigurnost informacijskih sustava može se diskutirati s različitih stajališta. Avery i suradnici pišu o potrebi koordinacije najvažnijih elemenata sigurnosti sa stanovišta kliničke prakse (razna upozorenja, ponovljeno propisivanje i sl.) [8]. Honyeman i suradnici diskutiraju potencijalni pristup pacijenta svome elektroničkom medicinskom zapisu [9]. U eri novih tehnologija kao što je primjerice Grid-tehnologija koja omogućava razvoj distribuiranih elektroničkih zapisa, nužno je posebnu pozornost obratiti upravo zaštiti, sigurnosti i etičkim aspektima [10], [11]. Zaštita informacija u informacijskim sustavima u obiteljskoj medicini u Hrvatskoj nije dovoljno dobra. Temeljem procjena samih liječnika, manje od 50 % liječnika štiti svoje elektroničke informacije na zadovoljavajući način. Međutim, to je tek vrlo gruba procjena.

S obzirom na rezultate ovog istraživanja, liječnici koji rade u ordinacijama obiteljske medicine nisu dovoljno vješti u korištenju informacijskih i komunikacijskih tehnologija u svom profesionalnom radu. Iako internet smatraju vrlo korisnim (kao i u razvijenim zemljama), koriste elektroničku poštu i surfaju internetom, oni još uvijek nedovoljno čitaju stručne časopise u elektroničkom obliku kao i sekundarne publikacije. Međutim, to je slično kao i u razvijenim zemljama [12]. Postoji doduše pozitivni pomak među mladim liječnicima zaposlenim u domovima zdravlja, u ordinacijama koje još nisu privatizirane, što se može smatrati ohrabrujućim jer stvari idu pozitivnim tokom.

S druge strane neki liječnici investiraju mnogo u suvremenu tehnologiju kako bi unaprijedili svoj posao i odnos prema pacijentima. Više je pozitivnih primjera edukacijskih tečajeva koje organiziraju medicinski fakulteti, Hrvatska liječnička komora, ali i kompanije koje proizvode i/ili prodaju takvu tehnologiju. Web-tehnologija se često koristi pri predstavljanju ordinacije. Ima međutim i primjera liječnika koji tu tehnologiju koriste da bi se približili pacijentima informirajući ih o njihovu zdravlju i učeći ih kako ostati zdrav [13].

Zahvala

Autori se zahvaljuju doktorici Milici Katić, profesorici obiteljske medicine na Medicinskom fakultetu Sveučilišta u Zagrebu koja je pomogla objasniti specifičnosti organizacije obiteljske medicine u Hrvatskoj.

[1] Končar M., Gvozdanović D.: Primary Healthcare Information System – the Cornerstone for the Next Generation Healthcare Sector in Republic of Croatia. Int J Med Inform 2006; 75, pp.306-314.

[2] Kern J., Strnad M.: Informatics in the Croatian Health Care System. Acta Med Croatica 2005; 59, pp. 161-168. [in Croatian]

[3] Golec B, Krajačić S.: Project of Automated Data Processing for Outpatient Health Organization. Health Centre Remetinec-Zagreb. Zagreb, Centre for economic development of the city of Zagreb, 1980. [in Croatian].

[4] Rosandić D.: Report on Introducing Computerization in the General Hospital "Dr. J.Kajfeš" – Zagreb, Zagreb, Commision for computerization of health care system 1975. [in Croatian].

[5] Reform of Healthcare System: Strategy and Plan of Reform of Health and Health Insurance System in Croatia. Zagreb, Ministry of Health 2002. [in Croatian]

[6] Stevanović R., Stanić A., Varga S.: Information System in Primary Health Care, Acta Med Croatica 2005; 59, pp. 209-212. [in Croatian].

[7] Western M. C., Dwan K. M., Western J. S., Makkai T., Del Mar C.: Computerisation in Australian General Practice. Aust Fam Physician 2003;32, pp. 180-5.

[8] Avery A.J.,.Savelyich B.S, Sheikh A., Cantrill J., Morris C.J., Fernando B., Bainbridge M., Horsfield P., Teasdale S.: Identifying and Establishing Consensus on the Most Important Safety Features of GP Computer Systems: e-Delphi Study, Inform Prim Care 13 (2005) 3-12.

[9] Honeyman A., Cox B., Fisher B.:.Potential Impacts of Patient Access to Their Electronic Care Records, Inform Prim Care 2005; 13, pp. 55-60.

[10] Kalra D., Singleton P., Milan J., Mackay J., Detmer D., Rector A., Ingram D.: Security and Confidentiality Approach for the Clinical E-Science Framework (CLEF), Methods Inf Med 2005; 44, pp. 193-197.

[11] Claerhout B., De Moor G. J.: Privacy Protection for HealthGrid Applications. Methods Inf Med 2005; 44, pp. 140-143.

[12] Bennett N. L., Casebeer L. L., Kristofco R., Collins B. C.: Family Physicians' Information Seeking Behaviors: a Survey Comparison with Other Specialties. BMC Med Inform Decis Mak 2005; 5, p. 9.

[13] http://www.ordinacije-lazic.hr/. Accessed on August 14, 2006.

EJBI 2007 ISSN 1801 - 5603

European Journal for Biomedical Informatics

Volume 3 (2007), Issue 1

Polish version

www.ejbi.eu

EuroMISE s.r.o.

