IJBH

International Journal
on Biomedicine and Healthcare

IJBH 2017 ISSN 1805-8698

International Journal on Biomedicine and Healthcare

Volume 5 (2017), Issue 1

Main Topic

Systems for Medical Decision Support and Forensic Identification

Editors

Arie Hasman and Jana Zvárová

An Official Journal of the EuroMISE Mentor Association

www.ijbh.org

Published by EuroMISE s.r.o., Prague, Czech Republic

Aims and Scope

The *International Journal on Biomedicine and Healthcare* is an online journal publishing submissions in English and/or Czech languages. The journal aims to inform the readers about the latest developments in the field of biomedicine and healthcare, focusing on multidisciplinary approaches, new methods, results and innovations. It will publish original articles, short original articles, review articles and short format articles reporting about advances of biomedicine and healthcare, abstracts of conference submissions, case-studies and articles that explore how science, education and policy are shaping the world and vice versa, editorial commentary, opinions from experts, information on projects, new equipment and innovations.

Editorial Board

Publisher

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

Office

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

Contact

Jana Zvárová
zvarova@euromise.cz

Secretary: Anna Andrlová
E–mail:andrlova@euromise.cz
URL: www.euromise.net

Instructions to Authors

General Remarks

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

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

Authorship

All authors should have made

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

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

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

Conflicts of interest

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

Protection of human subjects and animals in research

Authors who submit a manuscript on research involving human subjects should indicate in the manuscript whether the procedures followed were in compliance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the World Medical Association Declaration of Helsinki on Ethical Princi-

ples for Medical Research Involving Human Subjects (www.wma.net/en/30publications/10policies/b3/).

International Journal on Biomedicine and Healthcare does not publish original articles that has already appeared elsewhere. Submitted manuscripts should not be submitted in parallel to any other journal. All authors should submit Copyright transfer agreement (www.ijbh.org/copyright). Manuscripts using mathematical symbols should be prepared in Latex.

Manuscript preparation

Authors are kindly requested to carefully follow all instructions on how to write a manuscript. The manuscript can be written in Word according to the instructions (www.ijbh.org/word) or in LaTeX according to the instructions (www.ijbh.org/latex). In cases where the instructions are not followed, the manuscript will be returned immediately with a request for changes, and the editorial review process will only start when the paper has been resubmitted in the correct style.

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

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

The journal is publishing the following types of articles: short original articles, original articles, review articles, reports (on projects, education, new methods, new equipment, innovation, electronic healthcare issues), opinions (on management of research, education, innovation and implementation of new methods and tools in biomedicine and healthcare), abstracts (of conferences, workshops and other events), commentary. Manuscript of original articles should follow more detail instructions (www.ijbh.org/original-article).

Kindly send the final and checked source and PDF files of your paper to the secretary `andrlova@euromise.cz` with the copy to editor in chef `zvarova@euromise.cz`.

Contents

Systems for Medical Decision Support and Forensic Identification

Arie Hasman[1,3], **Jana Zvárová**[2,3]

[1] Dept. of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

[2] Charles University in Prague, 1st Faculty of Medicine, Prague, The Czech Republic

[3] EuroMISE Mentor Association

The last three years each year an international conference about a selected topic in Medical Informatics was organised by the EuroMISE Mentor Association. The main activity of the Association is to further the active cooperation of Czech and foreign experts in the domain of Biomedical Informatics and related fields in order to promote the development of the field and passing on knowledge and experience to younger scientists, researchers and pedagogical workers, health services professionals, and the general public. To achieve this goal also a Mentoring Course is organized in addition to the International conference.

On 23. – 24. February 2017 the international conference *Systems for medical decision support and forensic identification* was held. Each of the two topics was treated on a separate day. Each topic is introduced by a keynote given by an outstanding scientist: by *Mor Peleg (Izrael)* the first day and by *Norah Rudin (USA)* the second day. The day devoted to medical decision support focuses on practical and theoretical aspects of decision making. The day devoted to forensic identification also includes contributions related to other subjects.

The previous three international conferences were devoted to the topics

1. *Big Data Challenges for Personalised Medicine*,
2. *Information-based Prevention in Healthcare* and
3. *Electronic Healthcare Documentation*

and a number of good contributions of these conferences were published in the International Journal on Biomedicine and Healthcare (`www.ijbh.org`). The international conference *Systems for medical decision support and forensic identification* was held under the auspices of Charles University, First Faculty of Medicine in Prague and in cooperation with the Czech Society of Biomedical Engineering and Medical Informatics J.E. Purkyně and the Czechoslovak Society of Forensic Genetics. Again a number of good contributions were presented during the conference. This is also apparent from the original articles and abstracts published in this issue of the International Journal on Biomedicine and Healthcare. The last five abstracts in this issue belong to the lectures given at the Mentoring Course organized in addition to the International conference.

The editors would like to thank all the authors for their excellent work as well as to the reviewers for lending their expertise to the conference. We wish the readers much pleasure with reading the articles.

The MobiGuide Clinical Guideline Based Decision-Support System

Objectives, Methods, and Assessment

Mor Peleg[1]

[1] University of Haifa, Haifa, Israel

Correspondence to:

Mor Peleg
Department of Information Systems, Faculty of Social Sciences, University of Haifa, Haifa, Israel
Address: University of Haifa, Haifa, Israel, 3498838
E–mail: morpeleg@is.haifa.ac.il

IJBH 2017; 5(1):2
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

Abstract

MobiGuide was a large-scale 4-year European project, with over 60 researchers, clinicians and engineers, from 13 different organizations in five countries, in the area of guideline-based personalized medicine. The MobiGuide system is a scalable, secure, ubiquitously accessible, and user-friendly mobile solution for designing, deploying, and maintaining a clinical decision-support system (CDSS) for patients and their care providers. The novelty of the approach is in patient-centrality, personalization, and distribution of decision-support for patients who use a mobile CDSS that includes a Smartphone and wearable biosensors that interacts with the main web-based CDSS of the physicians. The CDSS is based on clinical guidelines and personal health records, provides personalized evidence-based clinical recommendations, and has demonstrated in our proof of concept implementation (gestational diabetes patients in a hospital in Spain and atrial fibrillation patients in Italy) an increase in patients' satisfaction and in their compliance to evidence-based clinical guidelines as well as an impact on clinician's decisions.

In this talk I will present the objectives of the project, an overview of the system and its innovation, followed by the main results of our 9-month long evaluation study with patients and clinicians. I will conclude with the main lessons learned from this project.

Author Biography

Mor Peleg is Assoc. Prof at the Dept. of Information Systems, University of Haifa, Israel, since 2003, and has been Department Head in 2009-2012. Her BSc and MSc in Biology and PhD in Information Systems are from the Technion, Israel. She spent 6 years at Stanford BioMedical Research during her post-doctoral studies and Sabbatical, and is currently on Sabbatical at Stanford (till July 2017). She was awarded the New Investigator Award by the American Medical Informatics Association (AMIA) for her work on modeling and execution of the knowledge encoded in clinical guidelines and is International Fellow of the American College of Medical Informatics since 2013. She is Associate Editor of Journal of BioMedical Informatics and a member of the editorial board of Methods of Information in Medicine. Her research concerns knowledge representation, decision support systems, and process-aware information systems in healthcare, and appeared in journals such as JAMIA, International Journal of Medical Informatics, Journal of Biomedical Informatics, IEEE Transactions on Software Eng, TKDE, Bioinformatics. She was the coordinator of the FP7-ICT large-scale project MobiGuide (`http://www.mobiguide-project.eu/`) in 2011-2015. She has edited a number of special issues related to process support in healthcare and artificial intelligence in medicine. Mor has served in program committees of numerous conferences, including, among others, AI in Medicine (Where she chaired the scientific PC in 2011), Medinfo, ER. She has been co-chair of the BPM ProHealth Workshop seven times and an organizing committee member of Knowledge Representation for Healthcare Workshop five times. `http://mis.hevra.haifa.ac.il/~morpeleg/`

Complex DNA Profile Interpretation

Stories from across the Pond: The Current State of Forensic DNA Interpretation and Statistics in the U.S.

Norah Rudin[1]

[1] Forensic DNA, Mountain View, CA, U.S.

Correspondence to:

Norah Rudin
Forensic DNA
Address: 650 Castro St. Suite 120-404,
Mountain View, CA 94041, U.S.
E–mail: norah@forensicdna.com

IJBH 2017; 5(1):3
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

Abstract

With the continued increase in the sensitivity of DNA testing systems comes a commensurate increase in the complexity of the profiles generated. Numerous sophisticated statistical tools intended to provide an appropriate weight of evidence for these challenging samples have emerged over the last several years. While it seems clear that only a likelihood ratio-based probabilistic genotyping approach is appropriate to address the ambiguity inherent in these complex samples, the relative merits of the different approaches are still being investigated.

I will summarize the history of approaches typically used by working analysts in the US, and discuss the current state of the practice. In 2005 and 2013, NIST distributed sets of mixtures to working laboratories and collected their interpretations and statistical weights. They found a wide range of variation both within and between laboratories in calculating the weight of evidence for the same sample in both surveys. Most disturbing was the continued use of simplistic tools, such as the CPI/CPE (RMNE), long considered inadequate for specific types of profiles. A number of publications and reports over the last 15 years have commented on the interpretation and statistical weighting of forensic DNA profiles. These include the ISFG commission papers of 2006 and 2012, the NAS 2009 report, the 2010 SWGDAM STR interpretation guidelines, and the 2015 SWGDAM probabilistic genotyping software validation guidelines. Several high profiles criticisms of laboratory protocols (e.g. Washington D.C. and the TX laboratory system) have emerged that have fueled debate. Recently, PCAST published a report commenting on the state of forensic science disciplines in the US, including DNA. An updated draft of the SWGDAM STR interpretation guidelines is currently posted for comment. Finally, the state of admissibility of PG in the U.S. will be discussed.

Author Biography

Norah Rudin holds a B.A. from Pomona College and a Ph.D. from Brandeis University. She is a member of the California Association of Criminalists, a fellow of the American Academy of Forensic Sciences, and a Diplomate of the American Board of Criminalistics. After completing a post-doctoral fellowship at Lawrence Berkeley Laboratory, she served three years as a full-time consultant/technical leader for the California Department of Justice DNA Laboratory and has also served as consulting technical leader for the Idaho Department of Law Enforcement DNA Laboratory, the San Francisco Crime Laboratory DNA Section, and the San Diego County Sheriff's Department DNA Laboratory. Dr. Rudin has co-authored An Introduction to DNA Forensic Analysis and Principles and Practice of Criminalistics: The Profession of Forensic Science. She is also the author of the Dictionary of Modern Biology. Dr. Rudin has taught a variety of general forensic and forensic DNA courses for the University of California at Berkeley extension and on-line. She is frequently invited to speak at various legal symposia and forensic conferences, and recently served a gubernatorial appointment to the Virginia Department of Forensic Science Scientific Advisory Committee. She is currently co-chair of the Constitution Project Committee on DNA Collection. She remains active as an independent consultant and expert witness in forensic DNA.

Compliance of Patient's Record with Clinical Guidelines

Arnošt Veselý[1], **Jana Zvárová**[2]

[1] Department of Information Engineering, Czech University of Life Sciences, Prague, Czech Republic

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

Abstract

In the paper we present an algorithm for comparing the patient's data record with clinical guidelines formalized in the EGLIF model. EGLIF is a simple enhancement of the standard GLIF model. We use the EGLIF model to make the design of the comparison algorithm more clear and convenient. If the patient's record completely describes the carried out treatment, then the comparing algorithm is able to recognize if the patient's treatment complies with the guidelines or not.

Keywords

Medical guidelines, Electronic health record, GLIF model, Remainder system

Correspondence to:

Arnošt Veselý
Czech University of Life Sciences
Address: Kamýcká 129, Praha 6
E–mail: vesely@pef.czu.cz

IJBH 2017; 5(1):4–7
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

Clinical guidelines (CG) contain a set of care decisions to assist the physician with patient care decisions about appropriate diagnostic, therapeutic, or other clinical procedures. They are intended to ensure high quality clinical practice [1]. CG are developed as textual recommendations by groups of medical experts on topics selected by a local scientific authority, e.g. an expert medical society or a national health institution. Usually their text focuses on a specific group of physicians or health professionals.

Paper-based guidelines have to be translated into a computer based representation for further computer processing. For this purpose a handful of formalization means have been developed. In the paper we use the EGLIF model that is based on the frequently used GLIF model.

When we know the patient treatment history, the natural question is, if the patient was treated in compliance with recommendations given in the guidelines.

The overview of methods designed to solve this task can be found in [2]. These methods are based on comparing of clinical actions performed by a physician and recorded into his health record with a predefined set of actions recommended by CG. In [3] the Guidelines Definition Language (GDF) has been defined that enables to transform recommendations given in the clinical guidelines into *when-then* rules. Fulfilment of these rules then may be tested on patient's health record data.

In our paper we assume that relevant information about the patient treatment is stored in the patient data record. We will describe an algorithm that compares the patient data record with the EGLIF guideline model and that is able to determine if the treatment was in compliance with it.

The comparison may be ex post or on-line. In the case of ex post comparison we have at our disposal patient records from a long time period and we want to know ex post if patients were treated according to the appropriate standard described in guidelines.

Online comparison means that patient data record is compared with the standard each time when it is updated with a new data item. An on-line reminder system, which warns the physician if his action does not comply with the treatment standard, might be based on such online comparison.

2 Comparing patient's record with EGLIF model

When a patient is under medical care, physicians carry out examinations, laboratory tests and apply different medications or therapies. From this point of view the patient is subjected to different actions. Each action is carried out at some time t and it has its result r. Actions will be denoted $A(r, t)$.

The sequence of actions the patient was subjected to, ordered according to time, we call procedure. A procedure is denoted as

$$R = A_1(r,t), \ldots, A_n(r,t)$$

Not all procedures are during the care acceptable from a medical point of view. The class of all recommended procedures is defined in clinical guidelines. Clinical guidelines are originally written in everyday language. To be able to process it by computer, guidelines must be converted into a formal model. For formalization the GLIF model is often used [4].

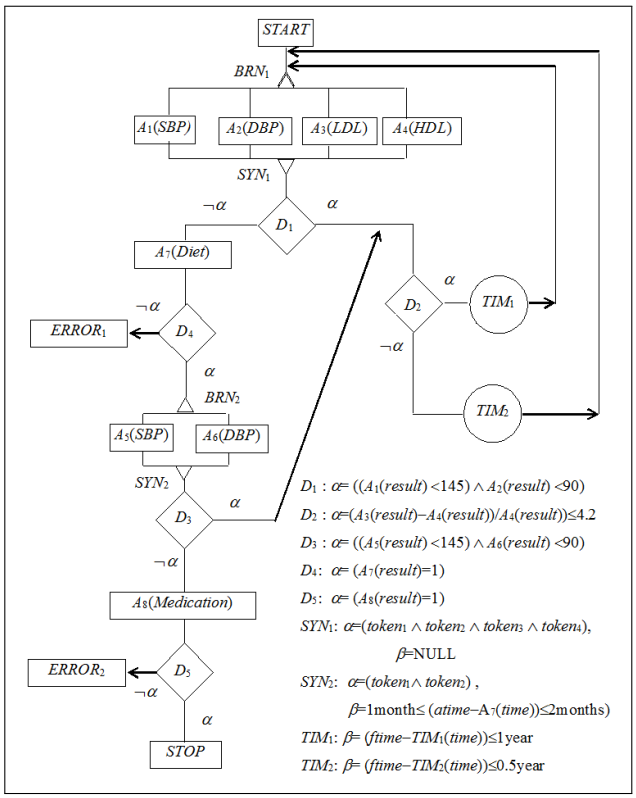

Figure 1: The EGLIF graph model of the Small guidelines for heart failure prevention.

The GLIF model [5] is a weighted directed graph. The nodes of the graph are guideline steps and edges represent continuation from one step to the other. The main guideline steps are *action step, decision step, branch and synchronization steps*. *Action steps* specify clinical actions that are to be performed. It can be an application of some therapy, carrying out some examination or measurement etc. Action step also may be sub-guidelines, which provide greater detail for the action. *Decision steps* are used for conditional branching. A decision step specifies the criteria for each possible alternative decision. *Branch and synchronization steps* enable concurrency in the model. Guideline steps that follow a branch step could be performed concurrently. Branches from a branch step eventually converge in a synchronization step. In this step all branches are synchronized. It means, that actions, that follow the synchronization step, could not be performed, unless all actions, following branch step and preceding the synchronization step, are finished.

In our paper we use the EGLIF model [6]. EGLIF is basically a GLIF model. The main difference is that we added a Time node to render the comparison with a procedure more transparent and easier. Time node sets time limits to the next action. When a token passes a time node, the current time is written into its parameter *time* and condition β assigned to the Time node is a time condition that the next action must fulfil.

Sync node in EGLIF has its continuation condition α that is sufficient condition for passing the Sync node and time condition β. Time condition β must be fulfilled for the time parameters of all actions carried out between Branch node and the corresponding Sync node.

Each branch of decision node has its in-condition. If the in-condition is fulfilled, then continuation through the corresponding branch is recommended. We assume that during passing the decision node one or more in conditions are fulfilled.

Some nodes in EGLIF are able to store a token or tokens. By means of tokens we simulate the passage through the model, when we compare the model with some procedure. Nodes that store tokens are Action nodes, Synchronization nodes and Start and Stop nodes. At the beginning the only token is in Start node. Then during comparison tokens move from nodes, where they are stored, to next nodes that can store tokens. Next node here means the first following node on the recommended branch.

We give here only brief description of the comparing algorithm. More detailed description can be found in [6].

2.1 Description of CA algorithm

Comparison of guidelines G with procedure

$$R = A_1(r,t), \ldots, A_n(r,t)$$

1. Action nodes, Start and Stop nodes and Sync nodes store tokens. The remaining nodes only propagate tokens.

2. At the beginning there exists only one token in the Start node.

3. In each step algorithm CA deletes the first action from sequence R and moves tokens from the nodes labeled with this action to the next nodes in the graph that store tokens. If no token can be moved, the algorithm stops.

4. When a token goes through a Branch node, new tokens are created, one for each branch.

5. As soon as Sync condition α is fulfilled, one token is propagated from Sync node.

6. When a token goes through a decision node, all recommended branches propagate the token. If more

than one branch is recommended, then new tokens are created.

7. If the run of CA algorithm ends and the sequence R is void, then procedure R is in compliance with guidelines G.

Example Small guidelines for heart failure prevention.

When a patient comes for a visit, his physician examines patient's blood pressure parameters SBP (systolic blood pressure), DBP (diastolic blood pressure) and let his cholesterol parameters LDL (low density cholesterol), HDL (high density cholesterol) be determined in laboratory.

1. If blood pressure is not normal, i.e. if the condition

$$\alpha = (SBP < 145) \wedge (DBP < 90)$$

is not satisfied, physician prescribes a diet and invites the patient for repeated examination after 1-2 months:

   (a) If the patient's blood pressure is not normal again, the physician prescribes medication.

   (b) If patient's blood pressure is normal, the physician evaluates patient's risk index

$$i_R = (LDL - HDL)/HDL$$

   If the risk index is small ($i_R < 4.2$), the patient is invited for the next examination not later than after a year. If the risk index is greater than 4.2, the patient is invited not later than after half a year.

2. If the blood pressure is normal, i.e. condition α is satisfied, physician evaluates patient risk index i_R. If the risk index is small ($i_R < 4.2$), the patient is invited for the next examination not later than after a year. If the risk index is greater than 4.2, the patient is invited not later than after half a year.

The EGLIF graph model of Small guidelines for heart failure prevention is given in Figure 2.

Assume that the procedure carried out by the physician is
$R = SBP(150, 1.1.01), DBP(85, 1.1.01), HDL(1, 2.1.01), LDL(6, 2.1.01), Diet(1, 2.1.01), DBP(85, 10.2.01), SBP(140, 10.2.01), SBP(130, 1.5.01), DBP(85, 1.5.01), HDL(1, 2.5.01), LDL(5.52.5.01), SBP(130, 1.4.02), DBP(90, 1.4.02), LDL(7, 2.4.01), HDL(2, 2.4.01)$

Here, to keep the presentation clear and simple, action time is expressed using day time scale, i.e 1.1.01 means the 1st January 2001.

After little thinking we can see that treatment of the patient does not comply with guidelines. In the procedure we have HDL (1, 2.5.01) and LDL (5.5, 2.5.01) and therefore the risk index during patient's visit at 2.5.01 had value $i_R = 4.5$. Hence the patient's following visit should have been sooner than after half a year. But his next visit was at 1.4.02 as we can see from the data item SBP (130, 1.4.02).

Algorithm CA is able to discover this discrepancy. For more detail see [6].

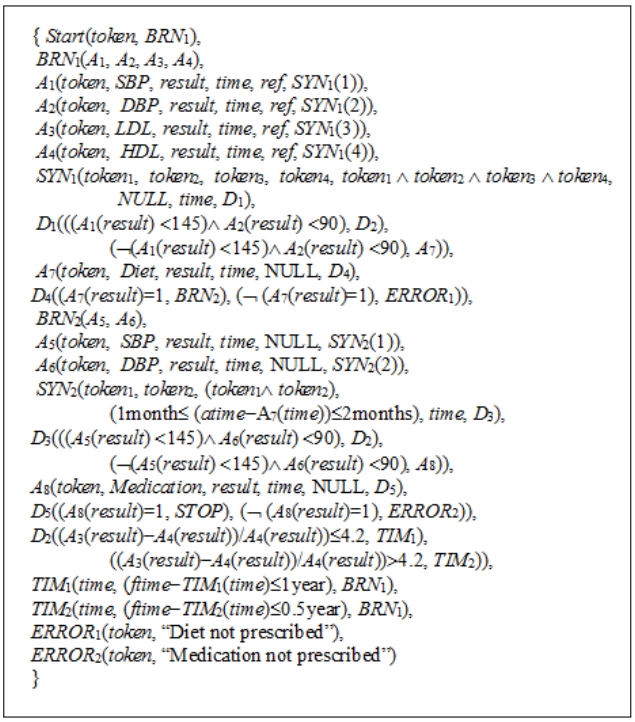

Figure 2: Coded EGLIF model of the Small guidelines for heart failure prevention.

3 Clinical record compliance with guidelines

Assume we have at our disposal a complete description of patient treatment stored in a patient database and an EGLIF model of guidelines. Then we may use the CA algorithm for checking compliance of the treatment with the guidelines.

3.1 Compliance problem

Given procedure (treatment)

$$R = A_1(r, t), \ldots, A_n(r, t)$$

and an EGLIF model G of guidelines. Is the treatment R in compliance with the guidelines?

For solution of compliance problem we can use the CA algorithm described above.

Consider as an example a reminder system. The current design of a reminder system that is usually built in information systems is outlined in Figure 3.

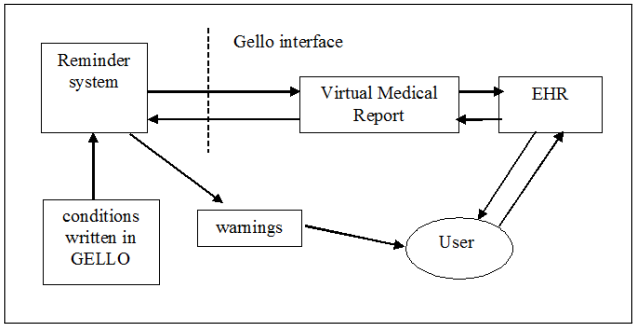

Figure 3: Architecture of today remainder system.

Here if the new item is put into a patient database (Electronic Health Record (EHR)), then all conditions that contain new item are automatically checked and if necessary, warnings are displayed. The set of conditions can be created on the basis of some guidelines or may be given separately by physician specialists.

In a reminder system it would be possible to use the guideline model instead of the set of conditions and compare stored data directly with the model, as it is outlined in Figure 4. Here, if the new item is stored into the EHR, a description of the previous treatment is generated from EHR and then it is compared with EGLIF model using CA algorithm.

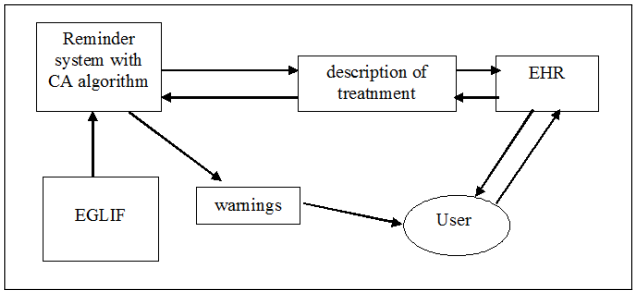

Figure 4: Remainder system using direct comparison of the input item and guidelines.