Content

Polish version

Rangowe modelowanie sekwencji chorób wątroby

Leon Bobrowski[1,2], Tomasz Łukaszuk[1], Hanna Wasyluk[3]
[1]Białystok Technical University, Faculty of Computer Science, Poland,
[2]Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw, Poland,
[3]Medical Center of Postgraduate Education, Warsaw, Poland

Streszczenie: Model rangowy w postaci liniowej transformacji wielowymiarowych wektorów cech na linię prostą może odzwierciedlać relacje przyczynowo-skutkowe pomiędzy chorobami wątroby. Wiedza medyczna o takich relacjach pomiędzy chorobami wątroby została uwzględniona w definiowaniu wypukłych i odcinkowo-liniowych (CPL) funkcji kryterialnych. Liniowe transformacje rangowe są projektowane poprzez minimalizacje funkcji kryterialnych typu CPL.

Słowa kluczowe: wzorce sekwencyjne, rangowe transformacje liniowe, wypukłe i odcinkowo-liniowe (CPL) funkcje kryterialne, liniowa separowalność zbiorów danych, sekwencja chorób wątroby.

1. Wstęp

Odkrywanie regularności w wielowymiarowych zbiorach danych lub w bazach danych jest jednym z głównych celów analizy eksploracyjnej lub metod rozpoznawania obrazów [1], [2]. Odkrywanie trendów w temporalnych bazach danych jest szczególnie interesującym problemem związanym z wieloma ważnymi zastosowaniami.

Odpowiednia standaryzacja danych jest wymagana przed stosowaniem narzędzi analizy eksploracyjnejdanych. Dane kliniczne w standaryzowanej postaci reprezentowane są jako wektory cech o tej samej wymiarowości lub jako punkty w wielowymiarowej przestrzeni cech. Poszczególne wzorce w danych są reprezentowane jako separowalne zbiory wektorów cech lub jako konstelacje punktów w przestrzeni cech.

Metody analizy regresyjnej grają znaczącą rolę w eksploracji danych [3]. Model regresyjny może opisywać zależność jednej cechy od wybranego podzbioru innych cech. Metoda regresji rangowej może być zastosowana do podobnych celów [4], [5], [6]. Modele rangowe są szczególnie użyteczne wtedy, gdy wartości cechy zależnej nie mogą być zmierzone precyzyjnie. Zamiast wartości cechy zależnej może być dostępna tylko informacja, że wartość rozpatrywanej cechy jest większa u wybranego pacjenta w porównaniu z drugim pacjentem. Na tej podstawie budowane są rangowe (porządkowe) relacje pomiędzy wektorami cech reprezentującymi wybrane pary pacjentów. Takie relacje porządkowe mogą być traktowane jako wiedza priori o wzorcach ukrytych w danych. W tym kontekście, indukcja modeli rangowych ze zbiorów danych może być traktowana jako problem rozpoznawania obrazów. Indukowane modele rangowe mogą być stosowane dla celów prognozy lub wspierania decyzji.

Metoda indukcji liniowych modeli rangowych ze zbiorów wektorów cech wraz z relacjami porządkowych pomiędzy wybranymi pacjentami została opisana we wcześniejszych publikacjach [4] [5]. Metoda ta bazuje na minimalizacji wypukłych i odcinkowo-liniowych (CPL) funkcji kryterialnych. Właściwości tej metody w kontekście analizowania przyczynowej sekwencji chorób wątroby są analizowane w prezentowanej publikacji. Wektory cech z hepatologicznej bazy danych sytemu Hepar oraz wiedza dodatkowa w postaci sekwencji przyczynowo-skutkowej chorób wątroby zostały użyte w projektowaniu liniowej transformacji rangowej [7].

2. Wektory cech oraz zorientowane dipole

Weźmy pod uwagę zbiór danych C zbudowany z *m* wektory cech $\mathbf{x}_j=[x_{j1},...x_{jn}]^T$ które są indeksowane w ustalony sposób:

$$C=(\mathbf{x}_j)\,(j=1,...,m) \qquad (1)$$

Wektory $\mathbf{x}_j$ należą do *n*-wymiarowej przestrzeni cech. $F[n]$ $(x_j \in F[n])$. Składowa x_{ji} wektora $\mathbf{x}_j$ jest wartością liczbową pomiaru diagnostycznego *i*-tej cechy $\mathbf{x}_i$ $(i=1,...,n)$ pacjenta O_j $(j=1,...,m)$. Wektory cech $\mathbf{x}_j$ są typu mieszanego jeżeli reprezentują różnego typu pomiary diagnostyczne $(x_i \in \{0,1\})$ lub $(x_i \in R)$.

Niech symbol "$\prec$" oznacza relację "*follows*" ("następuje po"), która jest spełniona wewnątrz uporządkowanych par $\{\mathbf{x}_j, \mathbf{X}_j'\}$ $(j<j')$ wektorów cech $\mathbf{x}_j$ i $\mathbf{x}_j'$ z indeksami z pewnego zbioru *J*:

$$(\forall (j,j') \in J) \quad \mathbf{x}_j \prec \mathbf{x}_{j'} \Leftrightarrow (\mathbf{x}_{j'} \;\; follows \;\; \mathbf{x}_j)$$
$$or \quad \mathbf{x}_{j'} \prec \mathbf{x}_{j'} \Leftrightarrow (\mathbf{x}_j \;\; follows \;\; \mathbf{x}_{j'}) \qquad (2)$$

Relacja $\mathbf{x}_j \prec \mathbf{x}_j'$ pomiędzy wektorami cech $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ oznacza, że wektor $\mathbf{x}_{j'}$ następuje po wektorze $\mathbf{x}_j$ w pewnej sekwencji. Ta relacja powinna być określona na bazie pewnej informacji dodatkowej o zbiorze wybranych (niekoniecznie wszystkich) par wektorów cech $\mathbf{x}_j$. Na przykład, lekarz może porównać dwu pacjentów i stwierdzić, że jeden z nich znajduje się w bardziej zaawansowanym stadium badanego schorzenia.

W przedstawionej pracy analizowany jest problem projektowania takich liniowych transformacji wektorów cech $\mathbf{x}_j$ na (*rangową*) prostą $y=\mathbf{w}^T\mathbf{x}$, która zachowuje w możliwie dużym stopniu relację "$\prec$" (2).

$$y_j = y_j(\mathbf{w}) = \mathbf{w}^T \mathbf{x}_j, \qquad (3)$$

gdzie $\mathbf{w}=[w_1,...,w_n]^T$ jest wektorem wag.

Rodzina relacji (2) definiuje pewien *wzorzec sekwencyjny* S($\mathbf{x}$) wektorów $\mathbf{x}_j$ w przestrzeni cech $F[n]$ $(x_j \in F[n])$.

Definicja 1: Wzorzec sekwencyjny $S(\mathbf{x})$ jest liniowy w przestrzeni cech $F[n]$ wtedy i tylko wtedy, gdy istnieje taki n-wymiarowy wektor wag $\mathbf{w}$ ($\mathbf{w} \in R^n$), że spełnione są poniższe implikacje:

$$(\forall (j,j') \in J) \quad \mathbf{x}_j \prec \mathbf{x}_{j'} \Rightarrow \mathbf{w}^T\mathbf{x}_j < \mathbf{w}^T\mathbf{x}_{j'} \quad and \quad \mathbf{x}_{j'} \prec \mathbf{x}_j \Rightarrow \mathbf{w}^T\mathbf{x}_{j'} < \mathbf{w}^T\mathbf{x}_j \tag{4}$$

gdzie J jest zbiorem par indeksów (j, j') uporządkowanych par wektorów cech $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j<j'$).

Procedura odkrywania liniowych wzorców sekwencyjnych $S(\mathbf{x})$ oraz projektowania odwzorowań rangowych może być oparta na koncepcji dipoli $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$($j<j'$) dodatnio lub ujemnie zorientowanych [4], [5].

Definicja 2: Uporządkowana para $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j<j'$) wektorów cech $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ tworzy *dodatnio zorientowany dipol* $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($(j,j' \in J^+$), wtedy i tylko wtedy, gdy $\mathbf{x}_j \prec \mathbf{x}_{j'}$.

$$(\forall (j,j') \in J^+) \quad \mathbf{x}_j \prec \mathbf{x}_{j'} \tag{5}$$

Definicja 3: Uporządkowana para $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j<j'$) wektorów cech $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ tworzy *ujemnie zorientowany dipol* $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($(j,j' \in J$), wtedy i tylko wtedy, gdy $\mathbf{x}_j \prec \mathbf{x}_{j'}$.

$$(\forall (j,j') \in J^-) \quad \mathbf{x}_{j'} \prec \mathbf{x}_j \tag{6}$$

Definicja 4: Odwzorowanie liniowe $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) jest w pełni rangowe wtedy i tylko wtedy, gdy

$$(\forall (j,j') \in J^+) \quad y_j(\mathbf{w}) < y_{j'}(\mathbf{w}) \quad and \quad (\forall (j,j') \in J^-) \quad y_j(\mathbf{w}) > y_{j'}(\mathbf{w}) \tag{7}$$

gdzie J^+ i J^- są zbiorami par indeksów (j, j') odpowiednio pozytywnie oraz negatywnie zorientowanych dipoli $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j<j'$), gdzie $J^+ \cup J^- = J$, $J^+ \cap J^- = \emptyset$.

3. Designing ranked models through minimization of a CPL criterion function

Wprowadźmy zbiór dodatni R^+ oraz zbiór ujemny R^- różnic wektorów cech $\mathbf{r}_{jj'}=(\mathbf{x}_{j'}-\mathbf{x}_j)$ utworzone odpowiednio na bazie zbiorów indeksów J^+ (6) oraz J^- (7).

$$R^+ = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j,j') \in J^+\} \quad R^- = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j,j') \in J^-\} \tag{8}$$

Interesuje nasz możliwość separacji zbiorów R^+ i R^- za pomocą hiperpłaszczyzny $H(\mathbf{w})$ przechodzącej przez początek układu współrzędnych przestrzeni cech $F[n]$.

$$H(\mathbf{w}) = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = 0\} \tag{9}$$

Definicja 5: Zbiory R^+ i R^- (8) są separowalne przez hiperpłaszczyznę $H(\mathbf{w})$ (9) wtedy i tylko wtedy, gdy zachodzą poniższe nierówności

$$(\exists \mathbf{w}) \quad (\forall (j,j') \in J^+) \quad \mathbf{w}^T\mathbf{r}_{jj'} > 0 \quad (\forall (j,j') \in J^-) \quad \mathbf{w}^T\mathbf{r}_{jj'} < 0 \tag{10}$$

Jeżeli wszystkie powyższe nierówności są spełnione dla pewnego wektora $\mathbf{w}$, wtedy hiperpłaszczyzna $H(\mathbf{w})$ (9) separuje zbiory R^+ i R^- (8).

Lemat 1: Odwzorowanie liniowe $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) jest w pełni rangowe (7) wtedy i tylko wtedy, gdy hiperpłaszczyzna $H(\mathbf{w})$ (9) separuje (9) zbiory R^+ i R^- (8).

Dowód: Jeżeli hiperpłaszczyzna H(w) (9) separuje (9) zbiory R^+ i R^- (8), to wtedy wszystkie rangowe nierówności (7) są zachowane ma prostej $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3). Z drugiej strony, spełnienie wszystkich nierówności (7) zapewnia spełnienie relacji (10).