However, the treatment description often will be incomplete. If we used the CA algorithm, the description should be complete. Otherwise a reminder system would generate a lot of superfluous warnings.

Nevertheless, it is easy to see that also in the case of incomplete treatment description some discrepancies with guidelines can be discovered. To discover them, we need to modify CA algorithm in order to be able to solve the following general compliance problem.

3.2 General compliance problem

Given subsequence S of the treatment R and guidelines model G. Is it possible so that R be in compliance with guidelines?

4 Conclusion

In this paper we introduced an algorithm that compares a patient treatment described with a patient data record with EGLIF model of clinical guidelines and determines if the patient treatment was in compliance with the guidelines or not. The designed algorithm assumes that the patient data record is complete.

It is clear that in the situation where we know that only some data about patient treatment were recorded, the possibility to test compliance of his treatment with the guideline model is strongly limited. However, in some cases non-compliance can be discovered in spite of missing data. This happens if the data record would be non-compliant with the guidelines model whatever the missing data were.

In some applications, e.g. in building a reminder system, we might admit missing data and we might want to get warnings only if the non-compliance is obvious from the remaining data. The design of effective algorithm solving comparison of an incomplete data record with guideline model is subject of further research.

References

[1] Isern D, Moreno A. Computer-based Execution of Clinical Guidelines: A Review, *International Journal of Medical Informatics 77*, Maastricht, 2008, 787-808.

[2] Peleg M. Computer-interpretable Clinical Guidelines: A Methodological Review, *Journal of Biomedical Informatics*, 46, 2013, 744-763.

[3] Anani N, Chen R, Moreira T, Koch S. Retrospective Checking of Compliance with Practice Guidelines for Acute Stroke Care: a Novel Experiment using Open EHR's Guideline Definition Language, *BMC Medical Informatics and Decision Making* 14:39, BioMed Central, 2014,1-18

[4] Patel V, Branch T, Wang D, Peleg M, Boxwala, Analysis of the Process of Encoding Guidelines: A Comparison of GLIF2 and GLIF3, *Methods Inf. Med.*, no.2, 2002, 105-113.

[5] Ohno-Machado,L . Gennari JH, Murphy SN. Jain SL, Tu SW, Oliver SD, et al. The GuideLine Interchange Format: A model for representing guidelines, *Journal of the American Medical Informatics Association*, 5(4), 1998, 357-372.

[6] Veselý A, Zvárová J. Determination of Guidelines Compliance: Comparison of Clinical Guidelines with Patient's Record, *EJBI* 8(1), 2012, 16-28.

What is a Relation between Data Analytics and Medical Diagnostics?

František Babič, Ján Paralič, Michal Vadovský, Miroslava Muchová, Alexandra Lukáčová, Zuzana Vantová

[1] Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia

Abstract

Data Analytics provides various methods and approaches to extract hidden and poten-tially useful knowledge for different purposes. It means that we can use this knowledge for decision support, e.g. to identify crucial inputs and relations, to predict some future state, to confirm or reject our hypothesis. The medical diagnostics deals with all the mentioned tasks to provide a right diagnosis for the patient and to ensure the effective diagnostic process. In this paper, we briefly describe some of our activi-ties oriented to support medical diagnostics by means of Data Analytics approaches; we selected only key points here, more details can be found in our previously pub-lished articles.

Keywords

Patients, Records, Prediction, Cut-of values, Exploratory analysis

Correspondence to:

František Babič
Technical university of Košice
Address: Letná 9, Košice, Slovakia
E–mail: frantisek.babic@tuke.sk

IJBH 2017; 5(1):8–12
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

The medical diagnostic is in some cases a complex and time-consuming process because the final decision depends on many input variables and hidden relations between them. It is a primary task for every general practitioner (GP) or specialist to identify and interpret this information in the right way, i.e. confirm or reject the expected diagnosis. The Data Analytics provides a set of various approaches and methods how to understand the medical data, how to process the data in a required format, how to apply suitable methods or algorithms on the prepared data and how to understand the obtained results. We selected some of the algorithms and methods from the domains like machine learning, statistics, artificial intelligence and exploratory data analysis. The GP may not have a deep knowledge about these analytical operations because of he (she) only specifies a task and expected result's format. It is important to point out that the GP or specialist is responsible for the final diagnosis, not the machine or software.

In the Centre of Business Information Systems, we focus on the possibly useful application of the various analytical methods to support the medical diagnosis of the selected diseases, e.g. Mild Cognitive Impairment (MCI), Metabolic Syndrome (MetSy), Hepatitis or Parkinson's disease (PD). For this purpose, we have used some data samples collected by our medical partners or freely available.

In this paper, we present some key points and interesting findings from our previous experiments published in other conferences like ITBAM (International Conference on Information Technology in Bio- and Medical Informatics). We selected here a classification model, an example of identification of the new cut-off values and an exploratory data understanding, which we consider as important means to support medical diagnosis. More details can be found in our previous papers that we mention with each experiment described in this paper.

2 Classification

Classification is a Data Analytics task that assigns input records to the target class with some suitably estimated accuracy. In our case, we solved a binary classification, i.e. we wanted to confirm or to reject the possible disease diagnosis. In the case of MCI [5] and MetSy [1], we used a data sample collected in a family practice located in an urban area of the town of Osijek, the north-eastern

part of Croatia. This sample contained records about 93 patients described by more than 60 variables. These variables represented routinely collected information from the patients' health records (e.g. age, sex, chronical diseases, drug use, etc.) and relevant laboratory tests' results e.g. information on inflammation, nutritional status, metabolic status, chronic renal impairment, latent infections, humoral (antibody-mediated) immunity and neuroendocrine status.

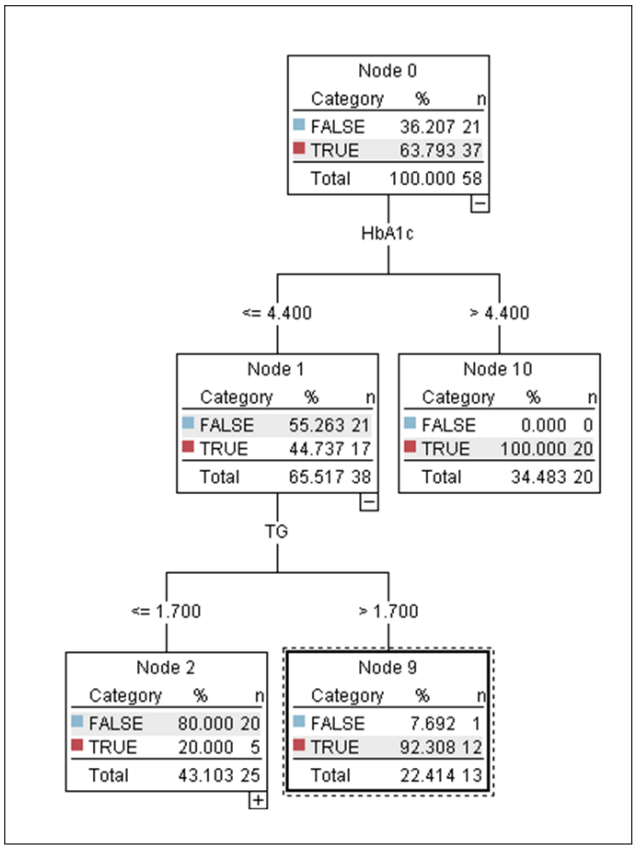

Figure 1: Example of the generated decision tree model (FALSE = MetSy diagnosis no, TRUE = MetSy diagnosis yes).

We chose a decision trees classification model because of the ability to extract the crucial knowledge for the decision in simply understandable visual form. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent target classes or class distributions [8]. The C4.5 algorithm builds the decision tree model using the concept of information entropy. At each node, it selects the attribute and its split criterion, which divides the data into subsets to distinguish the target classes as good as possible [9]. The C5.0 algorithm represents an improved version of the C4.5 that offers a faster generation of the model, less memory usage, smaller trees with similar information value, weighting and support for the boosting [10]. The CART (Classification and Regression trees) algorithm builds a model by recursively partitioning the data space and fitting a simple prediction model within each partition [11]. This method becomes popular as an alternative to the regression and discriminant analysis. The result is a binary tree using a greedy algorithm to select a variable and related cut-off value for splitting with the aim to minimise a given cost function. The CHAID (Chi-squared Automatic Interaction Detector) algorithm is one of the oldest tree classification methods originally proposed by Kass [12]. We decided for CHAID because obtained output is visually very easy to interpret. It usually creates simple conservative trees that give good results on the test set. This algorithm constructs non-binary trees and uses the Chi-square test to determine the best next split at each step. CHAID is recursive partitioning method.

We present one example from the many generated decision trees evaluated by the participated medical expert to confirm their correctness and usefulness. We used a 10-fold cross validation and discussed with the expert impact of the low number of input records to the obtained models. Figure 1 visualises the decision model for female patients consisting of two important variables. From this figure, we can extract some decision rules to reject or to confirm the MetSy diagnosis, e.g. if a patient has an average level of blood glucose during last three months more than 4.4 and level of triglycerides more than 1.7, with the 92.3% accuracy she has a positive MetSy diagnosis. In addition, we can state that in a similar group of people with similar health characteristics are women more prone to diabetes. We have published more experiments and relevant results in [2].

We used similar approach (but this time including the cost matrix) to improve the diagnostic process of hepatitis B (HBV) and C (HCV) based on collected questionnaires from patients hospitalised in all regional infectology departments in Slovakia. More than 4.5 thousand patients filled an anonymous questionnaire containing three sections: demographic data, epidemiological data and blood tests results. We aimed to confirm or reject some expected relationships between input variables; also to generate the prediction models in an early stage of the diagnostic process and finally to evaluate the economic effectivity of the necessary treatment. The dataset contained only 79 patients with confirmed HBV and 65 patients with confirmed HCV. This fact motivated us to propose several types of experiments: e.g. with the whole datasets, with separate male and female patients, with different ratios division into training and test set, with random or stratified sampling, with an oversampling or subsampling dataset based on target attribute distribution.

Figure 2 shows a part of the generated model for HBV diagnosis based on male patients only. The classification accuracy of this model was around 88%. The same experiment with female patients resulted in 90% accuracy. From the whole experiments, we extracted the most interesting finding evaluated by the participated expert, i.e. the probability of hepatitis B infection is higher in patients with a liver diagnostic test (ALT) ≥ 0.56 μkat/l. This knowledge can lead GP to test patients for HBsAg antigen even with normal levels of this test. This approach can early confirm

the diagnosis, which would significantly reduce the costs of their later treatment, and prevent further disease progression and worsen the health of patients. More details can be found in one of our published papers [4].

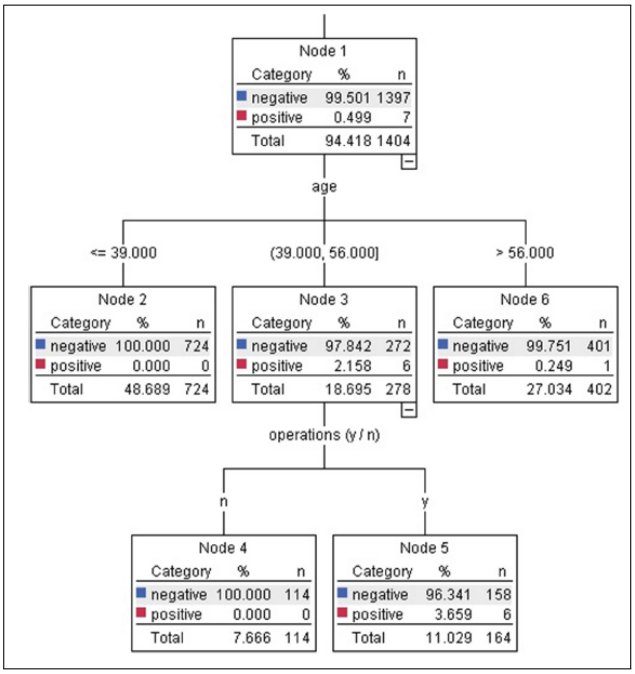

Figure 2: A part of the generated decision model for HBV diagnosis (negative = diagnosis no, positive = diagnosis yes).

We performed all above mentioned experiments as off-line analysis, i.e. we used the historical datasets to extract potentially useful knowledge for future diagnostics of the relevant diseases for a new group of patients with similar characteristics.

3 Cut-off Values Identification

In the second group of experiments, we used the same datasets to identify the new cut-off values for variables with higher impact to the expected diagnosis. In the case of MetSy, we aimed at variables selected by the participated expert as e.g. FOLNA (Folic acid), HbA1c (average blood glucose during last three months), MO (Monocytes % in White Blood Cell differential) and TSH (Thyroid-stimulating hormone). For this purpose, we used the measure called Youden index [3] and obtained results confirming the previous ones. The cut-off value in decision tree model for variable FOLNA was 14.7 and the Youden method calculated 15.6 (sensitivity 95.65%, specificity 83.33%). The similar situation appeared for variable HbA1C: 4.41 and 4.5 resp. (sensitivity 39.13%, specificity 100%). Because of these findings, we can consider the variable FOLNA as a new biomarker of MetSy, particularly suitable for screening in general male population.

The advantage of Youden method is in offering the best result with respect to the highest overall correct classification by maximizing the sum of sensitivity and specificity. The range of Jounden index is ¡0, 1¿, where the value 1 stands that all positive and negative patients are correctly classified. On the other hand, the value 0 means that the selected cut-off value is completely ineffective.

For MCI diagnosis, we slightly improved this approach, i.e. at first we constructed ROC curve (Receiver-Operating Characteristic) representing a graphical plot illustrating the performance of a binary classifier system as its discrimination threshold is varied (MCI diagnosis = yes/no) and calculated the corresponding AUC (Area Under ROC). Next, we applied again the Youden method to identify the new cut-off values (Table 1).

Table 1 contains the new cut-off values for variables like age, Body Mass Index, a level of Immunoglobulin E and one of the globulins separated through a serum protein electrophoresis used to identify patients with multiple myeloma and a sign of chronic low-grade inflammation (ALFA2). We concluded that an ability of the decision model based on these variables is better for the healthy persons as for patients with positive diagnosis. It may indicate that the model classifies some negative cases as positive, i.e. false alarms. This fact can be caused by a low number of positive records in our dataset and will be an objective of our future work to increase the TPR value. In addition, we can visualize these results in the simple understandable graphical form (Figure 3).

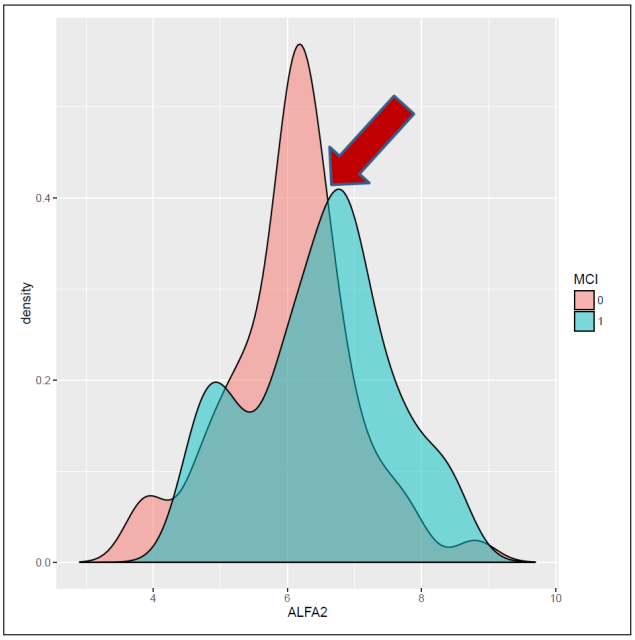

Figure 3: The graph of distribution function for the variable ALFA2 with identified cut-off value (red arrow).

In the case of hepatitis C, we selected attribute ALT as an illustrative example. Its typical cut-off values used in medical practice are 0.8 for men and 0.6 for women. We used our own cost matrix to evaluate the effectiveness of the generated models. Participated expert specified the following values for this matrix: TP (true positive) = 0, TN (true negative) = 0, FP (false positive) = 7€ for

Table 1: The most important variables for MCI diagnosis with new cut-off values.

Variable	AUC	Cut-off value	TNR (%)	TPR (%)	Accur (%)
Age	0.74	70.50	82.14	54.05	70.97
ALFA2	0.63	06.75	82.15	48.65	65.59
BMI	0.62	29.02	76.79	54.05	67.74
IGE	0.60	32.30	64.29	64.86	64.52

TNR - true negative rate (specificity): % of correctly classified people with negative MCI diagnosis (healthy persons)
TPR - true positive rate (sensitivity): % of correctly classified patients with positive MCI diagnosis

HCV and 9€ for HBV (costs of the corresponding investigations), FN (false negatives) = 1 400€ for F4 (cirrhosis treatment) or 740€ for F3 stage of the hepatitis C treatment. It was interesting that the calculated cut-off values for the male population were in all experiments the same (0.59 for both F3 and F4 stages), but there was a big difference in female population (0.24 for F4 stage and 3.94 for F4). Figure 4 shows a boxplot of the patients with normal ALT level according to the hepatitis B diagnosis (red line means the new cut-off value for F4 stage costs and red dotted line is for F3 stage costs).

Based on the overall results we concluded that the new cut-off points for HBV diagnosis is more cost sensitive. Using the new ALT threshold should lower the overall costs implied by the late diagnosis of hepatitis B, which is associated with much higher treatment costs.

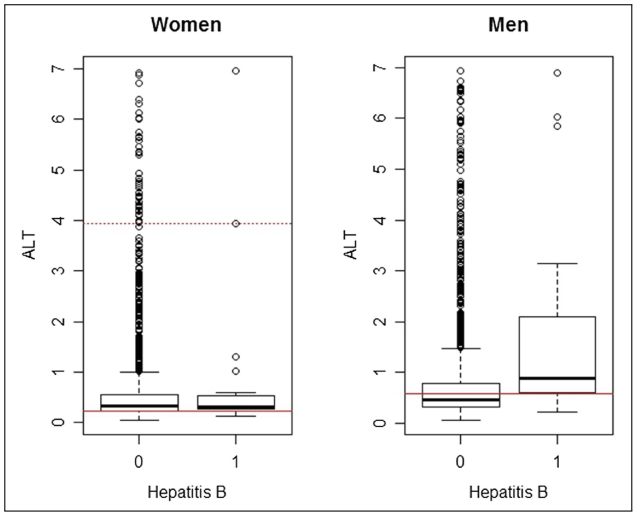

Figure 4: Boxplot for variable ALT in relation with target diagnosis HBV.

4 Exploratory Data Understanding

This type of understanding we used to explore the attribute's dependency in MCI related data sample. We applied a two sample Welch's t-test on nominal and a Pearson chi-square independence test on numerical attributes [6]. For both cases, we specified a hypothesis, which we wanted to confirm or to reject, i.e. two populations have equal means. As the most significant numerical attributes for the target diagnosis we confirmed ALFA2 (p-value = 0.0458, significance level = 0.05), Clear (0.0535, 0.1) and Skinf (0.0953, 0.1). Clear represents a creatinine clearance test, which doctors use to evaluate the rate and efficiency of kidney filtration, i.e. to detect and diagnose kidney dysfunction and/or the presence of decreased blood flow to the kidneys. Skinf represents a triceps skin-fold thickness, i.e. a value used to estimate body fat, measured on the right arm halfway between the olecranon process of the elbow and the acromial process of the scapula. In the case of nominal attributes, we confirmed the dependency only for a therapy with nonsteroidal anti-inflammatory drugs. More details can be found in [7].

5 Conclusions

In this paper, we briefly overviewed some of our research activities dealing with exploitation of the selected analytical methods to support the medical diagnostics for various diseases as MCI, MetSy and hepatitis. We can conclude that all obtained results are plausible; some of them the participated experts confirmed by related literature or their personal experiences. On the other hand, they marked some findings as interesting, new, potentially useful and needed for further verification. These experiences help us to prepare some new project proposals applying for financing from Slovak national funds or H2020. In our future work, we will focus on multivariate methods, clustering and also on development of an intelligent diagnostic system for the selected diseases. This is a quite challenging task.

Acknowledgments

The work presented in this paper was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0493/16 and by the Cultural and Educational Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 025TUKE-4/2015.

References

[1] R.A. Eckel, S.M. Grundy, P.Y. Zimmet, The metabolic syndrome, Lancet 365 (2005), 1415-1428.

[2] F. Babič, L. Majnarić, A. Lukáčová, J. Paralič, A. Holzinger, On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive Modelling Based on Machine Learning, Information Technology in Bio- and Medical Informatics LNCS 8649 (2014), 118-132.

[3] W.J. Youden, Index for rating diagnostic tests. Cancer 3 (1950), 32-35

[4] A. Lukáčová, F. Babič, Z. Paraličová, J. Paralič, How to Increase the Effectiveness of the Hepatitis Diagnostics by Means of Appropriate Machine Learning Methods, Information Technology in Bio- and Medical Informatics LNCS 9267 (2015), 81-94.

[5] M.S. Alber, S.T. DeKosky, D. Dickson, B. Dubois, H.H. Feldman, N.C. Fox, A. Gamst, D.M. Holtzman, W.J. Jagust, R.C. Petersen, P.J. Snyder, M.C. Carillo, B. Thies, C.H. Phelps, The Diagnosis of Mild Cognitive Impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's disease. Alzheimer's & Dementia 7(3) (2011), 270-79.

[6] B. Shahbaba, Biostatistics with R: An Introduction to Statistics through Biological Data (2012).

[7] M. Vadovský, F. Babič, M. Muchová, Systém na podporu rozhodovania pomocou jednoduchého a efektívneho pochopenie medicínskych záznamov. WIKT & DaZ 2016, 11th Workshop on Intelligent and Knowledge Oriented Technologies, 35th Conference on Data and Knowledge (2016), 89-93.

[8] K. S. Murthy, Automatic construction of decision tress from data: A multidisciplinary survey. Data Mining and Knowledge Discovery (1997), 345-389.

[9] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers (1993).

[10] N. Patil, R. Lathi, V. Chitre, Comparison of C5.0 & CART Classification algorithms using pruning technique. International Journal of Engineering Research & Technology 1(4) (2012), 1-5.

[11] L. Breiman, J.H. Friedman, R.A. Olshen, Ch.J. Stone, Classification and Regression Trees (1999), CRC Press.

[12] G.V. Kass, An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics 29(2) (1980), 119–127.

Knowledge Representation and Knowledge Management as Basis for Decision Support Systems

Bernd Blobel [1,2]

[1] Medical Faculty, University of Regensburg, Germany

[2] eHealth Competence Center Bavaria, Deggendorf Institute of Technology, Germany

Abstract

Deciding on things is a knowledge-based activity. In the context of clinical decision support systems (DSS) this means that representation and management of the related knowledge about underlying concepts and processes is foundational for the decision-making process. A basic challenge to be mastered is the language problem. For expressing and sharing knowledge, we have to agree on terminologies specific for each of the considered domains. For guaranteeing semantic consistency, the concepts, their relation and underlying rules must be defined, deploying domain-specific as well as high-level ontologies. Ontology representation types range from glossaries and data dictionaries through thesauri and taxonomies, meta-data and data models up to formal ontologies, the latter represented by frames, formal languages and different types of logics. Based on the aforementioned principles, special knowledge representation and sharing languages relevant for health have been introduced. Examples are PROforma, Asbru, EON, Arden Syntax, GELLO, GLIF, Archetypes, HL7 Clinical Statements, and the recently developed FHIR approach. With increasing complexity and flexibility of decision challenges, DSS design has to follow a defined methodology, offered by the Generic Component Model Framework meanwhile internationally standardized. This paper deals in detail with the basics and instances for knowledge representation and management for DSS design and implementation, thereby referencing related work of the author.

Keywords

Knowledge representation, Decision support systems, Artificial intelligence, System theory, Architecture, Ontologies, Standards, GCM

Correspondence to:

Bernd Blobel
Medical Faculty, University of Regensburg
Address: Regensburg, Germany
E–mail: bernd.blobel@klinik.uni-regensburg.de

IJBH 2017; 5(1):13–20
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

Decision making is a process of making a choice in the context of a business process at its planning, management and operation level involving people, technology and system development. Usually, it applies in changing or underspecified, i.e., in dynamic environments, also called unstructured or semi-structured decision problem [1]. Decision making deploys traditional data access and retrieval with modeling and analytics techniques. There are two different cycles to be understood when dealing with decision support systems: the decision making process cycle (Figure 1a) and the information cycle (Figure 1b) [2, 3]. Running both cycles requires knowledge. Therefore, knowledge creation, knowledge representation (KR) and knowledge management (KM) is crucial for the decision making process.