Projektowanie separującej hiperpłaszczyzny $H(\mathbf{w})$ (9) może być oparte na minimalizacji wypukłej i odcinkowo-liniowej (*CPL*) funkcji kryterialnej, która jest podobna do perceptronowej funkcji kryterialnej [2]. Wprowadźmy w tym celu pozytywna $\varphi_{jj'}^+(\mathbf{w})$ i $\varphi_{jj'}^-(\mathbf{w})$ (w) negatywną funkcje kary:

$$(\forall (j,j') \in J^-) \quad \varphi_{jj'}^-(\mathbf{w}) = \begin{cases} 1+\mathbf{w}^T\mathbf{r}_{jj'} & if \quad \mathbf{w}^T\mathbf{r}_{jj'} > -1 \\ 0 & if \quad \mathbf{w}^T\mathbf{r}_{jj'} \le -1 \end{cases} \tag{11}$$

$$(\forall (j,j') \in J^-) \quad \varphi_{jj'}^-(\mathbf{w}) = \begin{cases} 1+\mathbf{w}^T\mathbf{r}_{jj'} & if \quad \mathbf{w}^T\mathbf{r}_{jj'} > -1 \\ 0 & if \quad \mathbf{w}^T\mathbf{r}_{jj'} \le -1 \end{cases} \tag{12}$$

Funkcja kryterialna $\Phi(\mathbf{w})$ jest dodatnio ważoną sumą funkcji kary $\varphi_{jj'}^+(\mathbf{w})$ i $\varphi_{jj'}^-(\mathbf{w})$:

$$\Phi(\mathbf{w}) = \sum_{(j,j') \in J^+} \gamma_{jj'} \varphi_{jj'}^+(\mathbf{w}) + \sum_{(j,j') \in J^-} \gamma_{jj'} \varphi_{jj'}^-(\mathbf{w}) \tag{13}$$

gdzie $\gamma_{jj'}$ ($\gamma_{jj'}>0$) jest dodatnim parametrem (*cena*) związanym z dipolem $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j<j'$).

$\Phi(\mathbf{w})$ (13) jest wypukłą i odcinkowo-liniową (*CPL*) funkcją jako suma tego typu funkcji $\varphi_{jj'}^+(\mathbf{w})$ i $\varphi_{jj'}^-(\mathbf{w})$. Algorytmy wymiany rozwiązań bazowych, zbliżone do programowania liniowego, pozwalają znaleźć minimum funkcji Φ(w) w sposób efektywny nawet w przypadku dużych, wielowymiarowych zbiorów R^+ ai R^- (8) [7]:

$$\Phi^* = \Phi(\mathbf{w}^*) = \min_{\mathbf{w}} \Phi(\mathbf{w}) \ge 0 \tag{14}$$

Optymalny wektor parametrów $\mathbf{w}^*$ oraz wartość minimalna Φ^* funkcji kryterialnej $\Phi(\mathbf{w})$ (13) może być stosowana w rozwiązywaniu wielu problemów rangowej analizy danych. W szczególności, wektor w* wyznaczający najlepszą linię rangową $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) może być znaleziony w ten sposób.

Wartość minimalna Φ^* funkcji kryterialnej $\Phi(\mathbf{w})$ (13) może być używana jako miara stopnia liniowości wzorca sekwencyjnego (Def. 1) w danej przestrzeni cech .

Lemat 2: Wartość minimalna Φ^* funkcji kryterialnej $\Phi(\mathbf{w})$ (13) jest równa zeru wtedy i tylko wtedy, gdy wzorzec sekwencyjny $S(\mathbf{x})$ (Def. 1) jest liniowy.

Dowód: Jeżeli istnieje taki wektor $\mathbf{w}^*$, że kolejność punktów $y_j(\mathbf{w}^*)$ na linii $y(\mathbf{w}^*)=(\mathbf{w}^*)^T x$ jest w pełni zgodna (7) z relacjami "$\prec$", wtedy zbiory R^+ i R^- (8) są odseparowane (10) przez hiperpłaszczyznę $H(\mathbf{w}^*)$ (9). W tym przypadku, wartość minimalna Φ^* funkcji $\Phi(\mathbf{w})$ (13) jest równa zeru jak to wynika z teorii rozpoznawania obrazów [1]. Z drugiej strony, jeżeli wartość (14) funkcji kryterialnej (13) jest równa zeru w punkcie $\mathbf{w}^*$, wtedy wartości nieujemnych funkcji kary $\varphi_{jj'}^+(\mathbf{w})$ i $\varphi_{jj'}^-(\mathbf{w})$ muszą być także równe zeru.

To oznacza, że zbiory R^+ i R^- (8) mogą być odseparowane (10) przez hiperpłaszczyznę $H(\mathbf{w}^*)$ (9). W rezultacie, kolejność punktów $y_j(\mathbf{w}^*)$ na linii $y(\mathbf{w}^*)=(\mathbf{w}^*)^T x$ jest w pełni zgodna (7) z relacjami "$\prec$"(5) i (6).

4. Sekwencja zbiorów uczących

Załóżmy, że kliniczna baza danych zawiera opisy m pacjentów $O_j(k)$ (j=1,...m) etykietowanych zgodnie z ich diagnozą kliniczną ω_k(k=1,...K). Każdy pacjent $O_j(k)$ jest reprezentowany przez n-wymiarowy wektor cech $\mathbf{x}_j(k)$. Wektor cech $\mathbf{x}_j(k)$ reprezentuje j-tego pacjenta $O_j(k)$ przypisanego do k-tej jednostki chorobowej ω_k. Zbiór uczący C_k zawiera mk etykietowanych wektor cech $\mathbf{x}_j(k)$, które zostały przypisane do k-tej jednostki chorobowej (klasy) ω_k:

$$C_k = \{x_j(k)\} \quad (j \in I_k) \tag{15}$$

gdzie I_k jest zbiorem indeksów j wektorów cech $\mathbf{x}_j(k)$ przypisanych do k-tej klasy ω_k.

Zbiory uczące C_k zostały uszeregowane w poniższą sekwencję przyczynowo-skutkową:

$$C_1 \rightarrow C_2 \rightarrow ... \rightarrow C_{K-1} \rightarrow C_K \tag{16}$$

gdzie symbol "$C_{k-1} \rightarrow C_k$" oznacza, że "schorzenie ω_k następuje po ω_{k-1}" lub "schorzenie jest przyczyną". W sekwencji (16) zostało zastosowane spójne indeksowanie zbiorów C_k oraz schorzeń ωk. Oznacza to, że:

$$(\forall k,k' \in \{1,...,K\}) \quad (k < k') \Rightarrow (C_k \rightarrow C_{k'}) \tag{17}$$

Relacja przyczynowa "$C_k \rightarrow C_{k'}$" (17) pomiędzy zbiorami uczącymi C_k i $C_{k'}$ może być użyta w celu określenia relacji "$\prec$" (2) pomiędzy wektorami cech $\mathbf{x}_j(k)$ ($\mathbf{x}_j(k) \in C_k$) i $x_{j'}(k')$ ($x_{j'}(k') \in C_{k'}$) (15):

$$(\forall k,k' \in \{1,...,K\}) \quad (C_k \rightarrow C_{k'}) \Rightarrow ((\forall \mathbf{x}_j(k) \in C_k) \; and \; (\forall \mathbf{x}_{j'}(k') \in C_{k'})) \; \mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k') \tag{18}$$

lub

$$(\forall k,k' \in \{1,...,K\}) \quad (k < k') \Rightarrow (\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')) \tag{19}$$

Zauważmy, że nie istnieje rangowa relacja "$\prec$" (2) pomiędzy wektorami cech $\mathbf{x}_j(k)$ i $x_{j'}(k)$ z tego samego zbioru uczącego C_k.

Możemy założyć, że indeksy j wektorów cech $\mathbf{x}_j(k)$ są spójne ze zbiorami uczącymi C_k (15). Oznacza to, że zbiór C_1 zawiera m_1 pierwszych wektorów cech $\mathbf{x}_j(k)$, zbiór C_2 zawiera m_2 kolejnych wektorów $\mathbf{x}_j(k)$, itd. W konsekwencji, zachodzi poniższa relacja spójnego indeksowania:

$$(\forall \mathbf{x}_j(k) \in C_k) and (\forall \mathbf{x}_{j'}(k') \in C_{k'}) \quad (\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')) \Rightarrow (j < j') \tag{20}$$

Lemat 3: W przypadku spójnego indeksowania (20), linia $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) jest w pełni rangowa (7) wtedy i tylko wtedy, gdy zbiór R^+ (8) różnic wektorów cech $\mathbf{r}_{jj'} = (\mathbf{x}_{j'}-\mathbf{X}_j)$ jest usytuowany po dodatniej stronie hiperpłaszczyzny $H(\mathbf{w})$ (9):

$$(\exists \mathbf{w})(\forall \mathbf{r}_{jj'} \in R^+) \quad \mathbf{w}^T \mathbf{r}_{jj'} > 0 \tag{21}$$

Dowód: Relacje (19) i (20) gwarantują, że wszystkie dipole {$\mathbf{x}_j$, $\mathbf{x}_{j'}$} dodatnio zorientowane (*Def. 2*) oraz, że zbiór negatywny R^- (8) jest pusty (R^-=ø). W rezultacie wektor $\mathbf{w}$ definiuje taką hiperpłaszczyznę $H(\mathbf{w})$, która separuje (10) zbiory R^+ i R^- (8). Oznacza to, że założenia *Lematu 1* są spełnione.

Załóżmy, że zbiór R^+ (8) jest kompletny tj. zawiera wszystkie dodatnio zorientowane dipole {$\mathbf{x}_j$, $\mathbf{x}_{j'}$}, które mogą być wygenerowane z dwu zbiorów uczących C_1 i C_2 (15) zgodnie z relacją $C_1 \rightarrow C_2$ (16) i przy spójnym indeksowaniu (20)

$$R^+ = \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'}(2) - \mathbf{x}_j(1)) : \mathbf{x}_j(1) \in C_1 \; and \; \mathbf{x}_{j'}(2) \in C_2\} \tag{22}$$

Definicja 6: Dwa zbiory uczące C_1 i C_2 (15) są liniowo separowalne wtedy i tylko wtedy, gdy słuszne są poniższe nierówności:

$$(\exists \mathbf{w}, \theta) \quad \begin{matrix} (\forall \mathbf{x}_{j'} \in C_2) & \mathbf{w}^T \mathbf{x}_{j'} > \theta \\ (\forall \mathbf{x}_j \in C_1) & \mathbf{w}^T \mathbf{x}_{j'} < \theta \end{matrix} \tag{23}$$

gdzie θ (θ $\in R^1$) jest *progiem*.

Powyższe parametry definiują hiperpłaszczyznę $H(\mathbf{w},\theta)$ w przestrzeni cech, gdzie:

$$H(\mathbf{w},\theta) = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} = \theta\} \tag{24}$$

Liniową separowalność (23) dwu zbiorów uczących C_1 i C_2 (15) można związać z pełną rangowością (Def. 4) odwzorowania liniowego $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3).