A simplification of the decision making process cycle (Figure 1a) consists of

- intelligence in defining the problem;
- design in developing and analyzing concurrent problem solutions;
- choice in selecting the appropriate action, and
- implementation in adopting that action in the decision making process.

The information cycle (Figure 1b) describes different aspects of information related to corresponding information definitions. The syntactic aspect is based on observation resulting from measurements and results in data recording, represented by Shannon's information defini-

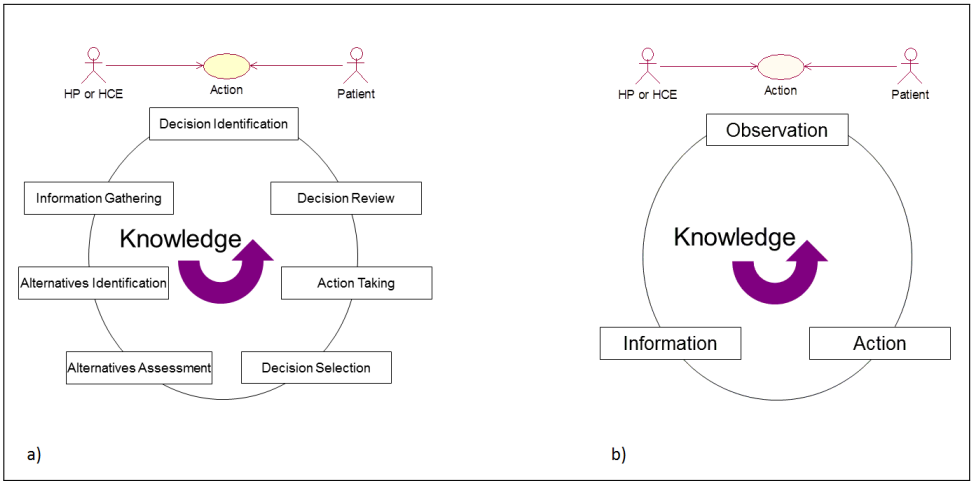

Figure 1: Cycles relevant for understanding decision making processes.

tion[1]. The semantic aspect dealing with the interpretation of data to information is based on conclusion derived from theory and results in data processing and decision support, related to Brillouin's information definition[2]. The pragmatic aspect is related to actions resulting from a practical experiment and provides output generation and process control according to Wiener's information definition[3]. [3]

Decision support systems (DSS) can be classified into passive and active ones. While the first only manage the data phase in the information cycle (Figure 1b), active DSS run the complete cycle. When including human beings in that process iteratively, active DDS are called cooperative [4]. The general architecture of a DSS comprises of the data or knowledge base, an inference or reasoning engine, and the user interface for interacting with the DSS.

2 Methods and Definitions

The paper is based on related work on DSS, KR and KM. presented in similar or different context to specific events and published, e.g., in [5, 6, 7]. For setting the ground and guaranteeing common understanding, some principles and definitions have to be considered.

Alter defines knowledge as "a combination of instincts, ideas, rules, and procedures that guide actions and decisions" [8]. Knowledge of a domain of discourse (discipline), representing that domain's perspective on reality to facilitate reasoning, inferring, or drawing conclusions, is created, represented, and maintained by domain experts using their methodologies, terminologies and ontologies. Initiated by cognitive sciences, knowledge representation and management happens on three levels: the epistemological (cognitive and philosophical) level, the notation (formalization and conceptual) level, and the computational or implementation level [9]. Deploying KR techniques such as frames, rules, tagging, and semantic networks, a good KR has to manage both declarative and procedural knowledge.

For describing any system, a model must be provided as partial representation of reality, focusing just on components, functions, internal and external relationships as well as environmental and contextual conditions the modeler is interested in for the considered business case. In other words, a model is defined as unambiguous, abstract conception of some parts or aspects of the real world corresponding to the modeling goals [10]. An ontology mentioned above is a formal explicit specification of a shared conceptualization of a domain of interest (after Gruber [11]), so describing an ordering system of entities of a domain and their relations. With the development of formalised and accepted ontologies, knowledge can be formally represented. Furthermore, knowledge creation, knowledge management and decision support can be decentralized.

A DSS can be modelled in two ways: a) We can focus on the system's functions by analyzing the system's output in relation to inputs and modifying conditions without detailing the system's architecture defined by structure, functions and interrelations of the system's components. Based on the functional representation of the system, we can decide on inputs and modifiers to get an intended output. This is the black-box approach of system theory, corresponding to the traditional, phenomenologically oriented medicine. b) Alternatively, we can consider the system's architecture to understand and to predict the system's behavior. This describes the white-box approach, represented by systems medicine or systems sciences in general, which also covers systems biology, sys-

[1]Information is the negative value of the logarithm of the probability of occurrence.

[2]Information is a function of the relation between possible answers before and after reception.

[3]Information is a name for the content of what is exchanged with the outer world as we adjust to it and make our adjustments felt upon.

tems pathology, etc., finally enabling the translational medicine approach [6]. As DSS can start with a huge amount of information from the molecule or even elementary particle up to the society, thereafter evidence-based reducing it step by step, we can overcome the so-called "anchoring bias", i.e., the human focus on one or a few pieces of information out of a series of characteristics from the beginning [12]. For prediction and decision support, deterministic, probabilistic, and uncertainty algorithms have to be deployed. With growing knowledge and experiences, the black-box approach can be transformed to the white–box one, exploiting the structure-function relationships of any system. Modeling in the context of personalized, predictive, participative precision medicine is discussed in some details in [6].

3 Knowledge Representation and Management in the DSS Context

The knowledge base can provide knowledge at different level of abstraction and expressivity, ranging from implicit knowledge up to fully explicit knowledge representation, i.e. from natural language up to universal logics as shown in Figure 2. Thereby, the Alternative Assessment, Decision Taking and Decision Review (verification) process steps in the decision making process cycle (Figure 1a) move from assistance through cooperation to automation, i.e. from human being to computer.

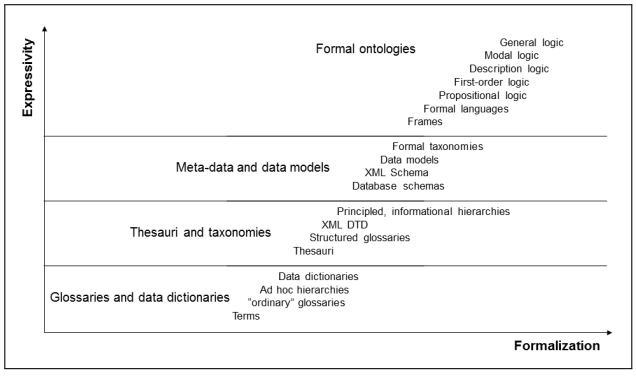

Figure 2: Ontology types for knowledge representation (after [13], changed].

As already stated in [6], KR is first of all a surrogate for the thing itself to enable an entity to determine consequences by thinking (reasoning about the world) rather than acting. KR is a set of ontological commitments to answer the question about the terms to be used to think about the world. KR is a fragmentary theory of intelligent reasoning, expressed in terms of three components: the representation's fundamental conception of intelligent reasoning; the set of inferences; the representation sanctions; and the set of inferences it recommends. KR is a medium for pragmatically efficient computation of thinking and a medium of human expression/language to describe the world (after Davis, Shrobe, and Szolovits [14]).

There are purpose related KR model types such as diagnostic models, connotative models, selective models, analytic models, instructive models, constructive models, or hybrid models.

Knowledge bases may represent inherent rules using set theory, Boolean logic, probability, Bayes rules, or informal logic according to the quality of relations of components and the strategy of the reasoning engine [15]. KR languages and related standards are shortly discussed in [6].

Regarding the language problem in KR with a special focus on formal languages coarsely addressed in Figure 2 already but also regarding the ontological framework needed, the reader is referred to [5].

4 How to Correctly Modeling, Designing and Implementing DSS

A DSS designed to control the behavior of a business system must correctly represent that business system on the basis of the representation constraints appropriate to meet the intended system goals, as mentioned in the former chapter. This is especially challenging for multidisciplinary business systems, where the system must be architecturally correctly modeled not just regarding the structural and functional properties of the system in general, but also with regard to the systems aspects from the different involved domains' perspectives represented through domain-specific concept instances derived from the corresponding domain ontologies. With increasing complexity and flexibility of decision challenges, DSS design has to follow a defined methodology. Here, the multidisciplinary system design approach of the Generic Component Model (GCM) and its Interoperability Reference Architecture Model and Framework – meanwhile standardized at ISO – should be used.

The GCM has been successfully applied in a series of international projects, specifications, and standards [16]–[21]. It offers a system-theoretical, architecture-centric, ontology-driven approach to model translational medicine, so also covering the challenge of knowledge representation. GCM is capable to describe the architecture of any systems, i.e. the composition and decomposition of its components. It allows multi-disciplinary considerations, i.e. the representation of different perspectives (partial systems or domains) of a system as established by domain experts using domain-specific methodologies, terminologies and ontologies.

In the context of intelligent system design it has been accepted that reasoning becomes simpler if the structure of the representation reflects the structure of the portion of reality being reasoned about [22, 23, 24]. Thereby, the representation of GCM components, structured objects, and their behavior, the processes knowledge representation (KR) deals with, must be mastered. In other words, the GCM is also used for modeling representation systems such as languages and ontologies. The three GCM main

axes represent different types of relationships, resulting in different properties of the ontological representation.

The architectural decomposition/composition represents structural and functional specialization or assemblage of components. Regarding the structural relations, constrained "is_a" and "is_part_of" associations are used, while functional relations deploy constrained "uses" – "is_used" associations. The domain axis represents dependencies between different domains' components, while the development process axis according to the ISO/IEC 10746 Reference Model – Open Distributed Processing (RM-ODP) describes transformations between the different RM-ODP views [25] (Figure 3a). Sub-systems derived (specialized) in the decomposition process are frequently represented by specialized sub-domains, leading to different representation styles of structural and behavioral aspects of the components involved (Figure 3b). For enabling a comprehensive representation of multidisciplinary, complex, scalable and flexible DSS systems, universal logic based on type theory is used as overarching framework [26, 27]. For more details in the context of GCM, see, e.g., [28, 29].

5 Standards for Medical Knowledge Representation

For completing the scope of the present paper, detailed information about standards and specifications for medical knowledge representation has been imported from [5] with some extensions. For sharing computable clinical knowledge and enabling intelligent cooperation in distributed environments, a common language for specifying expressions and criteria is inevitable. Therefore, the aforementioned principles and solutions for KR must be standardized. This is a basic requirement for all presented levels of KR from the high level and generic up to domain- and application-specific ones, thereby also developing defacto standards for corresponding tooling. Beside the basic standards tackling the challenge of KR, there are some health-specific ones addressed in the following. There are KR expression languages for guidelines representation and processing not considered in this paper because of the lack of international standardization. Here PROforma [30, 31], Asbru [32, 33], EON [34, 35] have to be mentioned.

5.1 Arden Syntax

Arden Syntax has been developed for sharing medical knowledge stored in technically differently implemented knowledge bases. It could be called a technology-independent (or platform-independent) knowledge exchange format. The Arden Syntax represents this knowledge using frame logics. Arden Syntax encodes medical knowledge about individual decision rules in knowledge base form as self-contained Medical Logic Modules (MLMs), which can be embedded into proprietary clinical information systems. The MLMs are implemented as event-driven alerts or reminders.

Expressed as semiformal language, MLMs contain three slots or categories: the Maintenance Category (identifying the module, author, version, evidence level, etc.), the Knowledge Category (medical concept represented), and the Library Category (references/evidences). The knowledge category has a data slot on the one hand and evocation, logic, and action slots on the other hand. The latter specify the aforementioned events that trigger the evocation of the MLM, the logical criterion evaluated, and the action performed when the logical criterion is met. These knowledge-category components define the logical rule that the MLM specifies. The concept representation for describing medical conditions or recommendations contains a production rule and a procedural formalism, enabling a logical decision. Processes can be managed by chaining MLMs

Arden Syntax has been originally developed by New York Columbia Presbyterian Medical Center (CPMC) and IBM Health Industry Marketing in Atlanta, and thereafter wider used at Regenstrief Institute as well as within the HELP (Health Evaluation through Logical Processing) system at Salt Lake City LDS Hospital. Advanced applications using Arden Syntax for generating clinical alerts and reminders, interpretations, diagnoses, screening for clinical research studies, quality assurance functions, and administrative support in so-called event monitors are meanwhile globally deployed, as also demonstrated in this volume.

Arden Syntax has been standardized first by ASTM (American Standards for Testing and Materials) [36] and thereafter at HL7 [37]. Since 2014, Arden Syntax Version 2.10 and a first version tackling fuzzy logic for production rule representation are available. It is a specification compliant with the HL7 RIM.

5.2 GELLO

GELLO is a typed object-oriented standard query and expression language that provides a framework for management and processing of clinical data. Based on the Object Management Group (OMG) Object Constraint Language (OCL), GELLO enables the specification of decision criteria, algorithms and constraints on data and processes [38]. By that way, it provides a standardized framework for implementing DSSs. Therefore, GELLO is sometimes also called an object-oriented clinical decision support language [39].

The GELLO language can be used to build queries to extract and manipulate data from medical records and construct decision criteria by building expressions to correlate particular data properties and values. These properties and values can then be used in decision-support knowledge bases that are designed to provide alerts and reminders, guidelines, or other decision rules [37]. For this purpose, GELLO expresses logical conditions and computations in an standardized interchange format for mod-

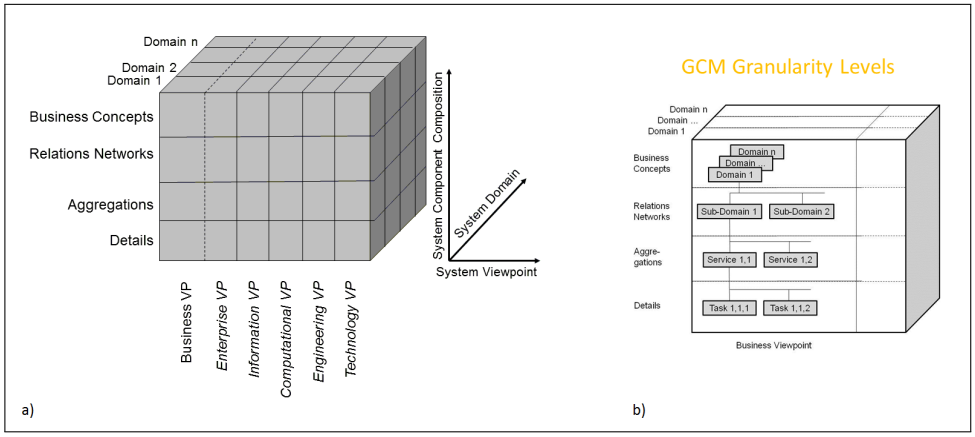

Figure 3: The Generic Component Model.

eling clinical guidelines, the GuideLine Interchange Format, v. 3 (GLIF3) [40]. Furthermore, it can be used for processing constraints, validation and calculated fields in Archetype data entries. It is also used to create complex data series for graphing or statistical analysis. For extracting data from any clinical database, a RIM-compliant virtual medical record has been defined as a mediator – similar to the RIM-based HL7 messaging framework enabling the communication of data between different health information systems. Thus, GELLO goes beyond the Arden Syntax which is limited to representing clinical rules.

GELLO is an HL7 International standard. Since 2010, GELLO Release 2 is available as formal, ANSI-approved specification. There are powerful GELLO compilers on the market, e.g., the Medical-Objects product [38].

5.3 GLIF

The GuideLine Interchange Format (GLIF) has been jointly developed at Stanford University, Brigham and Women's Hospital, and Columbia University to express and to share guidelines for prevention, diagnosis work-up, treatment, and patient-management processes (clinical pathways). They can be used as centrally stored sharable resource of knowledge, but also as directly executable guidance in response to network-based queries. Meanwhile, further institutions have joint the team.

GLIF3 [40] is an object-oriented expression and query language. Representing the description of complex multi-step guideline knowledge, the GLIF language can be also be translated into other languages established to execute clinical knowledge such as Arden Syntax. Using specific application interfaces (APIs), network-based clinical applications can directly access central decision support services executing approved guidelines based on the given data sets.

The GLIF3 specification consists of an extensible object-oriented model and a structured syntax based on the Resource Description Framework (RDF). GLIF3 enables encoding of a guideline at three levels: a conceptual flowchart, a computable specification that can be verified for logical consistency and completeness, and an implementable specification to be incorporated into local information systems. The GLIF3 model is represented using UML (Unified Modeling Language). Additional constraints are expressed in OCL. For enabling the integration into information system, GLIF uses HL7 RIM classes and data types. While Arden Syntax follows a bottom-up approach vs the top-down approach of GLIF, both specifications are complementary for representing medical knowledge for clinical decision support.

GLIF3 is application independent, executable, can be easily integrated into clinical information systems, extensible, and offers a layered approach for managing the complexity of knowledge. It has been standardized at HL7 International. Corresponding tools have been developed, e.g., by the InterMed Collaboratory [41].

5.4 Archetypes

Based on specification provided by the EU project Good European Health Record (GEHR), Australia with Thomas Beale as main actor developed the Good Electronic Health Record (GEHR), which meanwhile evolves under the auspices of the openEHR Foundation [42].

The Archetype approach supports semantically enriched EHR systems by encapsulating the domain expert's knowledge in archetypes, defined and expressed using the Archetype Definition Language (ADL) [43]. ADL is a member of the OCL family. The Archetype model provides a constraint data model, thereby reflecting the domain experts' view. The structural Reference Model used is documentation specific, tackling storage and retrieval of information. Thus, it represents an informational perspective contrary to clinical facts described by translational medicine and sophisticated medical ontologies [44]. Using the Archetype Query Language (AQL) [45], clinical information can be consistently and easily retrieved with high improve recall and precision, thereby constraining the data object instances according to the Archetype definition. Archetypes represent clinical knowledge using frame logics. The Header part contains identifying in-

formation and meta-data including external ones. The Body part contains the clinical concept represented. The Terminology part finally references Archetype classes to standard terminologies, by that way supporting harmonization between different environments.

Archetypes and the Archetype approach have been standardized at ISO and CEN in the context of the ISO/CEN 13606 "EHR communication" standards series. openEHR offers freely available ADL parser [42].

5.5 HL7 Clinical Statement Model

HL7 International has developed the Clinical Statement Model for representing clinical concepts in a single message or document according to the HL7 Version 3 methodology. For sharing documented clinical information in a standardized way, HL7 developed the Clinical Document Architecture (CDA), representing clinical documents as structured, persistent, human-readable and machine-processable objects for a specific purpose. A CDA document consists of the CDA Header and the CDA Body. The latter contains information about CDA Structure, CDA Entries and CDA External References. HL7 v3 CDA documents and messages are encoded using the meta-language Extensible Markup Language (XML). They derive their machine processable semantics from the HL7 RIM and use the HL7 Version 3 data types and class structures, thereby providing a mechanism for the incorporation of concepts from standard coding systems such as SNOMED CT and LOINC.

In an evolutionary process, different levels of granularity for encoding information into machine-processable data have been defined, represented as different Releases of CDA. The CDA interoperability level enhances with more structured CDA Releases from R1 up to R3, as roughly explained in the following. In R1, just the Header has been fully specified, while the body is represented in just one block. In R2, the Body has been separated into tagged sections for diagnosis and treatment. In R3, the Body part will be structured up to the level of atomic concepts. HL7 Templates are a constraint on the CDA R2 object model and/or against a Clinical Statement Model [37].

5.6 The Clinical Information Modeling Initiative

The Clinical Information Modeling Initiative (CIMI) is an international action to provide a common format and a common development process for detailed specifications for the representation of health information to enable creation and sharing of semantically interoperable information in health records, messages, documents, and for secondary data uses. CIMI is mainly based on the aforementioned Archetype approach. Additionally to the Archetype Object Model [46] and the expression means of ADL [43], an extended Reference Model [47] and the representation of the entire development process using UML and the SOA framework will be deployed. For more information, the reader is referred to [44].

5.7 HL7 FHIR Resources

The newest HL7 standard FHIR (Fast Healthcare Interoperability Resource) specifies reusable and easily implementable component enabling information exchange between newly developed or legacy health information systems [37]. Using Web representation and implementation tools such as XML, JSON (Java Script Object Notification) and the RESTful approach simplifies specification and implementation of FHIR resources. RESTful (Representational State Transfer conformant) (REST) approach describes a simplified architecture for Web services accessed through a unified interface by using a stateless protocol. Therefore, FHIR resources consist of a URL (Uniform Resource Locator) identifying the component, common metadata, a human-readable XHTML (Extensible Hypertext Markup Language) summary, a set of defined common data elements and finally an extensibility framework for adapting the resource to use-case-specific or national needs.

6 Discussion

The paper discusses KR and KM as key challenges of DSS architectures. The ongoing organizational, methodological and technological paradigm changes especially in, but not limited to, healthcare systems (see Table 1) also impact analysis, design and implementation of DSS.

From an IT perspective on the pathway to personalized and ubiquitous care technology paradigm, the IT paradigm changes can be classified according to Figure 4 [48].

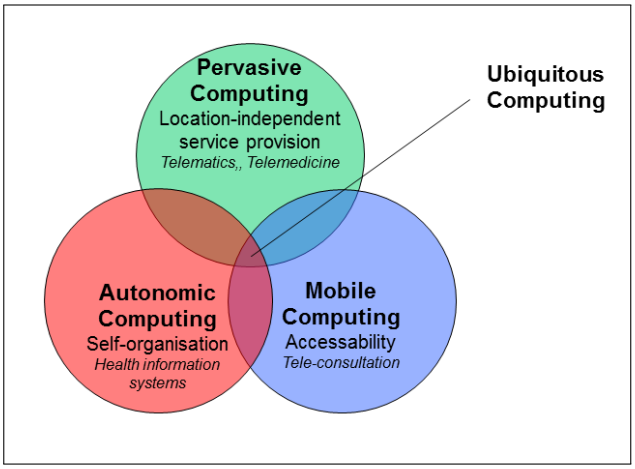

Figure 4: Ubiquitous computing paradigm.

So, DSS architectures will move to mobile, pervasive and autonomous systems as well. In detail, technological paradigm changes towards distributed systems, mobile

Table 1: Paradigm changes in health systems.

Organization	Methodology	Technology
Organization-centric care ↓ Process-controlled care (DMP) ↓ Person-centric care	General care addressing health problems (one solution fits all) → Phenomenological Approach ↓ Dedicated care (stratification of population for specific clinically relevant conditions) → Evidence-Based Medicine ↓ Personalized, preventive, predictive and participative care considering individual health state, conditions and contexts (stratification of population by risk profiles) → Systems Medicine, from art to multidisciplinary science, from elementary particle to society	Mainframe ↓ Client/Server ↓ Internet ↓ Distributed systems, mobile technologies, nano- and molecular technologies, knowledge representation & management, KI, Big Data & Business Analytics, Cloud Computing, Social Business

technologies, nano- and molecular technologies, knowledge representation & management, artificial intelligence (AI), big data & business analytics, cloud computing, and social business will also determine design and implementation of DSS. In consequence for example, intelligent molecules will correctly interpret a body component's modification (diagnosis) and intendedly interact with tissues (therapy).

In the context of the aforementioned paradigm changes in DSS design and implementation, new security and privacy challenges as well as safety issues have to be considered and solved [49].

Acknowledgments

The author is indebted to thank his colleagues and friends from the Standards Developing Organizations ISO TC 215, CEN TC 251, HL7 International, and especially Mathias Brochhausen, University of Arkansas of Medical Sciences, Little Rock, Arkansas, U.S.A., for continuous support and excellent cooperation.

References

[1] Wikipedia. Decision Support Systems. (last access 15 November 2016)

[2] UMass, 7 Steps to effective decision making. UMass, Dartmouth. `www.umassd.edu` (last access 15 November 2016).

[3] Van Bemmel J., Musen M. (eds) Handbook of Medical Informatics. Heidelberg: Springer; 2002.

[4] Haettenschwiler P. Neues anwenderfreundliches Konzept der Entscheidungsunterstützung. Gutes Entscheiden in Wirtschaft, Politik und Gesellschaft. Zurich: vdf Hochschulverlag AG; 1999, 189-208.

[5] Blobel B. Knowledge Representation and Management Enabling Intelligent Interoperability – Principles and Standards. Stud Health Technol Inform. 2013;186:3-21.

[6] Blobel B. Translational Medicine Meets New Technologies for Enabling Personalized Care. Stud Health Technol Inform. 2013;189:8-23.

[7] Blobel B. Analysis, Design and Implementation of Secure and Interoperable Distributed Health Information Systems. Series Studies in Health Technology and Informatics, Vol. 89. Amsterdam: IOS Press; 2002.

[8] Alter S. Information Systems – A Management Perspective. Reading: Addison-Wesley Publishing Co, Inc.; 1991.

[9] Dahn, BI, Dörner H, Goltz H-J, Grabowski J, Herre H, Jantke KP, Lange S, Posthoff C, Thalheim B, Thiele H. Grundlagen der Künstlichen Intelligenz. Berlin: Akademie-Verlag; 1989.

[10] M. Lankhorst et al., Enterprise Architecture at Work - Modelling, Communication and Analysis, 2nd Edition. The Enterprise Engineering Series, Dordrecht Heidelberg London New York: Springer; 2009.

[11] Gruber T. Ontology Definition. In: Liu L and Özsu MT (eds) The Encyclopedia of Database Systems. New York: Springer; 2009.

[12] Journal of Translational Medicine. `http://www.translational-medicine.com/` (last access 15 November 2016)

[13] Rebstock M, Fengel J, Paulheim H. Ontologies-Based Business Integration. Berlin – Heidelberg: Springer-Verlag; 2008.

[14] Davis R, Shrobe H, and Szolovits P. What is a Knowledge Representation? AI Magazine 1993;14,1:17-33.

[15] Spooner SA. Mathematical Foundations of Decision Support Systems. In: Berner ES (ed) Clinical Decision Support Systems – Theory and Practice. 2nd Edition. New York: Springer Science+Business Media; 2007.

[16] De la Cruz E, Lopez DM, Blobel B. A Reference Architecture for Sharing Documents in Colombia. European Journal for Biomedical Informatics 2012;8,3:en11-en17.

[17] Vida M, Stoicu-Tivadar L, Blobel B, Bernad E. Modeling the Framework for Obstetrics-Gynecology Department Information System. European Journal for Biomedical Informatics 2012;8,3:en57-en64.

[18] Yildirim Yayilgan S, Blobel B, Petersen F, Hovstø A, Pharow P, Waaler D, Hijazi Y. An Architectural Approach to Building Ambient Intelligent Travel Companions. International Journal of E-Health and Medical Communications 2012;3,3:86-95.

[19] Ruotsalainen P, Blobel B, Seppälä A, Sorvari H, Nykänen P. A Conceptual Framework and Principles for Trusted Pervasive Health. J Med Internet Res 2012;14,2:e52. URL: `http://www.jmir.org/2012/2/e52/` (last access 15 November 2016).

[20] Lopez DM, Blobel B. A development framework for semantically interoperable health information systems. International Journal of Medical Informatics 2009;78,2:83-103.

[21] Bernal JG, Lopez DM and Blobel B. Architectural Approach for Semantic EHR Systems Development Based on Detailed Clinical Models. Stud Health Technol Inform. 2012; 177:164-169.

[22] Szolovits P. Artificial Intelligence and Medicine. Chapter 1 in Szolovits, P. (ed) Artificial Intelligence in Medicine. Boulder, Colorado: Westview Press; 1982.

[23] Bobrow DG, and Winograd T. An Overview of KRL, a Knowledge Representation Language. Technical Report AIM-293, Stanford Artificial Intelligence Lab., Stanford, Ca; 1976. ftp://reports.stanford.edu/pub/cstr/reports/cs/tr/76/581/CS-TR-76-581.pdf (last access 15 November 2016).

[24] Szolovits P, Hawkinson L, and Martin WA. An Overview of OWL, a Language for Knowledge Representation. In: Rahmstorf G, and Ferguson M (eds) Proceedings of the Workshop on Natural Language Interaction with Databases, International Institute for Applied Systems Analysis, Schloss Laxenburg, Austria, 10 Jan 1977.

[25] International Organization for Standardization. ISO/IEC 10746 Information technology – Reference Model – Open Distributed Processing. Geneva; 1996. `www.iso.org` (last access 15 November 2016).

[26] Bloe R, Kamareddine F, Nederpelt R. The Barendregt Cube with Definitions and Generalized Reduction. Information and Computation 1996; 126,2:123–143.

[27] Kamareddine F, Laan T, Nederpelt R. A Modern Perspective on Type Theory. New York: Kluwer Academic Publishers; 2004.

[28] Blobel B, Pharow P. Analysis and Evaluation of EHR Approaches. Methods Inf Med 2009;48,2:162-169.

[29] Blobel B. Architectural approach to eHealth for enabling paradigm changes in health. Methods Inf Med 2010;49,2:123-134.

[30] Fox J, Rahmanzadeh A. Disseminating medical knowledge: the PROforma approach. Artificial Intelligence in Medicine 1998;14:157-181.

[31] PROforma. `www.openclinical.org/gmm_proforma.html` (last access 15 November 2016).

[32] Shahar Y, Miksch S, Johnson P. The Asgaard Project: A Task-Specific Framework for the Application and Critiquing of Time-Oriented Clinical Guidelines. Artificial Intelligence in Medicine 1998;14:29-51.

[33] Asbru. `www.openclinical.org/gmm_asbru.html` (last access 15 November 2016).

[34] Tu SW, Musen MA. A Flexible Approach to Guideline Modeling. Proc AMIA Symp 1999:420-424.

[35] EON. `www.openclinical.org/gmm_eon.html` (last access 15 November 2016).

[36] American Standards for Testing and Materials (ASTM). `www.astm.org` (last access 15 November 2016).

[37] Health Level 7 International, Inc. `www.hl7.org`

[38] Sordo M, Boxwala AA, Ogunyemi O, Greenes RA. Description and status update on GELLO: a proposed standardized object-oriented expression language for clinical decision support. Stud Health Technol Inform. 2004;107,1:164–168.

[39] Medical-Objects. GELLO.org. `www.gello.org` (last access 15 November 2016).

[40] Boxwala AA, Peleg M, Tu S, Ogunyemi O, Zeng QT, Wang D, Patel VL, Greenes RA, Shortliffe EH. GLIF3: A Representation Format for Sharable Computer-Interpretable Clinical Practice Guidelines. J Biomed Inform 2004 Jun;37,3:147-61.

[41] The InterMed Collaboratory. `http://mis.hevra.haifa.ac.il/~morpeleg/Intermed` (last access 15 November 2016).

[42] openEHR Foundation. `www.openehr.org` (last access 15 November 2016).

[43] Beale T, Heard S. Archetype Definition Language ADL 1.5. openEHR Foundation, January 2012, `www.openehr.org` (last access 15 November 2016).

[44] Blobel B, Goossen W, Brochhausen M. Clinical Modeling – a Critical Analysis. International Journal of Medical Informatics 2014;83,1:57-69

[45] Beale T. openEHR Archetype Query Language Description. `http://www.openehr.org/wiki/display/spec/Archetype+Query+Language+Description` (last access 15 November 2016).

[46] Beale T. Archetype Object Model AOM 2.1. openEHR Foundation, January 2012, `www.openehr.org` (last access 15 November 2016).

[47] CIMI Reference Model Report, Draft V 0.3, May 2012

[48] Blobel B, Pharow P, Norgall T. How to Enhance Integrated Care towards the Personal Health Paradigm? Stud Health Technol Inform. 2007;129:172-176.

[49] Blobel B, Lopez DM, Gonzalez C. Patient privacy and security concerns on big data for personalized medicine. Health and Technology 2016;6,1:75-81.

Parametric vs. Nonparametric Regression Modelling within Clinical Decision Support

Jan Kalina[1], **Jana Zvárová**[1,2]

[1] Institute of Computer Science CAS, Prague, Czech Republic

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

Abstract

Decision support systems represent very complicated systems offering assistance with the decision making process. Learning the classification rule of a decision support system requires to solve complex statistical task, most commonly by means of classification analysis. However, the regression methodology may be useful in this context as well. This paper has the aim to overview various regression methods, discuss their properties and show examples within clinical decision making.

Keywords

Decision Support Systems, Decision Rules, Statistical Analysis, Nonparametric Regression

Correspondence to:

Jan Kalina
Institute of Computer Science CAS
Address: Pod Vodárenskou věží 2, 182 07 Praha 8
E–mail: kalina@cs.cas.cz

IJBH 2017; 5(1):21–27
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Clinical Decision Support Systems

Decision support systems can be characterized as very complicated systems offering assistance with the decision making process. Using data and knowledge as main sources to obtain information [28], they are capable to solve a variety of complex tasks, to analyze different information components, to extract information of different types, and deduce conclusions from them. This section critically discusses principles of clinical decision support systems and reveals the aims of the paper.

As we have formulated in review papers on clinical decision making [11, 12], decision support systems in medicine compare different possibilities for the diagnosis, therapy or prognosis in terms of their risk. They represent an inherent tool of e-health technologies in the complex process of searching for an appropriate therapy. In practice, there exist specialized decision support systems for diagnosis and therapy in individual medicine fields and also specialized prescribing decision support systems. There has been less attention paid to decision support systems for prognosis, while there are still obstacles to apply decision support systems in healthcare routinely, although diagnostics and therapy would greatly benefit from reliable interdisciplinary and multidisciplinary systems.

Decision support systems have acquired an established place in various fields of clinical medicine with a certified ability to assist physicians with the decision making. Several studies proved that a decision support system can be useful for improving the quality of provided care, preventing errors, reducing financial costs and saving human resources. The system may bring the physician more comfort, a higher effectiveness and more time for the patient and also a reduction of errors. It saves also significant financial costs. It may be especially favorable during stress or for treating complicated patients. Particularly (but not only) a less experienced physician may benefit from using a decision support system, which exploits the level of knowledge reflecting the latest developments even in a narrow domain of medicine.

We endorse the concept of information-based medicine [3] for describing the future ideal state of health care. It goes far beyond the current clinical practices and reflects its constant development and enrichment by quickly emerging new results of basic research. Information-based medicine represents a new perspective paradigm in medicine, also overcoming the limitations of the popular evidence-based medicine. Although difficult to define it precisely, the concept of information-based medicine describes the effort to reach a distant aim, namely to transform the evidence for the (imaginary) averaged patient towards a real individual patient based on his/her individual data with clinical as well genetic or metabolic parameters measured by new technology.

Information-based medicine requires to extract information from massive data sets. Learning the classification rule of a clinical decision system is a particularly difficult task, commonly solved by methods of classification analysis. However, also the regression methodology may be useful in this context. This paper has the aim to overview various regression methods, discuss their properties and show examples of applications of regression models in clinical decision making. On the whole, regression methods within the task of clinical decision support allow to solve the following tasks.

- Description of the structure of the data across the whole clinical study,
- Prediction of a (continuous) response for an individual patient,
- Classification (if the regression method is suitable for such task),
- Hypothesis testing e.g. test of a treatment effect in the context of generalized linear models,
- Measuring the correlation between two variables or sets of variables (e.g. for dimensionality reduction purposes [10]).

This paper has the following structure. An overview of important classes of regression methods is presented in Section 2. The subsequent sections aim at discussing advantages and limitations of individual methods from the perspective of clinical decision support. Generalized linear models, whichrepresent the most common statistical methodology in biomedical applications, are discussed in Section 3. The linear regression model as a special case follows in Section 4. Other regression tools, which are much more popular in computer science (machine learning, predictive data mining) compared to statistics, can be characterized as nonparametric regression techniques (Section 5), although this concept is not unambiguously accepted.

2 Overview of Regression Methods

The aim of regression methodology is to model a continuous variable taking into account one or more independent variables (regressors). In biomedical applications, its tools are commonly used to predict values of an important variable for individual patients based on one or more (continuous and/or categorical) variables. Thus, regression modelling is a tool allowing to contribute to a targeted decision making.

The two communities of statisticians and computer scientists have the tendency to use techniques developed in their own environment and we pay attention to comparing advantages and disadvantages of these two rather different worlds. Unfortunately, comparisons of statistical methods with those proposed within the framework of data mining or machine learning are rather rare. Here, we bring together arguments in favor of statistical methods compared to data mining approaches. Thus, the overview based on our rich experience with regression modeling of biomedical data is rather unique.

Statistics as a discipline allows to take uncertainty into account within the process of inductive reasoning. The advantages of statistics over computer science, which has a tendency to combine ad hoc approaches ignoring the assumptions, are primarily:

- Theoretically proven optimality of methods,
- Ability to capture the multivariate structure of data.

We distinguish between the following main classes of regression models:

- Parametric regression models
  - Linear regression model
  - Nonlinear regression model
  - Generalized linear models
- Nonparametric regression (regression curve estimation, function approximation)
  - Regression trees
  - Multilayer perceptrons
  - Support vector regression
  - Kernel-based methods (kernel estimation of regression curve, or shortly kernel estimation)
  - Taut string [4]
  - Regularization networks

While nonparametric methods are more flexible, we present a list of important advantages of parametric regression models.

- No overfitting,
- Comprehensibility,
- Diagnostic tools and remedies,
- Efficient computation,
- Modifications for a high dimension (LASSO),
- Modifications robust to outliers,
- Available hypothesis tests,
- Confidence interval for parameter estimates,
- Confidence band (region) for the whole regression curve (or line).

It remains to be an open problem for the field of metalearning how to select the most suitable method for given data depending on their characteristics and also depending on the application domain.

3 Generalized Linear Models

This section is devoted to important examples of generalized linear models (GLM) [13], including the logistic regression, multinomial logistic regression, and LASSO-type estimation in logistic regression models. The last two models are natural generalizations of the basic logistic regression principle.

3.1 Logistic regression

Logistic regression is the most commonly used regression method in biomedical applications. It allows to predict a binary response and therefore may be used directly as a classification method.

Let us assume the total number of n measurements. A binary response Y_i is observed for each of them, while Y_i is considered to be a random variable following a binomial distribution with the probability of success π_i. Particulary, it is assumed that

$$\log \frac{\pi_i}{1-\pi_i} \tag{1}$$

depends linearly on the regressors. The parameters are estimated by the maximum likelihood principle and the computation is performed by means of an iterative algorithm (iterative reweighted least squares).

3.2 Multinomial logistic regression

We recall the multinomial logistic regression, which is an extension of the logistic regression to a model with a response with several different levels (categories). We assume the total number of n measurements

$$(X_{11},\ldots,X_{1p})^T,\ldots,(X_{n1},\ldots,X_{np})^T. \tag{2}$$

Each of them belongs to one of K different groups. The index variable $Y = (Y_1,\ldots,Y_n)^T$ contains the index $1,\ldots,K$ corresponding to the group membership of the n measurements. We consider a model assuming that Y follows a multinomial distribution with K categories.

Definition 1 *The data are assumed as K independent random samples of p-variate data. The multinomial logistic regression model is defined as*

$$\log \frac{P(Y_i=k)}{P(Y_i=K)} = \beta_{k1}X_{i1}+\cdots+\beta_{kp}X_{ip} \tag{3}$$

for $i = 1,\ldots,n$ and $k = 1,\ldots,K-1$, where where $P(Y_i = k)$ denotes the probability that the observation Y_i belongs to the k-th group.

An important generalization of the logistic regression is the ordinal logistic regression. It is one of GLMs assuming the response to be an ordinal variable. Its main difference from the logistic regression is replacing the probability $\mathsf{P}(Y = k)$ by modeling $\mathsf{P}(Y \leq k)$ for each $k \in \{1,\ldots,K-1\}$.

3.3 LASSO-logistic regression

LASSO-type estimators have been denoted as the most common classification method for molecular genetic data [29] and especially the LASSO estimator of parameters of the logistic regression [21] is important for solving classification problems for high-dimensional data. The estimator of the parameters $\beta = (\beta_1,\ldots,\beta_p)^T$ is obtained as the (constrained) maximum likelihood estimator with a fixed $t > 0$, maximized under the constraint

$$\sum_{j=1}^{p} |\hat{\beta}_j| \leq t \tag{4}$$

on the estimates $\hat{\beta}_j$ of regression parameters with a fixed value of the tuning parameter $t > 0$.

In spite of its popularity, a criticism of the LASSO has been reported [27] concerning e.g.

- High instability of the variable selection,
- Lack of satisfactory asymptotic properties,
- Too restrictive conditions on the design matrix,
- Low signal to noise ratios,
- Problems with measurement errors (technical noise) in the covariates,
- Instability for heterogeneous sample populations,
- Bias from unmeasured confounders.

4 Linear Regression

We consider the standard linear regression model

$$Y = X\beta + e \tag{5}$$

with n observations and p regressors. We discuss advantages and limitations of the least squares, LASSO, and partial least squares estimators from biostatistical perspectives. Another estimator of parameters of linear regression is the linear support vector regression, which will be treated separately later.

4.1 Least squares

Although linear regression is not suitable for classification tasks, some authors have used it e.g. in the analysis molecular genetic data, as overviewed in [8]. A broad scope of various practical aspects of regression modeling of medical data from the practical point of view is presented in an excellent book [24]. Specific questions which must be taken into account particularly in the analysis of biomedical data include:

- Transformation of covariates,
- Selection of the most relevant covariates (variable selection),
- Modeling nonlinear effects,
- Power calculations,
- Sample size calculations,
- Regression strategies for clustered or longitudinal data,
- Measurement errors or missing values,
- Techniques for measuring the prediction performance.

Because the least squares estimator can be derived by the maximum likelihood principle for Gaussian errors, it is quite flexible and allows modifications for specific violations of the assumptions, which may be detected by available diagnostic tools based on residuals of the ordinary least squares. Thus, we can say that linear model is more than a single universal model. Instead, we understand the least squares methodology as a set of approaches to modify the standard ("vanilla") model to reflect the context and variety of aspects of the data in relationship to the statistical assumptions. Important modifications of the standard least squares estimator include

- Cochrane-Orcutt transformation,
- Heteroscedastic regression,
- Instrumental variables estimator,
- LASSO estimator,
- Robust regression.

A direct generalization of the least squares to nonlinear regression with a specified (but nonlinear) parametric model is the nonlinear least squares estimator.

4.2 LASSO in linear regression

The least absolute shrinkage and selection operator (LASSO) estimator can be described as a popular procedure for simultaneous estimation and variable selection in the linear regression model, which has a high predictive ability also for high-dimensional data. LASSO is based on a combination of the classical estimator of regression parameters with additional constraints, which allow to reduce the dimensionality in an intrinsic way. Therefore, the method is popular in bioinformatics.

The estimator combines the idea of minimizing the classical sum of squared residuals with a requirement (penalization) on the sum of absolute values of estimated parameters not to exceed a constant (say) λ. This reduces the number of regressors with non-zero parameters. The method combines two principles, namely

- Shrinkage of regression parameters towards zero,
- Elimination of regressors.

The estimator is obtained as a solution of an optimization task, modifying the classical least squares estimator.

Formally, the estimator of the regression parameters $\beta = (\beta_1, \ldots, \beta_p)^T$ is obtained as the solution of minimization of squared residuals under the constraint (4). Let us use the notation $(.)_+$ for the positive part, i.e.

$$(f(x))_+ = \max\{f(x), 0\}. \tag{6}$$

Assuming $X^T X = I$ in (5), the LASSO estimator has an explicit expression in the form

$$\begin{aligned} \hat{\beta}_j &= \mathsf{sgn}(b_j^{LS}) \left(|b_j^{LS}| - \lambda\right)_+ = \\ &= \mathsf{sgn}(b_j^{LS}) \max\left\{|b_j^{LS}| - \lambda, 0\right\}, \end{aligned} \tag{7}$$

for $j = 1, \ldots, p$, where sgn denotes the sign function and

$$b_{LS} = (b_1^{LS}, \ldots, b_p^{LS})^T \tag{8}$$

the least squares estimator of β. The constant λ plays the role of a regularization parameter and is commonly found by a cross validation in practice. We can express $\hat{\beta}_j$ for $j = 1, \ldots, p$ as

$$\hat{\beta}_j = b_j^{LS} - \lambda, \quad \text{if } \lambda < b_j^{LS}, \tag{9}$$

$$\hat{\beta}_j = 0, \quad \text{if } -\lambda \leq b_j^{LS} \leq \lambda, \tag{10}$$

$$\hat{\beta}_j = b_j^{LS} + \lambda, \quad \text{if } b_j^{LS} < -\lambda. \tag{11}$$

4.3 Partial least squares

Partial least squares (PLS) regression was originally proposed for chemometrics applications [2] and later has spread to a variety of fields. It is a linear regression method allowing to consider a multivariate response Y. Various studies have shown the usefulness of the PLS under multicollinearity or in situations with the number of regressors exceeding the number of observations, particularly under the condition that random errors in the regression model have a large variability. In some applications, however, explaining the response by linear combinations of genes may not be biologically interpretable.

The computation is based on searching for latent variables extra for regressors and for a multivariate response Y, allowing to reduce the dimension intrinsically. The partial least squares method reduces the dimension to a certain number of components. Linear combinations of regressors are found, which contribute the most to explaining the variability within the matrix of regressors X. In the same way, linear combinations of the response variables are found contributing the most to explaining the variability within Y. Finally, the method explains the transformed response by the transformed regressors.

The PLS can be used also to solve classification problems [2, 15]. One possibility is to perform it in the standard way and apply one of standard classifiers to learn a classification rule over the components. Especially linear discriminant analysis (LDA) is suitable for this purpose, because it is also a linear method [17]. A preferable possibility is however to use an alternative version denoted as discriminant PLS (D-PLS). To explain the main advantage of D-PLS, its performance is ensured to be better or at least no worse than principal component analysis (PCA) for dimension reduction with the goal of achieving classification, which follows from a direct connection between the PLS and LDA. D-PLS is especially suitable when within-groups variability dominates the among-groups variability in the data.

4.4 Robust regression

The least squares estimator is too vulnerable to the presence of outlying measurements in the data. Therefore, its robust alternatives have been proposed. M-estimates or R-estimates represent important classes of robust estimators in the linear regression model [7, 19]. However, they do not possess robustness properties in terms of the breakdown point, which is a statistical measure of sensitivity against severe outliers in the data.

There are a few robust regression estimators with a high breakdown point, while is the least trimmed squares (LTS) estimator is the most common. In addition, its modification called least weighted squares estimator (LWS) has been proposed [26] which assigns nonnegative weights to individual observations based on ranks of residuals. The LWS estimator seems more promising than the LTS estimator for several reasons, mainly a high efficiency for normal (non-contaminated) samples, robustness to heteroscedasticity and local robustness to small changes of the data. Particularly, the LWS estimator attains a 100 % asymptotic efficiency of the least squares under Gaussian errors, if the data-dependent adaptive weights are used. The computation of LWS is however very tedious using an approximative algorithm comparing various possible permutations of the weights.

We have successfully used the LWS estimator in medical image analysis [9] or to define similarity measures suitable for reducing the dimensionality of molecular genetic data [10].

5 Nonparametric Regression

Nonparametric approaches are not based on a particular model, but the predicted value for each value of the regressor $X \in \mathbb{R}^p$ is an unknown parameter. In the regression framework, the task of nonparametric regression can be described as the function estimation, i.e. to estimate a continuous function of the regressors and the value of the function for each possible value of the regressors is an unknown parameter. In clinical applications, it may be useful for prediction of a given variable (e.g. a drug level) for an individual patient based on given values of regressors. We give an overview of some of the numerous available methods.

5.1 Regression trees

Regression trees represent a class of very traditional methods, which can be described as nonparametric. Their comprehensibility makes them a popular tool in biostatistics. The development of regression trees within the last 50 years was summarized in [14]. Their numerous algorithms have been well described and investigated theoretically [16]. Currently, their ensembles called random forests are popular, which are able to overcome the instability of individual trees. CART represents the most common algorithm for constructing a regression tree in the top-bottom approach, searching for individual variables with the greatest relevance for explaining the variability of the response.

5.2 Multilayer perceptrons

In the task of nonparametric regression, multilayer feedforward perceptrons represent the most commonly used form of neural networks.

Most commonly, they exploit the back-propagation algorithm minimizing the total error computed across all data values of the training data set. The algorithm is based on the least squares method, which is optimal (only) for normally distributed random errors in the data [18]. After an initiation of the values of the parameters, the forward propagation is a procedure for computing weights for the neurons sequentially in particular hidden layers. This leads to computing the value of the output and consequently the sum of squared residuals computed for the whole training data set. To reduce the sum of squared residuals, the network is sequentially analyzed from the output back to the input. Particular weights for individual neurons are transformed using the optimization method of the steepest gradient.

An intensive attention has been paid to the study of upper bounds for the approximation error of multilayer perceptrons giving arguments in favor of using them even in high-dimensional cases [5]. Such theoretical considerations are however very distant from the original biological motivation for artificial neural networks. In addition, there are no diagnostic tools available, which would be able to detect a substantial information in the residuals, e.g. in the form of their dependence, heteroscedasticity, or systematic trend.