Twierdzenie 1: Odwzorowanie liniowe $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) jest w pełni rangowe (7) zgodnie z kompletnym zbiorem R^+ (22) wtedy i tylko wtedy, gdy istnieje taki próg θ', że hiperpłaszczyznę $H(\mathbf{w},\theta')$ (24) separuje dwa zbiory uczące C_1 i C_2 (15).

Dowód: Jeżeli odwzorowanie liniowe $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$ (3) jest w pełni rangowe (7), to

$$(\exists \mathbf{w})(\forall \mathbf{x}_j \in C_1) and (\forall \mathbf{x}_{j'} \in C_2) \quad \mathbf{w}^T \mathbf{x}_{j'} > \mathbf{w}^T \mathbf{x}_j \tag{25}$$

Zdefiniujmy dodatni $\theta^+(\mathbf{w})$ oraz negatywny próg $\theta^-(\mathbf{w})$ na linii prostej $y(\mathbf{w})=\mathbf{w}^T\mathbf{x}$:

$$\theta^+(\mathbf{w}) = \min\{\mathbf{w}^T \mathbf{x}_{j'} : \mathbf{x}_{j'} \in C_2\} \tag{26}$$

$$\theta^-(\mathbf{w}) = \max\{\mathbf{w}^T \mathbf{x}_j : \mathbf{x}_j \in C_1\} \tag{27}$$

Poniższa nierówność wynika z relacji (25)

$$\theta^+(\mathbf{w}) > \theta^-(\mathbf{w}) \tag{28}$$

Próg θ' może być zdefiniowany jak poniżej:

$$\theta' = \frac{\theta^+(\mathbf{w}) + \theta^-(\mathbf{w})}{2} \tag{29}$$

Można bezpośrednio zweryfikować, że hiperpłaszczyzna $H(\mathbf{w},\theta')$ (24) separuje zbiory uczące C_1 i C_2 (15).

Z drugiej strony, hiperpłaszczyzna $H(\mathbf{w},\theta')$ separującą zbiory C_1 i C_2 (15) wyznacza taką linię $y(\mathbf{w})=\mathbf{w}$ (3), która jest w pełni rangowa (7). Jak wynika z definicji liniowej separowalności (23), każdy element $\mathbf{r}_{jj'} = \mathbf{x}_{j'}(2) - \mathbf{x}_j(1)$ zbioru R^+ (22) spełnia relację (21).

5. Przyczynowa sekwencja chorób wątroby

Baza danych systemu *Hepar* zawiera opisy pacjentów z przewlekłymi chorobami wątroby ω_k (k=1,...K) [7].

Wektory cech $\mathbf{x}_j$ z tej bazy danych są typu mieszanego, jakościowo-ilościowego. Wektory te zawierają zarówno liczbowe rezultaty testów laboratoryjnych ($x_i \in R$) jak również oznaki i symptomy danego pacjenta ($x_i \in \{0,1\}$). Każdy z wektorów $\mathbf{x}_j$ cech zawierał stałą liczbę n składowych (cech) $\mathbf{x}_i$, gdzie n ≈ 200. W oparciu o wiedzę lekarzy część cech xi została pominięta i w rezultacie wektory cech $\mathbf{x}_j(k)$ przypisane poszczególnym jednostkom chorobowym ωk miały wymiarowość n=62. Wyszczególnione poniżej jednostki chorobowe ω_k zostały uwzględnione przy formowaniu siedmiu (K=7) zbiorów uczących C_k. (15) z bazy danych systemu *Hepar*:

C_1 *Non hepatitis patients* - 16 patients
C_2 *Hepatitis acuta* - 8 patients
C_3 *Hepatitis persistens* - 44 patients
C_4 *Hepatitis chronica activa* - 95 patients
C_5 *Cirrhosis hepatitis compensata* - 38 patients
C_6 *Cirrhosis decompensata* - 60 patients
C_7 *Carcinoma hepatis* - 11 patients

Total: 272 patients (30)

Zbiory danych C_k (30) tworzą sekwencję przyczynowo-skutkową $C_1 \rightarrow C_2 \rightarrow ..... \rightarrow C_7$ (16), zgodną ze specjalistyczną wiedzą medyczną. Na bazie tej sekwencji została zbudowana relacja rangowa "≺" (2) pomiędzy wektorami cech (18). Taka relacja rangowa pozwoliła zarówno zdefiniować zorientowane dipole $\{x_i, x_j\}$ (5) (6) jak również zbiór dodatni R^+ i zbiór ujemny R^- (8) różnic wektorów $\mathbf{r}_{ji}(\mathbf{x}_j\text{-}\mathbf{x}_i)$. Zbiory R^+ i R^- zostały użyte w konstrukcji wypukłej i odcinkowo-liniowej (*CPL*) funkcji kryterialnej Φ(**w**) (13). Optymalny wektor parametrów **w*** (14), tworzący minimum funkcji kryterialnej Φ(**w**) (13) pozwala na zdefiniowanie modelu rangowego (3), który może być używany dla celów prognostycznych:

$$y_j = y_j(\mathbf{w}^*) = (\mathbf{w}^*)^T \mathbf{x}_j = w_1^* x_{j1} + ... + w_n^* x_{jn} \quad (31)$$

Procedury selekcji cech pozwalają określić takie cechy x_i, które są najbardziej znaczące w prognozowaniu dalszych stanów pacjenta, oraz pominąć cechy nieistotne. Rozwiązanie problemu selekcji cech może być także oparte na minimalizacji wypukłej i odcinkowo-liniowej (*CPL*) funkcji kryterialnej Φ(**w**) (13) [7].