Neural networks can be described as black boxes, i.e. it is impossible to explain why the algorithm yields particular results, although comprehensibility would be very desirable in a variety of fields in which they have been successfully applied, e.g. in phoniatrics [20] or chemical catalysis [1] to give a few examples. The extremely uninterpretable character is an attribute of deep multilayer perceptrons (deep networks), which have recently become popular without investigating any specific methodological issues regarding their learning.

5.3 Radial basis function networks

Radial basis function network represent another model for approximating a continuous nonlinear function. In contrary to multilayer perceptrons, the input layer transmits a measure of distance of the data from a given point to the following layer. Such measure is called a radial function. Typically, only one hidden layer is used and an analogy of the back-propagation is used to find the optimal values of parameters. The output of the network for

a given observation $\mathbf{x} \in \mathbb{R}^p$ has the form

$$f(\mathbf{x}) = \sum_{i=1}^{s} w_i \exp\{-\beta||\mathbf{x}-\mathbf{c}_i||^2\} \quad (12)$$
$$= \sum_{i=1}^{s} w_i \exp\{-\beta(\mathbf{x}-\mathbf{c}_i)^T(\mathbf{x}-\mathbf{c}_i)\}$$

where the network contains a total number s of neurons with weights $w_1, \ldots, w_s$, β is a parameter, and $\mathbf{c}_i$ is a given point corresponding to the i-th neuron. The radial basis function itself is defined as

$$\varphi(\mathbf{x}, \mathbf{c}_i) = \exp\{-\beta||\mathbf{x}-\mathbf{c}_i||^2\}, \quad \mathbf{x} \in \mathbb{R}^p, \quad (13)$$

and the points $\mathbf{c}_i$ can be interpreted as centers, from which the Euclidean distances are computed.

The output (13) is a sum of weighted probability densities of the normal distribution. The training of the networks requires to determine the number of radial units and their centers and variances. The formula (13) does not contain a normalizing constant for the density of the multivariate normal distribution, but it is contained in the weights for individual neurons. This type of network is however less suitable for high-dimensional data.

5.4 Support vector regression

While support vector machines (SVM) are more popular in the classification context, both SVM and support vector regression (SVR) were proposed in Vapnik's seminal book [25]. SVR is firmly grounded in the framework of statistical learning theory [6], which is in fact a theory of estimating an unknown continuous function. This is an advantage compared to neural networks, which are often criticized for their suboptimality and weak theoretical foundations compensated by sophisticated heuristics.

SVR explicitly formalizes the concepts solved implicitly by neural networks and can be considered a close relative of neural networks and an alternative approach to their training. The simplest form of SVR is the linear SVR, which solves the task of linear regression, while a so-called kernel trick allows to extend the linear SVR to the task of nonparametric regression. There exist also different methods exploiting the same ideas of Vapnik, e.g. the Least Squares Support Vector Machine (LS-SVM) [22].

The optimal values of the parameters of SVR are found by optimizing the prediction accuracy by means of solving a quadratic programming problem. Thanks to a suitable regularization, the SVR is suitable also for high-dimensional applications. Another advantage is that the prediction rule of SVR is based only on a small set of support vectors.

Acknowledgement

The project was supported by the project 17-01251S "Metalearning for extracting rules with numerical consequents" of the Czech Science Foundation.

References

[1] Baerns M, Holeňa M. Combinatorial development of solid catalytic materials. Singapore: World Scientific; 2009.

[2] Barker M, Rayens W. Partial least squares for discrimination. Journal of Chemometrics 2003; 17: 166-173.

[3] Borangíu T, Purcarea V. The future of healthcare–Information based medicine. Journal of Medicine and Life 2008; 1: 233-237.

[4] Davies PL, Kovac A. Local extremes, runs, strings and multiresolution. Annals of Statistics 2001; 29: 1-65.

[5] Gnecco G, Kůrková V, Sanguineti M. Some comparisons of complexity in dictionary-based and linear computational models. Neural Networks 2011; 24: 171-182.

[6] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2001.

[7] Heritier S, Cantoni E, Copt S, Victoria-Feser MP. Robust methods in biostatistics. New York: Wiley; 2009.

[8] Kalina J. Highly robust statistical methods in medical image analysis. Biocybernetics and Biomedical Engineering 2012; 32 (2), 3-16.

[9] Kalina J. Implicitly weighted methods in robust image analysis. Journal of Mathematical Imaging and Vision 2012; 44: 449-462.

[10] Kalina J, Schlenker A. A robust supervised variable selection for noisy high-dimensional data. BioMed Research International 2015; Article 320385, 1-10.

[11] Kalina J, Zvárová J. Decision support for mental health: Towards the information-based psychiatry. International Journal of Computational Models and Algorithms in Medicine 2014; 4 (2), 53-65.

[12] Kalina J, Zvárová J. Perspectives of information-based methods in medicine: An outlook for mental health care. Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies—Volume 5: HEALTHINF, 2016, 365-370.

[13] Lindsey JK. Nonlinear models in medical statistics. Oxford: Oxford University Press; 2001.

[14] Loh WY. Fifty years of classification and regression trees. International Statistical Review 2014; 82 (3): 329-348.

[15] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002; 18 (1): 39-50.

[16] Rokach L, Maimon O. Data mining with decision trees: Theory and applications. World Scientific Publishing, Singapore.

[17] Rencher AC, Christensen WF. Methods of multivariate analysis. 3rd edn. Wiley, Hoboken.

[18] Rusiecki A. Robust MCD-based backpropagation learning algorithm. In Rutkowski L, Tadeusiewicz R., Zadeh L., Zurada J. (Eds.): Artificial Intelligence and Soft Computing. Lecture Notes in Computer Science 2008; 5097: 154-163.

[19] Saleh AKMdE, Picek J, Kalina J. R-estimation of the parameters of a multiple regression model with measurement errors. Metrika 2012; 75 (3): 311-328.

[20] Šebesta V, Tučková J. The extraction of markers for the training of neural networks dedicated for the speech prosody control. In Novel Applications of Neural Networks in Engineering EANN'05, pp. 245-250.

[21] Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: A case study. Statistica Neerlandica 2001; 55 (1): 76-88.

[22] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters 1999; 9: 293-300.

[23] Tan Y, Shi L, Tong W, Hwang GTG, Wang C. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. Computational Biology and Chemistry 2004; 28: 235-244.

[24] Vach W. Regression models as a tool in medical research. Boca Raton: CRC Press; 2013.

[25] Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.

[26] Víšek JÁ. Consistency of the least weighted squares under heteroscedasticity. Kybernetika 2011; 47 (2): 179-206.

[27] Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 2006; 101 (476): 1418-1429.

[28] Zvárová J, Veselý A, Vajda I. Data, information and knowledge. In Berka P, Rauch J, Zighed D, editors: Data mining and medical knowledge management: Cases and applications standards. Hershey: IGI Global; 2009; 1-36.

[29] Zvárová J, Mazura I et al. Methods of molecular biology and bioinformatics. Prague: Karolinum; 2012. (In Czech.)

Data Collection Methods for the Diagnosis of Parkinson's Disease

Michal Vadovský[1], **Ján Paralič**[1]

[1] Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia

Abstract

In determining the symptoms and predicting disease in medicine, health outcomes of patients are used, which are obtained from different clinical tests. Among the initial symptoms of people suffering from Parkinson's disease we can include muscle rigidity, problems with speech (dysphonia), movement or writing (dysgraphia). In this article, we focus just on the data obtained from the primary symptoms of patients and their further use for early diagnosis and detection of Parkinson's disease using machine learning methods. We first describe basic characteristics of this disease and its typical features and then we analyze the current state and methods of collecting data used for creating decision models. Next, we also summarize the results of our previous research in this area and describe the future directions of our work.

Keywords

Parkinson's disease, Machine learning, Speech, Writing

Correspondence to:

Michal Vadovský
Department of Cybernetics and Artificial Intelligence
Technical University of Košice, Slovakia
Address: Letná 9, 042 00 Košice
E–mail: michal.vadovsky@tuke.sk

IJBH 2017; 5(1):28–32
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

Parkinson's disease [1] is a very serious neurological disease for which there is still no remedy. It affects more than one million patients across Europe. According to estimates, in Slovakia, there are about 12 – 15 thousand people suffering from this disease. Initial symptoms can occur even before reaching 40 years of age, but the disease usually affects people about 60 years old. The main cause of the disease is the death of nerve cells, producing an important chemical substance called dopamine [2]. This substance is important in the transmission of nerve signals, thus management and coordination of movement and muscles activity. Due to the lack of dopamine patient loses the ability to control movement functions, eventually becoming unable to perform normal activities, such as walking and moving [3].

Disease onset and initial symptoms [4] are often nonspecific. Typical resting tremor, slowed down movement and muscle rigidity precedes sleeping disorder, constipation, shoulder pain and failure to smell. These symptoms are difficult to recognize. Since the progression of this disease is slow sometimes, it may take several years before the resting tremor leads to the diagnosis of Parkinson's disease.

Nowadays, there is no satisfactory method, which could completely cure patients with Parkinson disease. Medication substituting dopamine and relieving the disease progress is partially helpful. The most frequent medication currently is L-Dopa [5], which, however, may cause serious adverse effects, in particular psychiatric ones, when used on a long-term basis. Side effects of the medication can sometimes be improved by changing the dose, the form of the drug or using other types of drug. Medication in form of patches is also popular. Various neurosurgeries [6], which aim to selectively destroy or stimulate individual nerve centers, can also be considered as means to control the disease progress. Cell replacements as well as stem cell therapies are the most up to date approaches for more efficient treatment of Parkinson's disease. These, nevertheless, are currently subject of further research.

The cause of Parkinson's disease in a patient is not yet known. There is significant amount of research aimed at evaluating genetics, toxins or ageing period. Currently, there is no objective quantitative method useful for clin-

ical diagnosis of the disease. It is assumed, that the disease can be definitely diagnosed after death, indicating the complexity of the diagnosis [7]. This is why the amount of research is focused on the creation of expert systems or decision support systems and assessing diagnosis of Parkinson's disease. These systems work with data which enable to build models with high accuracy. This article focuses on current methods of data collection and analyses useful for the development of successful models that can detect patients' initial symptoms.

2 Current state

Recently, many studies focus on the classification of patients with Parkinson's disease (PD). It is largely due to the fact that there is currently no suitable medicine, as well as its early diagnosis is difficult for doctors. As already mentioned above, the initial symptoms of PD would include muscle stiffness, problems with speech (dysphonia) [8] and movement, for example when writing (dysgraphia) [9]. This is the reason why many researchers focus on data collection capturing writing and speech of patients.

For example, in the work [10] P. Drotár et al. focused on the classification of patients with PD using data obtained from the writing. These data were collected at the First Department of Neurology, St. Anne's University Hospital in Brno, Czech Republic. Together they have captured data of 75 people, where 37 of them were suffering from Parkinson's disease and 38 were healthy. Each subject was writing with the right hand, attended at least 10 years of school and spoke fluent Czech. Sentence they were asked to write using digital tablet Intuos 4M (Wacom) was as follows: "Tramvaj dnes už nepojede." (The tram won't go today). The process of writing this sentence was processed into a number of indicators such as: stroke speed (trajectory during stroke divided by stroke duration), speed (trajectory during handwriting divided by handwriting duration), velocity (rate at which the position of a pen changes with time), acceleration (rate at which the velocity of a pen changes with time), jerk (rate at which the acceleration of a pen changes with time) and many other derived attributes. The authors recorded not only the movement of patients writing on the tablet, but also outside it (in the air). The goal of this publication was to compare the success rate of models from data obtained by writing on the tablet, the movement outside the surface of the tablet (in the air) or a combination of both. For categorization of the patients into healthy and suffering from PD they have used machine learning method called Support Vector Machines (SVM). The highest accuracy was achieved by model trained on the combination of both, data from writing and data form movement over the tablet (85.61%), followed by the model trained only on the data from the movement over the tablet in the air (84.43%), and finally the model trained only on the data about movement of the writing on the surface of the tablet (78.16%). Finally, it was found that the movement of the writing in the air has a major impact on classification accuracy of patients and confirms that the handwriting can be used as a biomarker for the diagnosis of Patient's disease.

Another publication [11], written by P. Drotár et al. also describes the classification of the same patients as above, according to the data obtained from the writing, but more records have been available. They focused on letters and simple words, where it is not necessary to interrupt the touch of the pen with the tablet. Each subject first drew Archimedean spiral, further wrote a letter "L", the syllable "le", the words "les" (forest), "lektorka" (lecturer), "porovnat" (to compare) and "nepopadnout" (not to catch). The last task for the subjects was to write the same sentence as above. The authors used three classification methods: K-nearest neighbors (KNN), AdaBoost and Support Vector Machines (SVM). The highest accuracy was achieved using SVM method again (81.3%), followed by AdaBoost classifier (78.9%) and finally KNN (71.7%). In addition, they divided individual attributes into kinematics and pressure features of handwriting and compared them using SVM method. The best results were achieved with models developed by attributes of pressure features (82.5%) while using the attributes of the kinematics features only achieved 75.4% accuracy. By combining the attributes of the two handwriting feature types they arrived at the success rate of model at 81.3% .

There are even more publications available from authors who worked with the data obtained from the records of patients' speech. Such kind of data is in fact freely available on the UCI Machine Learning Repository, where there are currently 3 similar datasets and include basic attributes derived from so called jitter and shimmer groups of parameters. Jitter represents voice frequency and shimmer voice amplitude perturbations and both are commonly used as part of a comprehensive voice examination [12]. For example, in the work [13] G. Yadav focused on the speech of PD patients. Their speech signals have been transformed into various numerical indicators, including the groups of jitter and shimmer attributes. For comparison resulted classification models of authors in [13] were as follows: decision trees (75%), SVM (76%) and logistic regression (64%).

3 Our research results

We have worked up to now only with data that is freely available on the Internet, specifically on UCI Machine Learning Repository. In this database there are currently available 3 different types of data, aimed to classify patients with PD using their processed speech records (it a group of numerical attributes sketched above). We have so far only worked with two of these datasets, where we firstly monitored and compared the classification accuracy obtained using different models of machine learning methods. In the second case, we focused on the particular

types of patients' speech comparing classification accuracy of created machine learning models.

3.1 The research focused on the overall accuracy

In the publication [14] we worked with data created by Max Little of the University of Oxford in collaboration with the National Center for Voice and Speech located in the Denver, Colorado [15]. The entire dataset contained the records of 32 patients (out of which 24 were with PD). For most individuals 6 records of their normal speech were available, and 3 individuals provided 7 records, which together resulted in 165 records. Several records of one patient in the data are taken independently of each other, while possible mutual correlation of such records will be the subject of our future research. Patients' speech was transformed into the number of attributes, such as the average (MDVP:Fo(Hz)) and minimum (MDVP:Fhi(Hz)) vocal frequency, level of variability in the frequency (Jitter group of attributes), the rate of variation in amplitude (Shimmer group of attributes), measurements of the noise and tonal components in the voice (NHR, HNR) and many others. Target attribute (Status) was in binary form 1/0 and represented the information, whether particular patient is suffering from the PD or not.

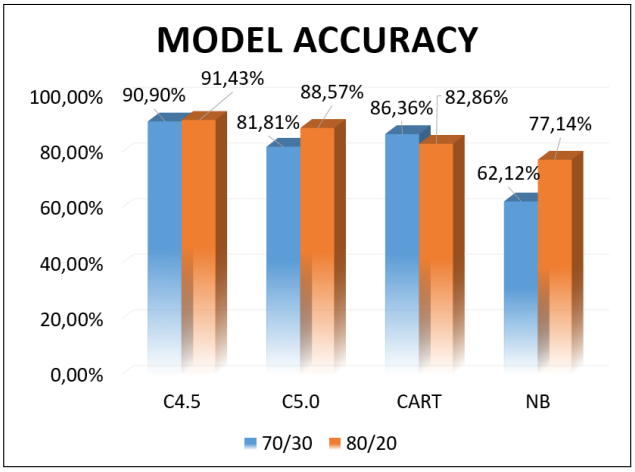

Figure 1: The results of the models created using machine learning methods.

In the first part, we monitored the dependence of individual attributes to the target attribute by using Two-sample Welch t-test that compares the average values of each attribute with respect to the class of the target attribute. If the average values of the two sets are the same, we are talking about independence of the attributes (zero hypothesis H_0). If there is a difference between the average values of the two sets, then there is a relationship between attributes (alternative hypothesis HA). In this test, the p-value is observed in order to decide how significant the relationship between particular attributes is. The lower the p-value, the higher is the probability of the target attribute dependence to the given numeric attribute. The lowest p-value (0.028) and therefore the highest dependency was observed between the attribute Status and the attribute expressing the maximum vocal frequency. In this case with the certainty of $(1 - p) * 100$ percent, we can reject H_0 and confirm HA – in this scenario, with the max. 97.2% certainty, we reject H_0.

The main part was aimed to create classification models using Naïve Bayesian classifier method and the methods of decision trees – Algorithms C4.5, C5.0, and CART. Before the modeling, we divided the data into training and testing set in the ratio 70/30 and 80/20. The data in the following sets were divided randomly according to the ratio, while the proportion of healthy patients and patients with PD was maintained in the two sets (approximately 1:3). Then we created 10 models using the same training data for each of the four different methods (C4.5, C5.0, CART and Naïve Bayesian classifier) and calculated their accuracies on testing set. The model with the highest accuracy for each method was chosen. The obtained results are shown in Figure 1, where we can observe that the highest accuracy of 91.43% has been obtained using the C4.5 algorithm with data distribution 80/20. The second highest accuracy at the level of 90.90% was achieved with data distribution 70/30. On the contrary, using the method of the Naïve Bayesian classifier we obtained the lowest accuracy of the model in both splits of the data into training and testing sets:

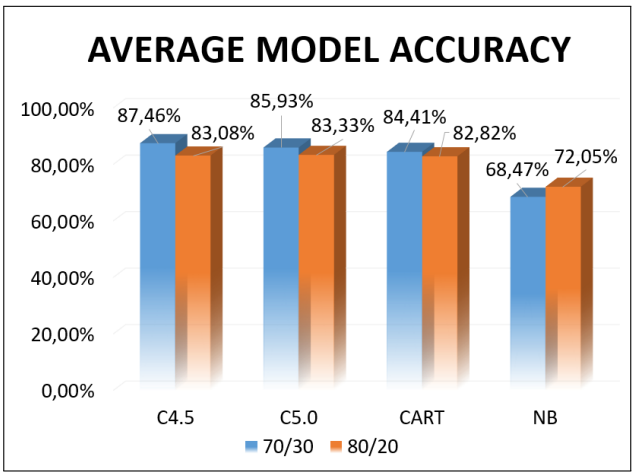

Figure 2: The average results of the models created using machine learning methods.

In order to reduce the effect of chance when random selecting the training and testing dataset, we repeated this process 10 times and the resulting accuracy of particular method was calculated as the average of all obtained accuracy results from 10 iterations of our training and testing process. The results are presented in Figure 2, where one can notice that the highest average accuracy of 87.46% was achieved using C4.5 algorithm again, however the for the training/testing data ratio of 70/30. The overall approach of decision trees and its algorithms C4.5, C5.0 and CART resulted in higher average accuracy in this data dis-

tribution. The lowest accuracy was achieved again using Naïve Bayesian classifier method.

For the best method of decision tree algorithm C4.5, the results were in the form of contingency table are presented in Table 1, where we can better understand the kind of errors of this model and follow the number of false positive and false negative cases. Since several contingency tables were obtained using the algorithm C4.5 we averaged the adjusted values in the table.

Table 1: Contingency table.

C4.5		**The real value**	
		0	1
Predicted value by the model	0	11	3
	1	4	41

This contingency table compares the real values in the testing set with predicted values obtained using C4.5 classification models. Values 0 and 1 represent the target attribute in binary form and served as the information on whether the patient suffered from Parkinson's disease (1) or not (0). For the records where patients do not suffer from PD, models were able to predict an average of 11 correct classifications to the target class and 4 records were incorrectly classified (false positives). Within the records of patients suffering from PD, models were able to classify the correct class in average of 41 cases and for 3 cases there was an error (false negatives).

According to the results, we can say that we can successfully classify patients by transformed indicators of their speech. In the publication [13] using same data and SWM method, the highest achieved accuracy was only 76

3.2 The research focused on a type of speech

In another publication [16] aiming for the classification of patients with Parkinson's disease we used freely available data from the Department of Neurology in Cerraphasa Faculty of Medicine, Istanbul University [17]. These data include basic attributes of Jitter and Shimmer, but in addition, other attributes are involved as well, such as NTH and HTN (two measures of the ration of the total noise component in the voice), Media pitch, Mean pitch, Standard deviation, Number of pulses etc. Sound recording was obtained from 40 subjects, half of whom suffered from PD. Each subject was represented by 26 recordings, while pronouncing the letters U, A, O, numbers from 1 to 10, four short sentences and nine words in the Turkish language. In our research, we focused to detect which of these types of speech signals can lead to a classification model with the highest accuracy. Target attribute was also in binary form 1/0, which divides subjects into patients with PD and healthy patients.

Firstly, we focused on the data understanding and the dependency monitoring, same as in the publication [15] using two-sample Welch t-test. The highest dependence to the target attribute was found in the following attributes: Jitter (local, absolute) (p-value = 0.0000000609), Shimmer (apq11) (p-value = 0.00000216) a Max pitch (0.00000644). In addition, we also observed the cut-off value of attributes that can classify patients with the highest possible accuracy. These values were calculated using Youden index [18]. By attribute called Jitter (PPQ5) we reached the cut-off point at the level of 1.0065, according to which we can divide subjects with 59.13

Our main aim was to ascertain what data in terms of the type of speech we should use to achieve the models with the highest accuracy. For creating models, we used the method of decision trees and corresponding algorithms C4.5, C5.0, and CART, as well as RandomForest. We used 4 and 5-fold cross validation for measuring of expected classification accuracy. From the results we found that the most accurate models (71%) were obtained employing the 4-fold cross validation, using RandomForest algorithm, where the subjects pronounced numbers from 1 to 10. When averaging the success rates of all the algorithms, we calculated the highest accuracy (66.5%) using 4-fold cross validation again, with the data of patients pronouncing numbers. On the contrary, the lowest average of all accuracies (51.86% and 50.63%) for both cross validations were obtained with data of individuals pronouncing the letter U.

We can conclude the second case with the observation that in all cases we can most reliably classify the PD patients according to the data, which record their pronunciation of numbers. In the future, this observation could help in data collection process for classification models. In addition, recordings of patients pronouncing the letter U can eventually cause the decrease in accuracy of the classification model.

4 Conclusion

In this article, we focused on various means of collecting and analyzing patients' data in order to create the models with the highest possible precision in the classification of PD. These methods were based on primary symptoms of people suffering from Parkinson's disease (muscle stiffness, problems with speech and writing). According to the results achieved in this article it can be observed that the classification of patients according to their writing and speech is possible with an average success rate at the level of 87.46%. This average percentage was obtained in pronouncing simple words of individuals where we used all available records to train the models. The data split by the type of speech achieved less precise results (71%). From this perspective, it makes more sense to use different types of speech of individual subjects. Similarly, writing records of subjects achieved a success rate of 85%, where a combination of attributes was used - screening movement of the pen on the tablet but also the movement of the hand over the tablet in the air.

In conclusion, the two types of data can be considered to have significant potential in the future and can help in the early and non-invasive diagnosis of PD patients.

Also, speech or writing of subject is not so costly and time-consuming compared to performing large amounts of medical tests. However, due to the large predominance of healthy individuals in the total population, the medical tests cannot be completely ruled out due to false alarms. Obtaining transformed attributes of patients' speech is quite simple. There are several tools available for this purpose, e.g. Praat Acoustic Analysis [19].

5 Future work

In the future research we would like to focus also on the classification of the different stages of the disease, not only to classify whether the patient is suffering from Parkinson's disease or not. We assume that determination of the initial stage will be the most difficult part, since the subjects are unlikely to be very different from healthy individuals. In addition, further focus will also be placed on the correlation rate of the data obtained for single individual patient and its impact on the final classification accuracy. More focus in the future will be involved in working with the data we have obtained recently from the company mPower: Mobile Parkinson Disease Study [20]. This company collects data from people using their mobile application, which records their demographic information as well as details of a voice, walking, memory and tapping on the mobile screen. With these data, we can expand our research in several areas and symptoms of Parkinson's disease. In addition, we have also agreed on cooperation with P. Drotar, who works with the data of subjects' movement while writing on a tablet.

Acknowledgement

The work presented in this paper was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0493/16 and by the Slovak Cultural and Educational Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic under grant No. 025TUKE-4/2015.

References

[1] L.M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," The Lancet Neurology, vol. 5, no. 6 (2006), pp. 525 – 535.