Model liniowy spełnia relacje rangowe (4) dla dużej części wektorów cech $\mathbf{x}_j$:

$$\mathbf{x}_j \prec \mathbf{x}_{j'} \Rightarrow (\mathbf{w}^*)^T \mathbf{x}_j < (\mathbf{w}^*)^T \mathbf{x}_{j'} \quad (32)$$

W rezultacie, sekwencja przyczynowo-skutkowa (16) zbiorów uczących C_k (30) jest w dużej części zachowana przez model (31). Zgodnie z tym modelem, każdy ze zbiorów uczących C_k (30) is transformowany w zbiór C_k' punktów $\mathbf{y}_j(k)$ na linii rangowej :

$$C_k' = \{y_j(k)\} \quad (j \in I_k) \quad (33)$$

Zbiory C_k' mogą zostać scharakteryzowane przez wartości średnie μ_k i wariancje σ_k^2:

$$\mu_k = \frac{\sum_j y_j(k)}{m_k} \quad (j \in I_k) \quad (34)$$

i

$$\sigma_k^2 = \frac{\sum_j (y_j(k) - \mu_k)}{m_k} \quad (j \in I_k) \quad (35)$$

Rezultaty obliczeń z modelem (31) opartym na zbiorach C_k (30) są zestawione w Tabeli 1:

Weźmy pod uwagę dodatkowo liniowe skalowanie $y'=\alpha y+\beta$ modelu $y=(\mathbf{w}^*)^T\mathbf{x}$ (31) w celu polepszenia jego funkcjonalności prognostycznych.

$$y_j'(k) = \alpha(\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \quad (36)$$

gdzie α i β są parametrami skalowania.

Możemy zauważyć, że implikacje rangowe (32) nie zależą od liniowego skalowania (36) modelu.

$$(\forall \alpha > 0)(\forall \beta) \quad (\mathbf{w}^*)^T \mathbf{x}_j < (\mathbf{w}^*)^T \mathbf{x}_{j'} \Rightarrow \alpha(\mathbf{w}^*)^T \mathbf{x}_j + \beta < \alpha(\mathbf{w}^*)^T \mathbf{x}_{j'} + \beta \quad (37)$$

Optymalne wartości α^* i β^* parametrów skalujących zostały wyznaczone przez minimalizację sumy $Q(\alpha,\beta)$ różnic $|k\text{-}\alpha(\mathbf{w}^*)^T\mathbf{x}_j(k)+\beta|$ dla wszystkich wektorów cech $\mathbf{x}_j(k)$:

$$Q(\alpha,\beta) = \sum_{k=1,..,K} \sum_{j \in I_k} \left| k - \alpha(\mathbf{w}^*)^T \mathbf{x}_j(k) + \beta \right| \quad (38)$$

Tabela 1. Wartości średnie μ_k i wariancje σ_k^2 zbiorów C_k' (33).

Zbiory uczące C_k' (33)	Liczby pacjentów m_k	Wartości średnie μ_k	Wariancje σ_k^2 (σ_k)
C_1'	16	-1,02	0,46 (0,68)
C_2'	8	-0,58	0,57 (0,76)
C_3'	44	0,12	1,1 (1,05)
C_4'	95	0,89	1,46 (1,21)
C_5'	38	2,11	2 (1,41)
C_6'	60	3,02	2,2 (1,48)
C_7'	11	3,78	0,62 (0,79)

gdzie I_k jest zbiorem indeksów j wektorów cech $\mathbf{x}_j(k)$ ze zbioru C_k (30).

Zauważmy, że $Q(\alpha,\beta)$ jest także funkcją wypukłą i odcinkowo-liniową. Tak więc algorytmy wymiany rozwiązań bazowych mogą być zastosowane do efektywnego wyznaczenia wartości α^* i β^* tworzących minimum funkcji $Q(\alpha,\beta)$. Rezultaty oceny wyskalowanego modelu pokazane są w Tabeli 2 i na Rys. 1.

Liniowy model rangowy $y=\alpha^*(\mathbf{w}^*)^T\mathbf{x}+\beta^*$ może być używany we wspieraniu diagnostyki nowego pacjenta $\mathbf{x}_0$. Lokowanie się punktu $y_0=\alpha^*(\mathbf{w}^*)^T\mathbf{x}_0+\beta^*$ na prostej rangowej (30) może dostarczyć cennych informacji diagnostycznych o pacjencie $\mathbf{x}_0$. W przypadku modelu wyskalowanego (Rys. 1) możemy oczekiwać, że punkt reprezentujący nowego pacjenta $\mathbf{x}_0$ z k-tej jednostki chorobowej ω_k będzie miał wartość bliską k.