[2] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, L. Defebvre, F. Grenez, "Low-frequency vocal modulations in vowels produced by Parkinsonian subjects," Speech Communication, vol. 50, no. 4 (2008), pp. 288-300.

[3] H.L. Teulings, J.L. Contreras-Vidal, G.E. Stelmach, C.H. Adler, "Parkinsonism reduces coordination of fingers, wrist, andarm in fine motor control," Experimental Neurology, vol. 146, no. 1 (1997), pp. 159–170.

[4] J.L. Contreras-Vidal, G.E. Stelmach, "Effects of parkinsonism on motor control," Life Sciences, vol. 58, no. 3 (1995), pp. 165-176.

[5] V. Ruonala, M.P. Tarvainen, P.A. Karjalainen, E. Pekkonen, S.M. Rissanen, "Autonomic nervous system response to L-dopa in patients with advanced Parkinson's disease," Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, 2015, pp. 6162-6165, ISBN 978-1-4244-9271-8.

[6] National Collaborating Centre for Chronic Conditions, Parkinson's disease, London, U.K.: Royal College of Physicians, 2006.

[7] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. Smekal, M. Faundez-Zanuy, "A new modality for quantitative evaluation of Parkinson's disease: In-air movement," Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, 2013, pp. 1-4, ISBN: 978-1-4799-3163-7.

[8] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," Biomedical Engineering, IEEE Transactions on, vol. 59, no. 5, pp. 1264-1271,2012.

[9] O. Geman and H. Costin, "Parkinsons disease prediction based on multi state markov models," International Journal of Computers, Communications & Control, vol. 8, no. 4 (2013), pp. 525-537.

[10] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. Smekal, M. Faundez-Zanuy, "Analysis of in-air movement in handwriting: A novel marker for Parkinson's disease," Computer Methods and Programs in Biomedicine, vol. 117, no. 3 (2014), pp. 405-411.

[11] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. Smekal, M. Faundez-Zanuy, "Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease," Artificial Intelligence in Medicine, vol. 67 (2016), pp. 39-46.

[12] J. P. Teixeiraa, A. Gonçalves: Accuracy of Jitter and Shimmer Measurements. In: CENTERIS 2014 / ProjMAN 2014 / HCIST 2014 - Int. Conf. on Health and Social Care Information Systems and Technologies. Elsevier Ltd, Procedia Technology 16 (2014) p. 1190 – 1199.

[13] A. Tsanas et al., "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," IEEE Transactions on Biomedical Engineering, vol. 57, no. 4 (2010), pp. 884-893.

[14] M. Vadovský, J. Paralič, "Predikcia Parkinsonovej choroby pomocou signálov reči použitím metód dolovania v dátach", In: WIKT & DaZ 2016: 11th Workshop on Intelligent and Knowledge Oriented Technologies 35th Conference on Data and Knowledge, Smolenice, Slovakia – Bratislava: STU, 2016, pp. 329-333. ISBN: 978-80-227-4619-9.

[15] UCI Macxhine Learning repository: Center for Machine Learning and Intelligent Systems – Parkinsons Data Set. Available at: `https://archive.ics.uci.edu/ml/datasets/Parkinsons`.

[16] M. Vadovský, J. Paralič, "Parkinson's Disease patients' classification based on the speech signals", unpublished.

[17] UCI Machine Learning repository: Center for Machine Learning and Intelligent Systems – Parkinson Speech Dataset with Multiple Types of Sound Recording Data Set. Available at: `https://archive.ics.uci.edu/ml/datasets/Parkinson+Speeh+Dataset+with++Multiple+Types+of+Sound+Recordings`.

[18] M.D. Ruopp, N.J. Perkins, B.W. Whitcomb, "Youden Index and Optimal Cut-point Estimated from Observations Affected by a Lower Limit of Detection," Biometrical Journal, vol. 55, no. 3 (2008), pp. 419-430.

[19] Praat: doing phonetics by computer. Available at: `http://www.fon.hum.uva.nl/praat/`.

[20] mPower: Mobile Parkinson Disease Study. Available at: `https://parkinsonmpower.org`.

Clinical Decision Support System in Dentistry

Michaela Bučková[1], **Tatjana Dostálová**[1], **Alexandra Polášková**[1], **Magdaléna Kašparová**[1], **Milan Drahoš**[2]

[1] Charles University 2nd Medical Faculty, Prague, Czech Republic

[2] Charles University 1st Medical Faculty, Prague, Czech Republic

Abstract

Dental treatment of special needs patients is more expensive and time-consuming than conventional dental treatment. Extensive research supported by the Ministry of Health (IGA : 9991-4) that focuses on the use of various therapeutic methods in the treatment of special needs patients takes place at the Department of stomatology Teaching Hospital Motol and 2nd Medical Faculty of Charles University.

A clinical decision support system for the treatment of the children with special needs was created. The system should be used for faster orientation and create a formula to treat patients who have a handicap due to their non-standard mode of therapy.

Keywords

Decision making system, Decision support system, Dentistry, Quality assesment

Correspondence to:

Michaela Bučková
Charles University 2nd Medical Faculty, Prague, Czech Republic
Address: V úvalu 84, 15006 Prague 5
E–mail: Stank@email.cz

IJBH 2017; 5(1):33–35
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

Making the right decision is becoming the key factor for successful diagnosis in all areas of medicine. Decision making is usually based on a combination of experiences from solving similar cases, the results of recent research and personal judgement. Decision support systems helping physicians are becoming a very important part of medical decision making, particularly in those situations where a decision must be made effectively and reliably [1]. The objective of this study is to create a tool for clinical dental providers, which helps them to make therapy decisions for children with special needs and multiple caries lesions. Nowadays there does not exist any system of information exchange between private dental offices and hospitals. Our simple decision support system should show the gold standard in the hospital to help the private dentist to inform their patients about the possibilities of the treatment and offer the possibility of the treatment with a new method of the preparation – the Er:YAG laser.

2 Methods

Finding a solution with the help of decision trees was started by preparing two groups of patients. From the daily experience in our department we see that young patients are referred to a special dental office with the request of treatment from their dentists because of uncooperation. So there were created two groups of patients - cooperative and uncooperative. The cooperative patients are treated under local anesthesia. Cooperative patients are usually treated by the private clinical dental providers and rarely come to the hospital for treatment.

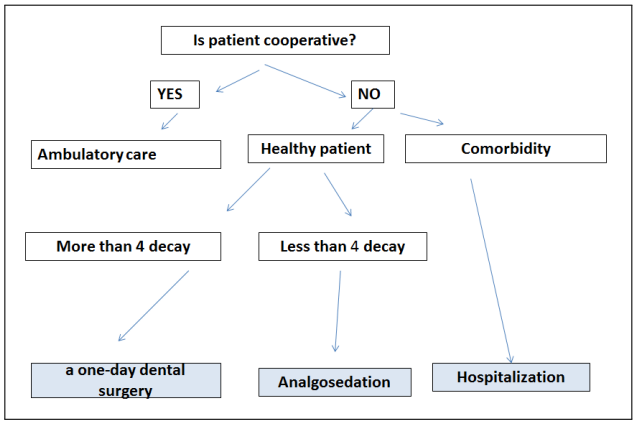

Figure 1: Simple decision scheme.

Our clinic can offer dental treatment with an Er:YAG laser, which is a new treatment method used instead of the common drill preparation. The Er:YAG laser has a lot of advantages and helps to prevent noncooperation of chil-

dren with multiple caries lesions. The patients are treated under general anesthesia or analgosedation. The possibilities and conditions are included in the support system. The scheme is based on the daily experience, expert knowledge and daily work at our clinic. Our standards are not known among clinical dental providers. Figure 1 shows the main idea of the scheme.

The proposed scheme to support decision-making during dental treatment in paediatric patients is conceived as a planning decision-making process without quantifying uncertainty. The input data are known, together with the objectives of the decision-making process (optimal patient treatment technique) and the core decision-making process is created by using rules, conditions to achieve optimal goals [2]. This is a simple inference network. The presentation schema is created in HTML with cascading style sheets, and the JavaScript language with a library jQuery. Pictures and explanatory pop-up windows are shown. The scheme is available at `http://decisionschema.cuni.cz/`. A test of the clinical decision support system was made in the year 2015 with 100 new patients. 50 uncooperative patients and 50 cooperative patients who came to the hospital because of the need of dental treatment but they could not find a private dentists who treats children in their town.

3 Results

For the assessment of the function of the system we chose 50 cooperative patients and 50 uncooperative patients. The dentist, who verified the decision support system, works in the hospital, so is familiar with the gold standards at the clinic. The author of the system was also present meanwhile the diagnose was made. To display and evaluate the agreement between dentist's decision and the decision schema output, pivot tables were used. Rows correspond to scheme decisions and columns correspond to the dentist's decisions. The coefficient of conformity c was defined as the ratio of sum of the diagonal elements and the sum of all elements in the pivot table

$$c = \frac{\sum_{i=1}^{n} x_{i,i}}{\sum_{i,j=1}^{n} x_{i,j}}$$

where $x_{i,j}$ represents an element of the table with size n. In the case, where the conformity between dentist and schema is random, the coefficient for table size n would score

$$c_{\text{random}} = \frac{n \cdot a}{n^2 \cdot a} = \frac{1}{n}$$

where a represents the average score for a single cell. In the opposite case, where the conformity between dentist and schema is ideal, the coefficient reaches

$$c_{\text{ideal}} = 1$$

The data from the non-cooperative group were registered in Table 1 and for the cooperative group in Table 2. The appropriate coefficients of conformity are

$$c_{\text{noncoop}} = 0.90$$

resp.

$$c_{\text{coop}} = 0.86$$

Table 1: Uncooperative patient - coefficient of conformity $c_{\text{noncoop}} = 0.90$.

	Dentist's decision			
Scheme result	A	AGA	HGA	NS
Analgosedation (A)	18	2	0	0
Ambulant General anaesthesia (AGA)	0	16	2	0
General anaesthesia within hospitalization (HGA)	0	1	15	0
The treatment is not suitable (NS)	0	0	0	1

Table 2: Cooperative patient - coefficient of conformity $c_{\text{coop}} = 0.86$.

Cooperative patient	Dentists decision	
Scheme result	Laser	Drill
Laser	45	6
Drill	3	5

4 Discussion and Conclusion

Within our study we have created, a decision support system for the treatment of patients with multiple caries lesions. When creating a knowledge base schema we started from a diagnostic and treatment protocol used in the Department for stomatology Teaching Hospital Motol and 2nd Medical Faculty of Charles University. It is important to decide which type of anaesthesia is suitable in case of anxious and uncooperative patients. In the group of cooperative patients it is possible to use the laser treatment instead of the conventional drilling protocol. The use of the Er:YAG laser in dental practice is not common but is advantageous and the patients, who are in danger of noncooperation in the future, can be referred to the Dental clinic to be treated. The user interface was created using Web technology. We have created a group of 100 patients who came for conservative dental treatment. The decision support system was tested by the dentist and the author. We used pivot tables to compare the results. We have defined the coefficient of conformity. Our scheme shows the coefficient for uncooperative patients, 0.90 and 0.86 for cooperating children. This result is considered to be good. In one case the dental treatment was not suitable

because of the bad health status of the patient. The treatment was postponed until after solving the other health problems. Some patients were treated under analgosedation but because of the paradoxreaction they had to be treated under the general anesthesia. The frequency of classes influences the results. In the case of uncooperative patients the results are balanced. In the case of cooperative patients we can see that the frequency of classes influences the results. The diagnosis of the degree of the decay is not simple and in some cases the dentist after the X-Ray examination changed the decision. In this part of the schema our simple approach need to be improved. The knowledge base may be developed and enriched by various less frequent problems occurring in paediatric patients and the scheme shows, what are the possibilities and procedures in the Dental clinic. 3984 uncooperative patients were treated in our clinic in the years 2006-2014. All these patients were informed by their dentists. We wanted to create the tool, which helps to standardize this information exchange. Future work building from this baseline assessment will measure the actual provider adherence to the tools and factors relating to overall implementation adherence [3].

Acknowledgements

Supported by "Conceptual development project of research organization 00064203" University Hospital Motol, Prague and IRP "E-learning and distance components of education 237984".

Conflicts of interest

All authors have no financial and personal relationships with other people or organizations that could inappropriately influence (bias) their actions.

References

[1] Mertz E, Bolarinwa O, et al. Provider Attitudes Toward the Implementation of Clinical Decision Support Tools in Dental Practice. Original Research Article. Journal of Evidence Based Dental Practice, 2015; 15 (4):152-163.

[2] Nguyen L, Bellucci E, Nguyen LT.Electronic health records implementation: an evaluation of information system impact and contingency factors. Original Research Article. Int J Med Inform. 2014 ; 83(11):779-96.

[3] Fontaine P, Ross S, Zink T, Schilling L: Systematic review of health information exchange in primary care practices. Original Research Article. J Am Board Fam Med 2010 ; 23(5):655-670.

New Derivation of Balding-Nichols Formula

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

[1] Institute of Health Information and Statistics of the Czech Republic

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

Abstract

In this paper we describe the influence of the subpopulation structure on the probability of observing homozygous and heterozygous genotype. Balding-Nichols formula using for this purpose coefficient θ is compared with the newly derived formula that reduces the bias of the calculated probabilities from probabilities valid in an unstructured population, which was described in several studies.

Keywords

Coancestry coefficient, Homozygote and heterozygote genotype

Correspondence to:

Dalibor Slovák
Institute of Health Information and Statistics
Address: Palackého náměstí 4, 128 01 Prague 2
E–mail: dalibor.slovak@uzis.cz

IJBH 2017; 5(1):36–37
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Coancestry coefficient

Imagine a population divided into several subpopulations. Allelic frequencies in individual subpopulations are usually different from allelic frequencies in the whole population and persons within the same subpopulation have a more similar gene pool than people from different subpopulations.

Measure of subpopulation kinship is expressed by coancestry coefficient θ, which indicates the probability that two alleles randomly chosen from the subpopulation will be *ibd* (identical by descent). Alleles are identified as *ibd* when they are copies of any allele in a common ancestor (if located in two different individuals), or when they are copies of an allele in a common ancestor of parents (if there is one person) [1].

The definition implies that the value of θ varies between different subpopulations of the same population, but also between the different loci in the same subpopulation. It is therefore a property of a particular subpopulation and a specific locus. For the correct interpretation, it is necessary to realize that this is a parameter relating to a pair of alleles: θ therefore can be interpreted as the probability of occurrence of an *ibd* pair, which corresponds to the relative frequency of *ibd* pairs in the subpopulation.

At loci which are used in forensic identification, population frequency of individual alleles are currently known for most populations. Then it is also easy to predict the occurrence of homozygous and heterozygous genotype.

However, if population is structured, it is not only possible to narrow the view of a particular subpopulation and use the same procedure as for the whole population. The frequencies of alleles in a subpopulation are generally not known and the only known entry characterizing a subpopulation is just coancestry coefficient θ. If we want to estimate the probability of occurrence of homozygous and heterozygous genotype in the subpopulation and we do not know the relevant allelic frequencies, it is necessary to start with population frequencies and use θ as a correction.

2 Formula for calculation

The process how to include the influence of a subpopulation to the calculation of probabilities of observing a homozygous and heterozygous genotype in the subpopulation, was suggested in their article by Balding and Nichols [2] (so-called Balding-Nichols formula).

Rohlfs et al. [3] provide an overview of articles that deal with comparing of theoretical results predicted by Balding-Nichols formula and truly observed frequencies. Although the inclusion of population structure makes the test conservative, some results suggest that this conservativeness might be too high. The observed values often lie somewhat closer to allelic frequencies expected when the influence of the subpopulation is ignored than the correction calculated using the Balding-Nichols formula would suggest. Rohlfs et al. [3] attempt to compensate this

difference by calculating the fraction of shared alleles between different subpopulations to a lesser extent.

In our view, however, such an adjustment should not enter into the calculation and the observed difference is likely due to improper construction of Balding-Nichols formula.

Table 1: The change of genotype frequencies in various models (for genotype AB is not taken into acount the order of alleles). The values pA = 0.6, pB = 0.4, and $\theta = 0.05$ are used.

Genotype	**AA**	**AB**	**BB**
unstructured population	36 %	48 %	16 %
Balding's and Nichols' formula	37.2 %	45.6 %	17.2 %
proposed formula	35.8 %	47.91 %	16.29 %

On the basis of mathematical derivation, we have suggested a formula that adjusts calculation of probabilities of observation of homozygous and heterozygous genotype in the subpopulation. Comparison of the values of the allelic frequencies 0.6 and 0.4 with the value of the coancestry coefficient 0.05 is shown in Table 1. We can see that the values obtained using our proposed formula reduces bias caused by Balding-Nichols formula.

References

[1] Evett I.W., Weir B.S., Interpreting DNA evidence, Sinauer, 1998.

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

[3] Rohlfs R.V., Aguiar V.R.C., Lohmueller K.E., Castro A.M., Ferreira A.C.S., Almeida V.C.O., Louro I.D., Nielsen R., Fitting the Balding-Nichols model to forensic databases, Forensic Science International: Genetics 11 (2014), 56-63.

Paternity and/or Maternity Assessment of Complete and Partial Hydatidiform Moles and Non-molar Triploids

Halina Šimková[1,3], **Jiří Drábek**[2,3]

[1] Charles University, Faculty of Science, dpt. of Anthropology and Human Genetics, the Czech Republic

[2] Institute of Molecular and Translational Medicine of the Faculty of Medicine and Dentistry of Palacký University in Olomouc, the Czech Republic

[3] Czechoslovak Society for Forensic Genetics, Olomouc, the Czech Republic

Correspondence to:

Halina Šimková
Czechoslovak Society for Forensic Genetics
Address: Karafiátová 27, Olomouc, Czech Republic
E–mail: halina.simkova@gmail.com

IJBH 2017; 5(1):38
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

Abstract

Complete hydatidiform moles (CHM) and partial hydatidiform moles (PHM), as well as non-molar triploids (NMT) are very rare pathological products of conception (POCs). While normal POC originates from one haploid egg and one haploid sperm and consequently consists of biparental diploid cells, i.e. cells containing one set of paternal chromosomes and one set of maternal chromosomes, the contribution of maternal and/or paternal genetic information differs in CHMs, PHMs and NMT.

Complete hydatidiform mole (CHM) consists of uniparental diandric nulligynic diploid cells and usually originates in fertilization of empty ovum by one normal sperm following endoreduplication or, fertilization of empty ovum by two normal sperms. Partial hydatidiform mole (PHM) consists of biparental diandric monogynic triploid cells and usually originates in fertilization of normal ovum by two normal sperms. Non-molar triploid (NMT) consists of biparental monoandric digynic triploid cells and usually originates in fertilization of diploid ovum by normal sperm.

Despite their incorrect ploidy, these aberrant ova often can divide and give rise to more or less pathologically organized tissues that may potentially become the object of forensic analysis. In those cases correct formulas for calculating probabilistic kinship parameters (paternity index, maternity index, etc.) must be drawn up and applied. Also we point out several other interesting particular issues concerning aberrant POCs, such as assessment of mono/dispermity of unipaternal diploids or distinction of digynic monoandric triploids from placental mixtures of normal POCs.

Using Mendelian Randomization Principle to Demonstrate Protective Effects of the Isothiocyanate in Cruciferous Plants in the Prevention of Malignant Neoplasms

Vladimír Bencko[1], **Ladislav Novotný**[1]

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

Correspondence to:

Vladimír Bencko
Institute of Hygiene and Epidemiology 1st Faculty of Medicine, Charles University, Prague, Czech Republic
Address: Studničkova 7, Praha 2, 128 00
E–mail: vladimir.bencko@lf1.cuni.cz

IJBH 2017; 5(1):39
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

Abstract

Though the methodology and designs of epidemiological studies and analyses of medical databases have improved, associations between modifiable exposures and the disease in observational epidemiological studies remain partly biased. Mendelian randomization principle, which is the random distribution of parental genes to offspring in meiosis during gametogenesis and at conception, represents a new method of evaluation of the causal relations between the external causes and the disease. The use of this principle assumes the association between the disease and the genetic polymorphism, and reflects the biological relation between the suspected exposure and the disease, and is generally less prone to the phenomenon of confounding and reverse causation that can impair the interpretation of results in conventional observational studies.

Authors describe explanatory options of the Mendelian randomization principle by using an example of isothiocyanate versus lung carcinoma. Though the use of Mendelian randomization principle has its limitations, it offers new possibilities to test causal relations and clearly shows that means invested into the Human genome project can contribute to the understanding and prevention of adverse effects of modifiable exposure to the human health.

Keywords

Genetic epidemiology, Risk factors, Causality, Glutathione-S-transferase, Brassica genus, Isothiocyanate, Lung carcinoma

Defective Collagen Type I production in Czech Osteogenesis Imperfecta Patients

Lucie Hrušková[1], **Ivan Mazura**[1]

[1] Department of Pediatrics and Adolescent Medicine, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic

Abstract

Osteogenesis imperfecta (OI) is heritable and clinical heterogeneous disease of connective tissue. Currently, OI classification includes fourteen OI types differed by clinical signs and genetic origin. The typical clinical feature is low bone mass resulting in high frequency of bone fractures. Other feature observed in OI patients are bone deformities, blue or grey sclerae, otosclerosis and dentinogenesis imperfecta (DI). Molecular-genetic testing of 34 Czech OI probands identified nine mutations, including 6 novel ones, of collagen type I genes.

Keywords

Collagen type I, COL1A1, COL1A2, Osteogenesis imperfecta

Correspondence to:

Lucie Hrušková
First Faculty of Medicine, Charles University
Address: Kateřinská 32, 128 08 Prague 2
E–mail: black.luca@seznam.cz

IJBH 2017; 5(1):40–41
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

Genetic origin of first four OI types (I-IV) is defect in one of collagen type I genes (COL1A1, COL1A2). About 90% of affected patients have causative mutation in one of these two genes. These mutations result in mild to lethal phenotype regard on type and position of the change in alpha chains encoded by collagen type I genes [1, 2].

Collagen type I is a heterotrimer composed of two alpha 1 chains (produced by COL1A1 gene) and one alpha 2 chain (encoded by COL1A2). Mutations of alpha chains result in either reduced production of collagen type I (typical genetic origin of patients affected by OI type 1) or in the synthesis of structurally abnormal protein [1, 3].

2 Material and Methods

Molecular genetic analysis of collagen I genes was performed in a cohort of 34 OI patients. The DNA samples were analysed by PCR and Sanger sequencing. Identified DNA changes were compared with wild-type sequences as submitted to Ensembl accession no. ENST00000225964 (COL1A1 gene) and no. ENST00000297268 (COL1A2 gene) and with Osteogenesis Imperfecta Variant Database, the Human Genome Mutation Database and the Ensembl database.

Table 1: Identified collagen type I DNA changes.

Gene	DNA change	Novelty
COL1A1	p.Tyr47X	yes
COL1A1	p.Arg131X	no
COL1A1	p.Arg415X	yes
COL1A1	p.Gln1341X	yes
COL1A1	p.Cys61Phe	no
COL1A1	p.Gly794Gly	yes
COL1A1	p.Pro1186Ala	no
COL1A1	c.1057-1G>T	yes
COL1A2	p.Gly814Trp	yes

X – nonsense mutation (STOP codon) resulting in reduced production of collagen type I; Tyr – tyrosine; Arg – arginine; Gln – glutamine; Cys – cysteine; Phe – phenylalanine; Gly – glycine; Pro – proline; Ala – alanine; Trp – tryptophan

This study was performed in accordance with principles of the Declaration of Helsinki and approved by the Ethics Committee of General University Hospital in Prague (project 83/14). Participants provided a written informed consent for their involvement in the study.

3 Results

Molecular genetic analysis identified 9 mutations. 8 of them occurred in COL1A1 gene, the last one was situated in COL1A2 gene (Table 1). Further, 8 DNA changes was found in coding sequences (exons), one was identified in non-coding (intronic) part of one of collagen type I genes. This data was previously described in Hrušková et al. (2015) [4] and Hrušková et al. (2016) [5].

Disclosure

The author reports no conflicts of interest in this work.

Acknowledgement

This study was supported by the grants SVV-2016-260267, PRVOUK P24/1LF/3 and UNCE 204011 from the Charles University.

References

[1] Forlino A, Cabral WA, Barnes AV, Marini JC, New perspectives on osteogenesis imperfecta, Nat Rev Endocrinol 7 (2011), 540–557

[2] Endotext [homepage on the Internet]. MDText.com, Inc. Available from: `http://www.endotext.org/chapter/osteogenesis-imperfecta/7/`. Accessed April 15, 2015. Accessed June 16, 2014.

[3] Dalgleish R, The human type I collagen mutation database, Nucleic Acids Res 25 (1997), 181 187.

[4] Hrušková L, Mařík I, Mazurová S, Martásek P, Mazura I, COL1A2 gene analysis in a Czech osteogenesis imperfecta patient: a candidate novel mutation in a patient affected by osteogenesis imperfecta type 3, Advances in Genomics and Genetics 5 (2015), 275-281.

[5] Hrušková L, Fijalkowski I, Van Hul W, Mařík I, Mortier G, Martásek P, Mazura I, Eight mutations including 5 novel ones in the COL1A1 gene in Czech patients with osteogenesis imperfecta. Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub 160(3) (2016), 442-447.

Direct Home BP Telemonitoring System – Suitable Tool for Repeated Out-of-Office BP Monitoring

Jan Peleska[1], Jan Muzik[2], Marek Doksansky[2], David Gillar[2], Jan Kaspar[2], Karel Hana[2], Milan Polacek[2]

[1] 3rd Department of Medicine, General Faculty Hospital, Prague, Czech Republic

[2] Faculty of Biomedical Engineering, Czech Technical University, Prague, Czech Republic

Abstract

Direct home BP telemonitoring can eliminate the not always reliable BP values reported by the patient due to intentional or unintentional transcription errors. The presented telemedicine system transfers data directly, without any patient interaction, from a BP measuring device (BPMD) via a Bluetooth interface and sends them to a telemedicine server. Measurements can be sent either directly using a Intel Compute Stick mini-PC or indirectly via a mobile phone application that uses Apple HealthKit as an intermediate storage. The web logbook is based on ESH standardised logbook transferred to Excel. This enables an easy calculation of the average BP across several days. A chart and table with a daytime BP profile partially mimics 24-h ambulatory BP monitoring (ABPM). The patient's logbook is accessible to both the patient and the physician via a web application. It can be also generated as a pdf and sent to the physician by email, alternatively it can be printed. Moreover, the proposed system offers direct information about the detection of an irregular heartbeat rhythm during a BP measurement that can be easily distinguished in the logbook. Using the latest HL7 standard, the FHIR, the measurements can be directly sent to a hospital information system. This may help in the early detection of asymptomatic atrial fibrillation and in the prevention of its serious complications.

Keywords

Direct home blood pressure telemonitoring, HealthKit, Irregular heartbeat rhythm, Telemedicine

Correspondence to:

Jan Peleška
3rd Department of Medicine, General Faculty Hospital
Address: Karlovo nám. 32, Prague, Czech republic
E–mail: jan.peleska@seznam.cz

IJBH 2017; 5(1):42–44
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

The current European Society of Hypertension Practice Guidelines for home blood pressure monitoring (HBPM) contain a standardised BP logbook [1]. The PC presentation of these Guidelines for indirect BP telemonitoring has already been shown [2]. However, BP values reported by the patient may not always be reliable due to transcription errors. Some patients even select the more optimistic lower BP values from multiple measurements. Mistakes can altogether reach up to 30 % of all reported BP values. Therefore, direct HBP telemonitoring can overcome this obstacle.

The 2016 ESC Guidelines for the management of atrial fibrillation [3] describe this dangerous arrhythmia (AF), which is independently associated with a two-fold increased risk of all-cause mortality in women and a 1,5-fold increase in men. Death due to stroke can be largely mitigated by anticoagulation. AF appears with greater prevalence in older individuals, in patients with hypertension (the most frequent cardiovascular disease) and other conditions.

The diagnosis of AF requires rhythm documentation using an electrocardiogram (ECG). Individuals with AF may be symptomatic or asymptomatic ("silent AF"). The detection of asymptomatic AF by new technologies including BP machines with AF detection algorithms has not yet been formally evaluated against an established arrhythmia detection method.

When the BP measuring monitor detects an irregular rhythm two or more times during the measurement, the irregular heartbeat symbol will appear on the display with the measurement values.

However, some patients do not notice or report its presence to their physician.

The proposed solution enables direct telemonitoring of both BP values and information about detected irregular

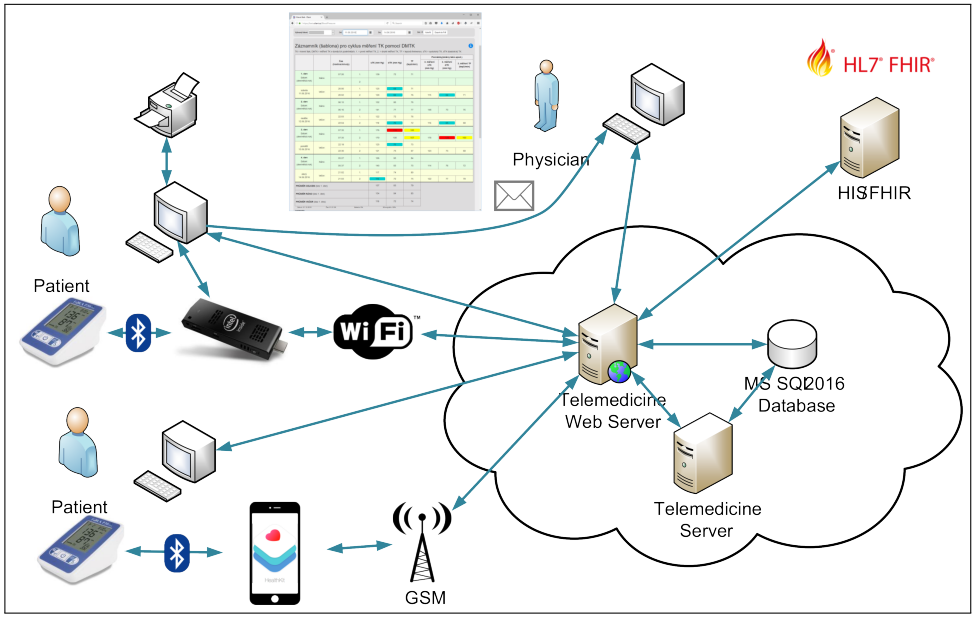

Figure 1: System architectures.

heartbeats corresponding to various arrhythmias, one of the most dangerous among them being AF.

2 Methods

The patient's set consists of the validated BPMD Fora P30+ (ForaCare, USA, available in the Czech Republic) that is wirelessly connected to a low cost mini PC – currently we are using the Intel Compute Stick (Intel Corp., USA) - but any computer able to run Windows 7 or later with Bluetooth (BT) 2.0 and which has internet connection can be used. Since the mini PC does not have a display, a special method for connecting to Wifi has been implemented. We have developed separate software that can be downloaded from the internet (without installation, using ClickOnce technology) that generates an encrypted file with Wifi credentials and saves it to a flash disk that is that used to transfer the file to the mini PC. For patients without an internet connection a set with a GSM modem can be used.

The mini PC contains software that awaits a BT connection from the BPMD and immediately saves it to the local database and then attempts to upload the measured values to the central server via secured web services (Figure 1). Both patients and physicians have access to the measured data via a web application.

During the last year more BPMD models supporting data transfer using the BT interface were introduced. Some of them with AF detection algorithms can diagnose AF with high probability, some generally detect an irregular heartbeat rhythm (IHR). This is defined e.g. as a rhythm with a 25% lower or 25% higher heart rate than that of the average heart rate detected while the monitor is measuring the systolic and diastolic BP.

Internally, (IHR) is treated as another measured variable and is transferred to the server. In the electronic logbook, measurements containing detected (IHR) are displayed with the bold-font-highlighted heart rate value and the irregular heartbeat symbol.

There is a chance to detect IHR when the recommended monitoring schedule for hypertension is used: seven-day home measurements before each clicic/office visit and once or twice per week in the long-term follow-up.

While in most cases the BPMD in the home environment is used more or less at the same location, in some cases it is more practical to use a mobile phone for the transfer of BP values between the BPMD and the server. For such cases we have developed the Diani Connector mobile application that automatically synchronises the data. The Diani Connector is currently only available for phones running on iOS since it makes use of Apple HealthKit (HK) as an intermediate storage. Many BPMD manufacturers are providing public mobile applications that store the data from the device in the HealthKit database where it can be read by any approved application. The advantage of such an approach is the HK providing a single unified interface regardless of the communication interface used by the BPMD. Therefore, the number of supported BPMDs is much higher and we do not need to know and implement its communication protocol. The disadvantage is that only a subset of data can be stored there. For example, systolic BP as well as the diastolic BP and heart rate can be stored there. On the other hand, information about an irregular heartbeat detected by the BPMD is very specific information that the HK cannot store.

3 Results and Discussion

The web logbook is based on ESH standardised logbook [1] transferred to Excel. This enables an easy calculation of the average systolic and diastolic BP (total, morning and evening) from several monitoring days. According to the current guidelines, the value of the total average BP is a criterion to diagnose hypertension, controlled hypertension with therapy or normotension (< 135/85 mmHg). The additional calculation of average morning and evening BP enables a better titration of pharmacotherapy. We implemented the version presented in the 2012 London European Meeting on Hypertension [4] that introduced differently coloured cells for both extremely high and extremely low BP and heart rate levels with warnings and recommendations for the patient to contact his/her physician within hours or days. Similarly, results of a mean BP evaluation are shown in different colours – an increased BP mean and a suitable range of the target BP mean.

Office blood pressure (BP) is usually higher than BP measured out of office, which has been attributed to the alerting response and anxiety (white coat effect).

24-h BP monitoring (ABPM) is currently considered the reference for out-of-office BP, but it is not so suitable for repeated BP monitoring. On the contrary, the direct home BP telemonitoring system with reliable BP values can fulfil the task of repeated BP monitoring during the titration of pharmacotherapy in the majority of cases easily.

Before printing or exporting to pdf, patients are asked to fill in their medication information. If a sequence of multiple daily measurements (e.g. each hour) is detected, the report will also contain a chart and table with a daytime BP profile for easier detection of hypotension at the time of an antihypertensive drugs' peak effect. The report can be viewed via the web application or can be generated as a pdf and sent to a physician by email or be printed and brought in paper form. Currently we are providing patients preconfigured sets and patients only need to setup a Wifi connection.

4 Conclusion

Advantages of the presented solution are:

- Measured values are transferred automatically – this eliminates the chance of patient transcription errors or a forgotten logbook when visiting the physician.
- Detection of irregular heartbeat – condition of in time diagnostics of AF and preventive therapy.
- Data are stored permanently and can be used for long-term follow up.
- Multiple deployments of the solution are possible – every physician or organisation can have the data under their control.

References

[1] G. Parati, G.S. Stergiou, R. Asmar, et al. European Society of Hypertension Practice Guidelines for home blood pressure monitoring. J Hum Hypertens 24 (2010), 779-785.

[2] J. Peleska, Z. Rotal. Home blood pressure monitoring, International Journal on Biomedicine and Healthcare 2015; 3 (2): 25–26

[3] P. Kirchhof, S. Benussi, D. Kotecha et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. European Heart Journal (2016) 37, 2893-2962

[4] J. Peleska. A new version of home blood pressure excel logbook for use in indirect telemonitoring, Journal of Hypertension, Vol 30, e-Supplement A, April 2012, e491

Detection of Unrecoverable Noise Segments in BSPM

Matěj Hrachovina[1], **Lenka Lhotská**[2]

[1] Department of Cybernetics, FEE, CTU, Prague, Czech Republic

[2] CIIRC, CTU, Prague, Czech Republic

Abstract

Thoughts on how randomly induced noise can be detected in BSPM processing.

Keywords

CRT, BSPM, Preprocessing, Noise detection

Correspondence to:

Matěj Hrachovina
Department of Cybernetics, FEE, CTU
Address: Karlovo náměstí 13, 121 35 Praha 2
E–mail: hrachmat@fel.cvut.cz

IJBH 2017; 5(1):45–48
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

In a project aimed at assessing the effectiveness of Cardiac Resynchronization Therapy (CRT), Body Surface Potential Mapping (BSPM) is one of the methods chosen to help quantify the therapy progress. A tool is being developed to evaluate the BSPM signals and model the propagation of the activation wave along the myocardium. BSPM signals are acquired on a system with 120 unipolar electrodes. Measured signals from individual leads are distorted by a combination of motion artefacts, mains hum and random noise arising from improper skin-electrode contact. Therefore we can't design a single filter to denoise all the signals in one batch and have to use more complex preprocessing techniques.

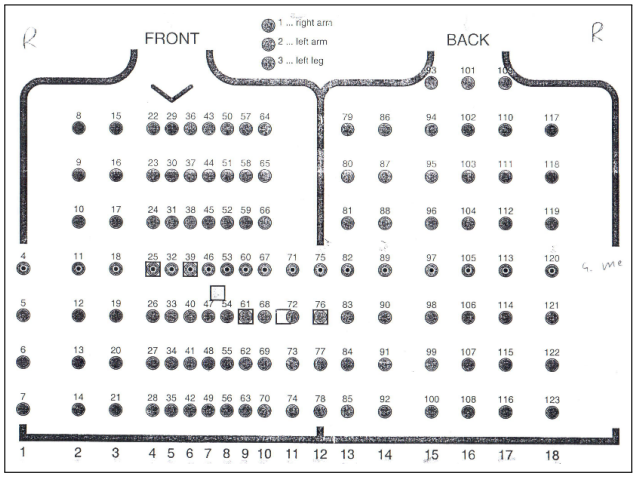

Figure 1: BPSM electrode placement.

Further processing techniques involve automatic QRS labeling. The labels are then used to calculate propagation maps and then to locate the origin of the electrical signal, so the stress on high preprocessing quality is crucial. We need to be able to label the waveforms with precision in the order of milliseconds (±1 sample with our sampling frequency).

2 CRT data

The signals are measured using Active Two hardware from BioSemi[1]. It has 120 unipolar electrodes placed unevenly along the chest and back in strips (Figure 1) plus three electrodes on the limbs to form Einthoven's triangle for reference.

The sampling frequency for all channels is 1024 Hz and samples are quantized into 24 bits. For each pacemaker setting we record approximately a 2 minute interval.

The signals suffer from 3 types of noise: motion artefacts and isoline drift, mains hum and noise induced by varying skin-electrode impedance.

Motion artefacts and isoline drift can be easily removed with moving average filters. We try to use FIR filters wherever possible in order to keep the phase of the signals intact.

Mains hum is best removed by not inducing it at all during measurement, but this is not always possible. Otherwise it's removed by IIR notch filter.

Varying skin-electrode impedance does not really induce noise. It distorts the measured signal's phase and amplitude. The signal cannot be recovered without knowing the exact impedance for each affected sample.

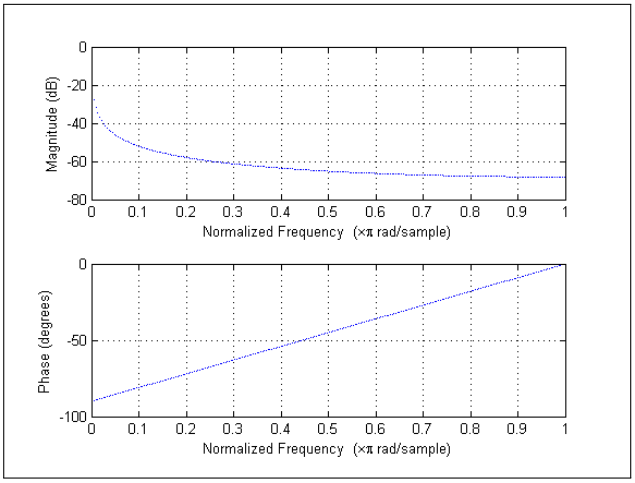

Figure 2: Impulse response of MA filter.

3 Basic preprocessing

Processing was conducted using MATLAB[2]. All signals are pre-processed using a sequence of filters. We assume all noise, except mains hum, has a random frequency distribution. We will also be averaging the signals into one beat. Therefore we don't need the preprocessing filters to have high attenuation in the stop band. Instead we prefer minimum phase distortion and use FIR filters.

3.1 Isoline removal

In the first phase we remove isoline drift caused by breathing. For this we use a moving average filter with a cut-off frequency of 0.6 Hz. Frequency characteristics of this filter can be seen in Figure 2.

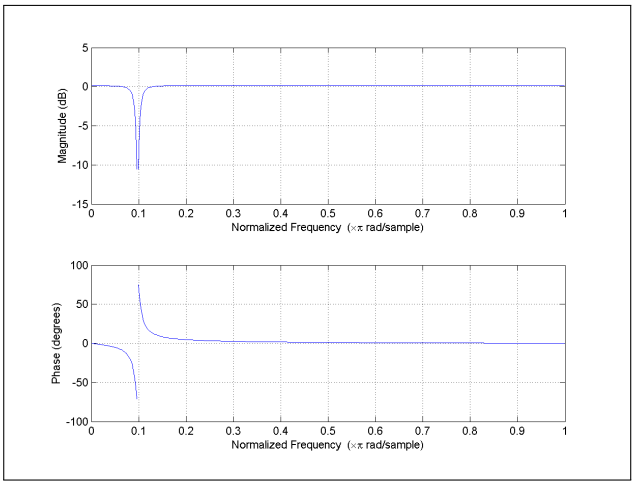

Figure 3: Impulse response of notch filter.

3.2 Mains hum removal

Mains hum usually has a narrow frequency distribution and so we need to use a narrow band FIR filter to remove it. We chose a notch filter centered at 50 Hz with 3 dB bandwidth of 6 Hz. Frequency characteristics of this filter can be seen in Figure 3.

3.3 Motion artefacts removal

To remove motion artefacts and smoothen the signal we use a moving average filter with cut-off frequency of 80 Hz in the final preprocessing phase.

Figure 4 shows the change in signal amplitude and frequency characteristics after passing through the filtering sequence.

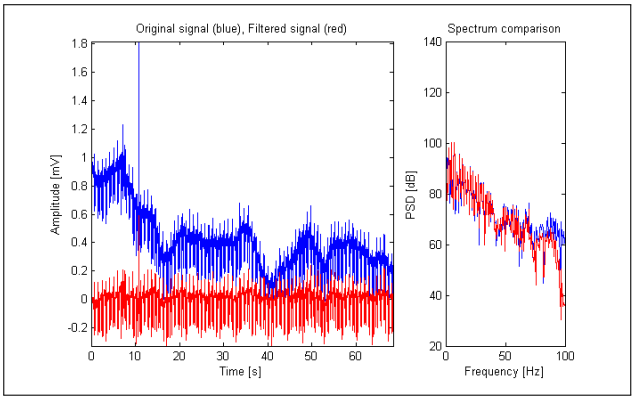

Figure 4: Comparison of signal before and after filtration.

4 Unrecoverable noise

The electrode strips are designed to reduce the time needed to apply them on the patient. Unfortunately they only come in one size. So for smaller patients they need to be folded to get the electrodes into desired positions, which increases the tension on the adhesive. The adhesive used on the electrodes also doesn't work well on patients with dry skin. This means, that the electrodes tend to lose contact unexpectedly, resulting in varying electrode-skin impedance.

In the samples where the impedance is unstable, the signal is lost and cannot be recovered by any filtration technique, unless we would also record the impedance of the electrodes for each sample.

Samples affected with this problem have varying amplitude, ie. the change in the nature of the signal is unpredictable. Averaging beats from signals containing this type of noise with unaffected beats would also not help, as the shifts in amplitude are of a greater order and propagate dominantly into the resulting average. We must therefore make sure that segments which contain noise induced by unstable skin-electrode impedance are marked and are excluded from further processing.

5 Unrecoverable noise detection

Samples where skin-electrode impedance instability ruined the signal must not get into further processing, so they don't bias the propagation maps.

In a first approach we tried to detect them by thresholding the signals' amplitude. In a second approach we segmented the signals into segments of 20 samples and thresholded the power of the segments. These approaches sometimes worked, but were very inaccurate.

For now we are working on two approaches based on the segmentation method. The first uses the power spectral density and the second uses the cepstrum to obtain the thresholded parameter.

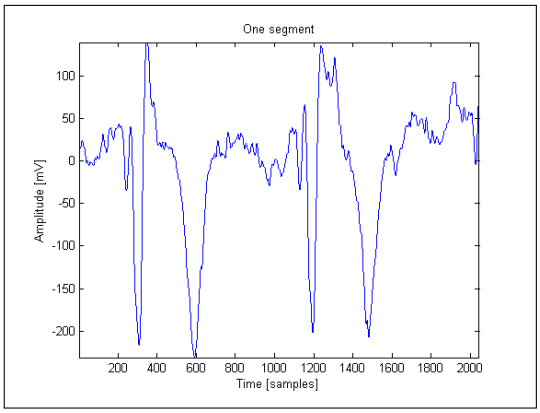

Figure 5: Segmentation.

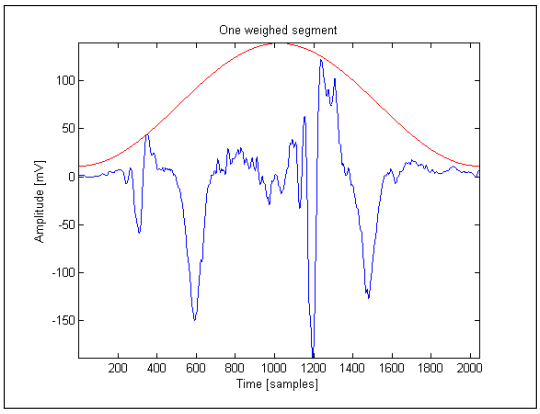

Figure 6: Weighing.

5.1 Segmentation

For both methods the signal is segmented into 2 seconds segments (2048 samples) and weighed with a Hamming window to minimize frequency overshoots in frequency analysis (FFT). The length of the segment was chosen to be 2 seconds, so that no matter where it is positioned, it will always contain at least one full beat. The segmenting window is moved with 50% overlap (1024 samples). Segmentation and weighing of the signal is outlined in Figure 5 and Figure 6.

5.2 Power spectral density

The thresholded parameter for this method is obtained by calculating the PSD of a segment and correlating it with PSD of a signal without noise (reference beat).

The power spectral density is calculated using the FFT algorithm implemented in Matlab. The number of DFT points does not matter as long as it's the same as for the reference beat. We used 1024 point DFT. For correlating the two PSDs we used Matlab's cross-correlation function and we take the central value corresponding to no offset between the signals.

The resulting value is stored in a time series which is then thresholded.

5.3 Cepstrum

The thresholded parameter for this method is obtained by calculating cepstral coefficients and calculating the cepstral distance.

Cepstral parameters are calculated as the cosine transform of the logarithm of the fourier transform of the segment (Figure 7).

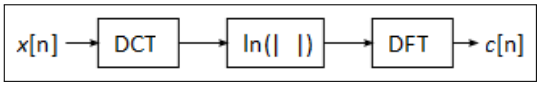

Figure 7: Cepstrum calculation.

To get a single value representing each segment, the Euclidean distance is calculated from the cepstral coefficients. The time series of the Euclidean distances is then thresholded.

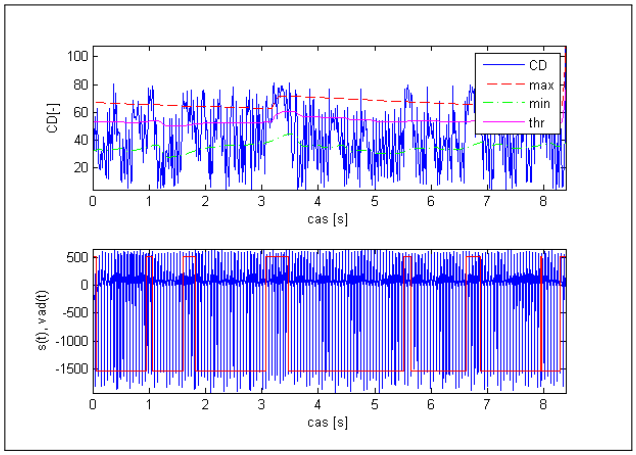

Figure 8: Thresholding.

5.4 Thresholding

To detect which segment of the signal is affected by unrecoverable noise, the time series obtained in the extraction phase (either from PSD or Cepstrum) is thresholded using dynamic thresholding.

It is basically a combination of two integrating filters:

$$y[n] = qx[n-1] + (1-q)x[n].$$

One is tracking the minima of the signal, the other is tracking maxima. If the difference between the maximum and minimum exceeds a given value, the segment is marked as noisy. An illustration of thresholding is displayed in Figure 8.

The resulting values for all leads are combined using logical OR into a single vector, which marks the times, where all leads in the recording are noise free.

6 Conclusion

We don't have enough data to make conclusions yet, but the cepstrum method looks more promising. It is also much less computationally demanding. The use of this method is crucial for labelling the beats and creating isochronous and potential maps.