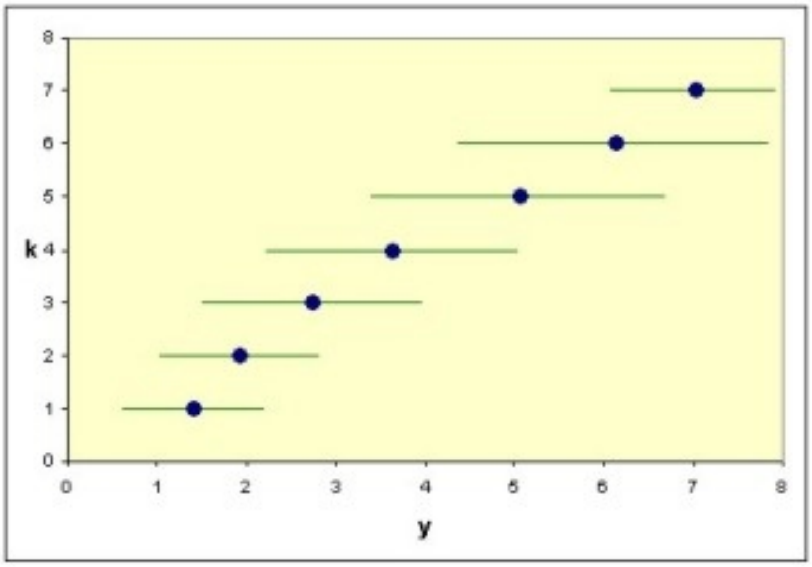
6. Uwagi końcowe

Liniowy model rangowy może być indukowany ze zbiorów uczących C_k (15) na bazie dodatkowej wiedzy medycznej w postaci sekwencji przyczynowo-skutkowej jednostek chorobowych ω_k (k=1,...,K). Relacja rangowa "≺" (2) pomiędzy wektorami cech $\mathbf{x}_j(k)$ z różnych zbiorów uczących. C_k i C_k' została zdefiniowana (18) na podstawie sekwencji przyczynowo-skutkowej (16). Taka relacja rangowa pozwala zdefiniować zarówno zorientowane dipole $\{x_j, x_j\}$ (5) (6) jak również zbiór dodatni R^+ i zbiór ujemny R^- (8) różnic wektorów $\mathbf{r}_{jj'}(\mathbf{x}_j-\mathbf{X}_j)$.

Zbiory R^+ i R^- (8) zostały użyte dla celów skonstruowania wypukłej i odcinkowo-liniowej (*CPL*) funkcji kryterialnej $\Phi(\mathbf{w})$ (13). Optymalny wektor parametrów $\mathbf{w}^*$ (14) wyznaczający minimum funkcji $\Phi(\mathbf{w})$ (13) definiuje liniowy model rangowy (31), który może być użyty dla celów prognostycznych. Na przykład, na bazie modelu rangowego mogą być oparte badania przesiewowe ukierunkowane na wyodrębnienie grup pacjentów wysokiego ryzyka, których należy poddać bardziej szczegółowym badaniom diagnostycznym.

Tabela 2. Wartości średnie μ_k i wariancje σ_k^2 zbiorów C_k' (33) otrzymanych z modelu (31) po jego skalowaniu (36) zgodnie z optymalnymi parametrami α^* i β^*.

Zbiory uczące C_k' (33)	Liczby pacjentów m_k	Wartości średnie μ_k'	Wariancje σ_k^2 (σ_k)
C_1'	16	1,41	0,64 (0,8)
C_2'	8	1,93	0,79 (0,89)
C_3'	44	2,75	1,51 (1,23)
C_4'	95	3,65	1,99 (1,41)
C_5'	38	5,08	2,74 (1,65)
C_6'	60	6,14	3,02 (1,74)
C_7'	11	7,03	0,85 (0,92)



Rys 1: Graficzna prezentacja wartości średnich μ_k' i wariancji σ_k^2 zbiorów C_k' (33) uzyskanych z modelu (31) po skalowaniu (36) zgodnie z optymalnymi parametrami α^* i β^*.

Duże znaczenie praktyczne może też mieć wyodrębnienie tych cech x_i, które miały największy wpływ na rozwój schorzeń w analizowanych grupach pacjentów. Tego typu zagadnienia mogą być rozwiązywane w ramach problemu selekcji cech przy wykorzystaniu wypukłych i odcinkowo-liniowych (*CPL*) funkcji kryterialnych.

Duże znaczenie praktyczne może też mieć wyodrębnienie tych cech x_i, które miały największy wpływ na rozwój schorzeń w analizowanych grupach pacjentów. Tego typu zagadnienia mogą być rozwiązywane w ramach problemu selekcji cech przy wykorzystaniu wypukłych i odcinkowo-liniowych (*CPL*) funkcji kryterialnych.

Praca częściowo finansowana z projektu KBN 3T11F01130, projektu 16/St/2007 z IBIB PAN, oraz z projektu W/II/1/2007 Politechniki Białostockiej.

References

[1] Duda O. R., Hart P. E., Stork D. G.: Pattern Classification, J. Wiley, New York, 2001.

[2] Fukunaga K.: Introduction to Statistical Pattern Recognition, Academic Press 1972.

[3] Johnson R. A., Wichern D. W.: Applied Multivariate Statistical Analysis, Prentice-Hall Inc., Englewood Cliffs, New York, 1991.

[4] Bobrowski L., Łukaszuk T.: Ranked Linear Modeling in Survival Analysis, pp. 61-67 in: Lecture Notes of the ICB Seminars: Statistics and Clinical Practice, ed. by L. Bobrowski, J. Doroszewski, N. Victor, IBIB PAN, Warsaw, 2005.

[5] Bobrowski L.: Ranked Modelling with Feature Selection Based on the CPL Criterion Functions, in: Machine Learning and Data Mining in Pattern Recognition, eds. P. Perner et al., Lecture Notes in Computer Science vol. 3587, Springer Verlag, Berlin, 2005.

[6] Bobrowski L.: Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych (Data mining based on convex and piecewise linear (CPL) criterion functions) (in Polish), Białystok Technical University, 2005.

[7] Bobrowski L., Wasyluk H.: Diagnosis Support rules of the Hepar system, pp. 1309-1313 in: MEDINFO 2001, eds: V. L. Petel, R. Rogers, R. Haux, IOS Press, Amsterdam, 2001.

[8] Bobrowski L.: Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique, Pattern