Acknowledgement

The research is supported by the project No. 15-31398A "Features of Electromechanical Dyssynchrony that Predict Effect of Cardiac Resynchronization Therapy" of the Agency for Health Care Research of the Czech Republic and by grant application No. OHK3-033/16 of the CVUT SGS program.

References

[1] BioSemi B.V., Amsterdam, Netherlands.

[2] MATLAB 2012b, The MathWorks, Inc., Natick, Massachusetts, United States

Expanding Functionality of a Diabetes Smartwatch Application

Miroslav Muzny[1,3], **Jan Muzik**[2], **Eirik Arsand**[3,4]

[1] Spin-off Company and Research Results Commercialization Center,
The First Faculty of Medicine, Charles University in Prague, Czech Republic

[2] Department of Information and Communication Technologies in Medicine,
Faculty of Biomedical Engineering, Czech Technical University in Prague, Czech Republic

[3] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromso, Norway

[4] UiT The Arctic University of Norway, Department of Clinical Medicine, Tromso, Norway

Correspondence to:

Miroslav Mužný
Spin-off Company and Research Results Commercialization Center,
The First Faculty of Medicine, Charles University, Czech Republic
Address: Studnickova 7, Prague 2, 12800
E–mail: mmuzny@gmail.com

IJBH 2017; 5(1):49–50
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Background

As a result of several research projects we have developed a smartphone Diabetes Diary application [1] and a companion application for the Pebble smartwatch [2]. Its main purpose is to provide a digital diabetes diary with advanced functions easily available from the users' wrist.

2 Introduction

Our existing Pebble smartwatch application offers two-way synchronization of registrations with both the iOS and Android version of the Diabetes Diary application. It offers basic physical activity tracking and provides alarms for remembering to measure blood glucose. However, the original design of the application was influenced by various limitations of the Pebble smartwatch hardware and software, which we address and present improvements on in this presentation.

3 Methods

Based on improvements in new generations of Pebble smartwatch, we have innovated several new functions, which we have implemented and tested as extensions of our existing Pebble smartwatch application. None of these functions are currently included in the public version of the application, which is available on the Pebble app store. No study on the usability of these new functionalities has been done on users yet.

4 Results

We have designed and implemented the following new functionalities:

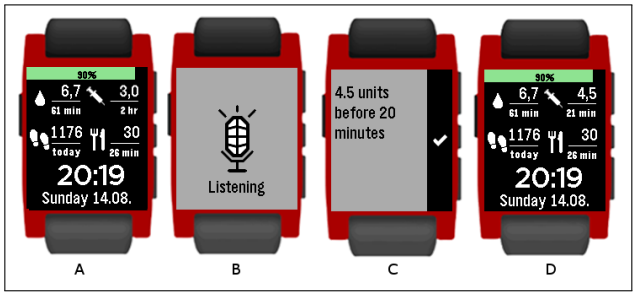

Figure 1: Making a new insulin registration using voice input.

Voice-based input method for entering new registrations Voice input has several advantages over the ordinary method using buttons – it allows to easily specify other properties of new registrations (e.g. time of the registration, insulin type etc.). Voice recognition is supported in multiple languages, see Figure 1.

Finding similar situations when entering new insulin registration We have utilized the integrated algorithm on the phone to provide a user overview of similar situations in the history of records when entering a new insulin registration. The most similar situation is transferred to the smartwatch and presented on the 'new insulin' registration screen, see Figure 2A.

Integration with the Nightscout application The Nightscout application was originally developed to allow parents of children with Type 1 diabetes to remotely monitor data from continuous blood glucose meters (CGM) [3]. We have modified our Pebble smartwatch application to show the real-time CGM value on the watch's main screen, see Figure 2B.

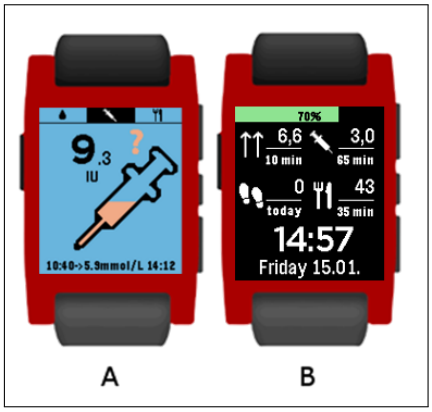

Figure 2: Showing the most similar situation (A) and Nightscout integration (B).

5 Discussion

Unfortunately, because of the company's economic problems, Pebble Technology Corporation, was recently acquired by a larger company, Fitbit Inc., resulting in a shutdown of the production of new Pebble smartwatches [4]. It is unclear whether the software will become a part of a new product. Therefore, to be able to continue doing research on innovative smartwatch-based self-management functions, we have to search for alternate open platforms such as Pebble, providing similar opportunities for implementation of health applications.

References

[1] Arsand E, Skrovseth SO, Joakimsen RM, Hartvigsen G. Design of an Advanced Mobile Diabetes Diary Based on a Prospective 6-month Study Involving People with Type 1 Diabetes. The 6th International Conference on Advanced Technologies and Treatments for Diabetes, February 27.-March 2. 2013, Paris. France.

[2] Arsand E, Muzny M, Bradway M, Muzik J, Hartvigsen G. Performance of the first combined smartwatch and smartphone diabetes diary application study. Journal of diabetes science and technology, 9(3), 556-563.

[3] Nightscout. Available from: `http://www.nightscout.info/`.

[4] "What Does Pebble Joining Fitbit Mean For Me?" Available from: `http://help.getpebble.com/customer/portal/articles/2663228-what-does-pebble-joining-fitbit-mean-for-me-?b_id=8309`.

Principles of Medical Decision Making. Quantifying Uncertainty: Bayesian Approach and Statistical Decision Support

Jana Zvárová[1]

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

[2] EuroMISE Mentor Association, Prague, Czech Republic

Correspondence to:

Jana Zvárová
Institute of Hygiene and Epidemiology, First Faculty of Medicine,
Charles University and General University Hospital in Prague
Address: Studničkova 7, Praha 2, 128 00, Czech Republic
E–mail: zvarova@euromise.cz

IJBH 2017; 5(1):51
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

Abstract

Medical care is often said to be the art of making decisions without adequate information. Physicians must frequently choose treatment long before they know which disease is present .Even when illness is known, one must usually select from among several treatment options and the consequences of each cannot be foretold with certainty. Indeed, uncertainty is intrinsic to the practice of medicine. Clinicians routinely make decisions for and with their patients that are complex, under time constraints, and involve risks or uncertain outcomes. These decisions may be made either with certainty or with some uncertainty. Essentially, there are three possible "operating systems" of reasoning in medicine under uncertainty.

First is probability theory and its application, especially using Bayes rule, second chaos theory and its uses in clinical research and third fuzzy logic and fuzzy set theory in handling imperfect or hard to interpret data. We focus on uncertainty and probability quantifying uncertainty and its estimate using data, using personal experience or using published experience. Finally, we mention some decision support systems based on probabilistic and statistical approaches.

References

[1] Sox HC, Blan MA, Higgins MC, Marton KI: Medical Decision Making. Butterworth-Heinemann, London 1988

[2] Jenicek M: Foundations of Evidence-Based Medicine. The Parthenon Publishing Group, New York 2003

[3] Van Bemmel JH. Musen MA: Handbook of Medical Informatics. Houten-Diegem 1997

[4] Zvárová J., Svačina Š., Valenta Z. et al: Systems for Medical Decision Support, Carolinum, Prague 2009

[5] Blobel B, Hasman A., Zvárová J (Eds). Data and Knowledge for Medical Decision Suport, Studies in Health Technology and Informatics 186, 2013

[6] Kalina J., Zvárová J.: Regression Modelling: A Fundamental Tool in Clinical Decision Support. International Journal on Biomedicine and Healthcare,5 (1), 2017, 21-27

From Clinical Practice Guidelines to Computer Interpretable Guidelines

Arie Hasman[1]

[1] Dept. of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Correspondence to:

Arie Hasman
Dept. of Medical Informatics, Academic Medical Center,
University of Amsterdam, Amsterdam, The Netherlands
Address: Meibergdreef 15, 1105 AZ Amsterdam
E–mail: a.hasman@amc.uva.nl

IJBH 2017; 5(1):52
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

Abstract

Guidelines are among us for over 30 years. Initially they were used as algorithmic protocols by nurses and other ancillary personnel. Many physicians regarded the use of guidelines as cookbook medicine. However, quality and patient safety issues have changed the attitude towards guidelines. Implementing formalized guidelines in a decision support system with an interface to an electronic patient record (EPR) makes the application of guidelines more personal and therefore acceptable at the moment of car

The goal of the presentation is to obtain, via a literature review, an insight into factors that influence the design and implementation of guidelines.

I will present studies that enable us to explain the characteristics of high-quality guidelines, and new advanced methods for guideline formalization, computerization, and implementation. We show how the guidelines affect processes of care and the patient outcome. We discuss the reasons of low guideline adherence as presented in the literature and comment upon them.

Developing high-quality guidelines requires a skilled team of people and sufficient budget. The guidelines should give personalized advice. Computer-interpretable guidelines (CIGs) that have access to the patient's EPR are able to give personal advice. Because of the costs, sharing of CIGs is a critical requirement for guideline development, dissemination, and implementation. Until now this is hardly possible, because of the many models in use. However, some solutions have been proposed. For instance, a standardized terminology should be imposed so that the terms in guidelines can be matched with terms in an EPR. Also, a dissemination model for easy updating of guidelines should be established. The recommendations should be based on evidence instead of on consensus. To test the quality of the guideline, appraisal instruments should be used to assess the guideline as a whole, as well as checking the quality of the recommendations individually. Only in this way optimal guideline advice can be given on an individual basis at a reasonable cost.

Data Mining and Machine Learning

Petr Berka[1,2]

[1] University of Economics, Prague, Czech Republic

[2] University of Finance and Administration, Prague, Czech Republic

Correspondence to:

Petr Berka
University of Economics
Address: W. Churchill Sq. 4, 130 67 Prague
E–mail: berka@vse.cz

IJBH 2017; 5(1):53–54
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

The rapid growth of data collected and stored in various application areas brings new problems and challenges in their processing and interpretation. While database technology provides tools for data storage and "simple" querying, and statistics offers methods for analyzing small sample data, new approaches are necessary to face these challenges. These approaches are usually called knowledge discovery in databases (KDD) or data mining. These two terms are often used interchangeably. We will support the view that knowledge discovery is a broader concept covering the whole process in which data mining (also called modeling or analysis) is just one step in which machine learning or statistical algorithms are applied to preprocessed data to build (classification or prediction) models or to find interesting patterns. We thus understand knowledge discovery in databases as the

Non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from data [3],

or as an

Analysis of observational data sets to find unsuspected relationships and summarize data in novel ways that are both understandable and useful to the data owner [5].

2 The Process of Knowledge Discovery in Databases

According to the CRISP-DM methodology [2] the KDD process consists of business understanding, data understanding, data preprocessing, modeling, evaluation and deployment steps.

Business understanding is the initial phase that focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives

The **data understanding** phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

The **data preparation** phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

In the **modeling** phase various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed. This phase corresponds to the data mining step in the narrow sense.

At the **evaluation** stage in the project the build model (or models) appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Creation of the model is generally not the end of the project. Depending on the requirements, the **deployment** phase can be as simple as generating a report or as complex as implementing a repeatable data mining pro-

cess. In many cases it will be the customer, not the data analyst, who will carry out the deployment.

3 Data Mining Tasks, Methods and Applications

Knowledge discovery in databases is commonly used to perform the tasks of data description and summarization (to find concise description of characteristics of the data, typically in elementary and aggregated form), segmentation (to find interesting and meaningful subgroups where all members of a subgroup share common characteristics), concept description (to find understandable description of concepts or classes), classification (to build classification models, which assign the correct class label to previously unseen and unlabeled objects), prediction (to build models that predict changes of a variable over time), dependency analysis (to find significant dependencies or associations between data items or events), or deviation detection (to find significant changes in the data from previously measured or normative values).

There is a wide range of methods and algorithms that can be used to solve these tasks (see e.g. [4] or [6]. First group are statistical data analysis methods. Among them the most used are regression analysis, that can be used for classification (in case of logistic regression) or prediction, discriminant analysis, that can be used for classification tasks, and cluster analysis, that can be used for segmentation tasks. Beside these methods that can be directly used for the modeling step, statistical techniques like factor analysis or principal component analysis can be used for attribute transformations during data preprocessing. The second group are methods and algorithms from the area of machine learning. Among them, most popular are algorithms for decision tree induction (usually used for classification), algorithms for learning neural networks (for classification or prediction in case of multilayer perceptrons or RBF networks, or for segmentation in case of SOM networks), Bayesian classifiers (naïve Bayesian classifiers or Bayesian networks), or algorithms that use instance-based approach (like e.g. k-NN algorithm). In last decade, the so called support vector machines (SVM) gained increasing popularity in the data mining community. However, we must stress here that there is no "best" algorithm that will outperform (in terms of classification or prediction accuracy) the other algorithms on any problem [7].

KDD can be applied in various domains: banking and finance, insurance, life sciences, retail, technical diagnostics, computer networks, social networks e.t.c. Let us consider an example from medical domain, the analysis of atherosclerosis risk factors data with the aim to build a model that will differentiate between normal and risky patients [1].

References

[1] Berka,P. – Rauch,J. – Tomečková,M. (2009). Data Mining in the Atherosclerosis Risk Factor Data. In: (Berka,P. – Rauch,J. – Zighed,D,A. eds.) Data Mining and Medical Knowledge Management: Cases and Applications. IGI Global.

[2] Chapman,P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS Inc.

[3] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (Eds). (1996). Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press.

[4] Han, J.& Kamber,M. (2001). Data Mining: Concepts and Techniques. Academic Press.

[5] Hand, D., Mannila, H. & Smyth, P. (2001). Principles of Data Mining. MIT Press.

[6] Witten, I.H. & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufman

[7] Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural Computation, 8(7), 1341–1390.

[8] Zvárová,J. – Svačina,Š. – Valenta,Z. et al. (2009) Biomedicínská informatika III. Systémy pro podporu lékařského rozhodování. Nakladatelství Karolinum, Universita Karlova, Praha.

Managing and Representing Knowledge in Decision Support Systems

Bernd Blobel[1,2]

[1] Medical Faculty, University of Regensburg, Germany

[2] eHealth Competence Center Bavaria, Deggendorf Institute of Technology, Germany

Correspondence to:

Bernd Blobel
Medical Faculty, University of Regensburg, Germany
Address: Regensburg, Germany
E–mail: bernd.blobel@klinik.uni-regensburg.de

IJBH 2017; 5(1):55–56
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

1 Introduction

Business processes are a matter of interoperability, i.e. communication and cooperation, between principals according to the definition of the Object Management Group (OMG) such as persons, organizations, devices, applications, and single components or objects [1], also called actors. In that context, the information cycle [2] must be mastered between the actors involved in the business process by observing process conditions, transforming those data into information for communication, and taking the right action to achieve the business objectives [3]. Performing the information cycle phases requires knowledge.

2 Knowledge Representation and Management for Decision Support Systems

Two basic services must be provided for enabling a reasonable business case: a) explicit, computable representations of knowledge about underlying business concepts and business processes, and b) services to facilitate knowledge sharing between all business system components. This also applies for DSS as business system components. Deploying KR techniques such as frames, rules, tagging, and semantic networks, a good KR has to manage both declarative and procedural knowledge.

In advanced healthcare settings, many different disciplines must cooperate for providing high quality, safe and efficient health services. Knowledge of a discipline, also called domain of discourse is created, represented, and maintained by domain experts using their methodologies, terminologies and ontologies. For designing and implementing DSS, knowledge representation and management must be realized at epistemological (cognitive and philosophical), the notation (formalization and conceptual) level, and computational or implementation level [4]. For harmonizing knowledge from different domains in an interoperability scenario, a common notation is inevitable, thereby enhancing expressivity and formalization of the representation in the continuum of languages from natural languages, glossaries and data dictionaries, thesauri and taxonomies, meta-data and data models and finally formal ontologies up to a level of sufficient commonality.

A crucial component of DSS is the knowledge representation formalism. That formalism is used to encode units of knowledge, i.e. specialized problem-solving expertise such as facts, rules, procedures, which are stored in the knowledge base. Each of those units of knowledge contains sufficient knowledge to make a single decision. Examples of such knowledge representation formalism for encoding units of knowledge in the knowledge base are medical knowledge representation languages such as PROforma, Asbru, EON, Arden Syntax, GELLO, GLIF, Archetypes, HL7 Clinical Statements, and the recently developed FHIR approach. For more details see, e.g., [5].

The knowledge base is linked to the inference engine or event monitor that executes knowledge units in combination with process-related data (e.g. patient data from EHR systems) to produce tailored context-specific process decisions (interventions). The interference engine deploys appropriate models according to available decision theories. Therefore, a DSS consists of three basic components: a) the knowledge base (database), b) the model base and analytical tools, and c) the user interface [6].

3 Future Challenges and Conclusions

Currently health systems are objects of fundamental organizational, methodological and technological paradigm changes. The organizational paradigm turns from the organization-centric towards the person-centric paradigm. Regarding the methodologies deployed, medicine advances from an empiric, phenomenological approach towards personalized, preventive, predictive, participative precision system medicine, covering the continuum from elementary particle to society. The technological changes cover distributed systems including the Internet of Things (IoT), mobile technologies, nano- and molecular technologies, knowledge representation and knowledge management, artificial intelligence, big data & business analytics, cloud computing, and social business. With increasing complexity and flexibility of decision challenges, the aforementioned methodologies and technologies have to be appropriately integrated in DSS design and implementation. This requires another approach to design, development and implementation of business systems including DSS, going beyond the current level of abstraction and formalization. In consequence, a mathematical, system-theoretical, architecture-centric, ontology-driven approach has to be deployed as demonstrated, e.g., in [3, 7] and standardized at ISO (e.g. [8, 9].

More information on the subject can be found in [10].

References

[1] Object Management Group, Inc. `http://www.omg.org`. (last access 15 November 2016).

[2] Van Bemmel J., Musen M. (eds) Handbook of Medical Informatics. Heidelberg: Springer; 2002.

[3] Blobel B. Analysis, Design and Implementation of Secure and Interoperable Distributed Health Information Systems. Series Studies in Health Technology and Informatics, Vol. 89. Amsterdam: IOS Press; 2002.

[4] Dahn, BI, Dörner H, Goltz H-J, Grabowski J, Herre H, Jantke KP, Lange S, Posthoff C, Thalheim B, Thiele H. Grundlagen der Künstlichen Intelligenz. Berlin: Akademie-Verlag; 1989.

[5] Blobel B. Knowledge Representation and Management Enabling Intelligent Interoperability – Principles and Standards. Stud Health Technol Inform. 2013;186:3-21.

[6] Wikipedia. Decision Support Systems. (last access 15 November 2016)

[7] Blobel B. Architectural approach to eHealth for enabling paradigm changes in health. Methods Inf Med 2010;49,2:123-134.

[8] International Organization for Standardization. ISO 13606 Health informatics – EHR communication. Geneva: ISO; 2016.

[9] International Organization for Standardization. ISO 22600 Health informatics – Privilege management and access control. Geneva: ISO; 2014.

[10] Blobel B. Knowledge Representation and Knowledge Management as Basis for Decision Support Systems. International Journal on Biomedicine and Healthcare 2017; 5(1):13–20

Bayesian Networks for Uncertain Knowledge Representation

Radim Jiroušek[1,2]

[1] Faculty of Management, University of Economics, Prague

[2] Institute of Information Theory and Automation Czech Academy of Sciences

Correspondence to:

Radim Jiroušek
Faculty of Management, University of Economics, Prague
Address: Jarošovská 1117/II, 377 01 Jindřichův Hradec
E–mail: radim@utia.cas.cz

IJBH 2017; 5(1):57–58
received: November 20, 2016
accepted: January 5, 2017
published: February 20, 2017

In the course of last fifty years, a number of different models for knowledge representation have been developed. When uncertain knowledge is considered – and in our opinion, deterministic knowledge applies to very specific situations only – one can consider models based on some of the calculi proposed specifically for this purpose. The oldest one is probability theory but many others appeared in the considered period; from many-valued and fuzzy logics, through rough sets theory to approaches based on non-additive measures, e.g., possibility theory [2] and Dempster-Shafer theory of evidence [9]. Though most of the content of the lecture could be expressed also in the framework of other uncertainty calculi [5], in the lecture we shall discuss only probabilistic models. Namely, its basic concept of a *multidimensional probability distribution* (or probability measure) is a suitable tool to represent relationships between (or among) features, characteristics, or generally random variables (events). Less formally said, multidimensional probability distributions are a tool to represent knowledge [7].

The strength of probability theory for knowledge representation has been shown in many publications, e.g., the seminal book by Judea Pearl [7]. If the reader prefers an informal way of presentation, then they can get acquainted with a probabilistic knowledge representation on a forensic application in [1]. In this paper the reader learns not only how the probabilistic apparatus is used to represent knowledge (the case of O. J. Simpson) , but mainly, how it can be used for inference. In the case of the cited paper, the reader can learn how to deduce a posterior probability (given the published evidences on the case) that O. J. Simpson is guilty. For this purpose the author employs Bayesian networks.

The family of Graphical Markov models [6], to which Bayesian networks [3] belong, have become a very popular way for multidimensional probability distribution representation and processing. In the lecture it will be shown on simplified examples (e.g. from rheumatology [4]) that it is its *independence structure* [10], and not the orientation of edges what influences the validity of models most. The fact that some of independence systems can be successfully represented with the help of graphs is reflected in the general title: *graphical modeling*.

The graphs used to describe Bayesian networks are (acyclic) *oriented graphs*, ie., graphs, edges of which are represented by arrows. The edges indicate the closest relationships (mutual dependence) of the variables (there is a one-to-one correspondence between the variables and the nodes of the graph). The variables, which are not directly connected by an edge are *conditionally independent*, and a system of conditional independence relations valid for the distribution defines the above mentioned independence structure. Therefore, one cannot understand Bayesian networks without understanding the notion of conditional independence. Let us stress that it is this very notion that makes representation of high-dimensional probability distributions possible.

Different graphical models employ different types of graphs to describe the independence structure of distributions. For example, decomposable models use for this purpose undirected graphs, and special models need more complicated graphical tools like chain graphs, hypergraphs, or, annotated (di)graphs. Using oriented graphs for Bayesian network definition has its advantages (especially their simplicity and comprehensibility), but also one disadvantage. Some of the users misinterpret the direction of edges as causality. It is true that causal networks (as

they are defined by Pearl [8]) use also acyclic directed graphs to represent their structure, but one must keep in mind that handling causal networks is different (more complicated) than handling common Bayesian networks. On the other hand, causal networks enable their users to deduce also the impact of an *intervention*, which cannot be done in usual Bayesian networks. The difference between the conditioning and the intervention can easily be described by a joke: *A statistician explains his friend that a probability that there are two bombs on board a plain is much less than there is one bomb. The next day, when meeting at a check-in counter at the airport, the statistician asks his friend why has a big suitcase. The friend replies: "A bomb. Just to decrease a probability that there will be another one on board."* Unfortunately, in practical situations the difference between intervention and prediction is not so obvious, and many people mix them up. This will also be one of the highlights of the lecture.

References

[1] Drábek, J., 2014. Analýza případu O. J. Simpsona pomocí bayesovský sítí. Část 1-3. *Kriminalistický sborník* 1-3, 61-66, 68-72, 58-67.

[2] Dubois, D. and Prade, H., 2001. Possibility Theory, Probability Theory and Multiple-valued Logics: A Clarification. *Annals of Mathematics and Artificial Intelligence* 32, 35-66.

[3] Jensen, F.V., 2001. *Bayesian Networks and Decision Graphs.* IEEE Computer Society Press. New York.

[4] Jiroušek, R., 1998. Bayesovské sítě - moderní technologie umělé inteligence. Lékař a technika, 29, 4, 79-88.

[5] Jiroušek, R. and Vejnarová, J., 2003. General framework for multidimensional models. *Int. J. of Intelligent Systems*, 18, 107–127.

[6] Lauritzen, S. L., 1996. *Graphical Models.* Oxford University Press.

[7] Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo.

[8] Pearl, J., 2009. *Causality: Models, Reasoning, and Inference.* (Second Edition) Cambridge University Press.

[9] Shafer, G., 1976. *A Mathematical Theory of Evidence.* Princeton University Press, New Jersey.

[10] Studený, M., 2005. *Probabilistic Conditional Independence Structures.* Springer, London.

[11] Zvárová, J. *et al.*, 2009. *Biomedicínská informatika III. Metody pro podporu rozhodování.* Karolinum. Kapitola 5, Bayesovské sítě. 295-340.

