IJBH

International Journal
on Biomedicine and Healthcare

IJBH 2015 ISSN 1805-8698

International Journal on Biomedicine and Healthcare

Volume 3 (2015), Issue 1

Main Topic:

Big Data Challenges for Personalised Medicine

Editors:

Pirkko Nykänen 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

Big Data Challenges for Personalized Medicine

Pirkko Nykänen[1], **Jana Zvárová**[2]

[1] University of Tampere, School of Information Sciences, Tampere, Finland

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

In June 2015 the International Joint Meeting EuroMISE 2015 was organized in Prague with the theme 'Big data challenges for personalized medicine'. The papers published in this issue have all been accepted and presented in this conference demonstrating the important research results and opinions on big data challenges. The conference demonstrated well that the challenges of big data in health informatics are not only in capturing and storing information but also in providing the methods and tools to analyze and manage big data. The identified challenges of big data include: Standards for consolidating, characterizing, validating and processing of data; ontologies for knowledge and relationships between knowledge entities such as genes, drugs, diseases, symptoms, patients and treatments; integration of various data sources and information systems and integration of environmental data with individual genomic measurements; and open access - availability, readability and usability of big data. However, many questions need to be solved with big data before we are able to improve research and exploit research outputs to improve health both at a public health level and at personalized medicine level. The papers published in this issue deal with many of these questions.

The keynote by Bernd Blobel tackles a very important issue of change and its impacts: when health care is changing towards ubiquitous personalized health it means also significant paradigmatic changes in the security and privacy ecosystem. The keynote by Edward Hammond outlines opportunities and challenges of big data e.g. how to establish quality and noise reduction sufficient to validate studies derived from big data and how to ascertain the completeness and consistency of data. The third keynote by Arie Hasman discusses structural equation modelling in relation to big data. Structural equation modelling enables determination of the extent to which a theoretical model is supported by sample data.

The paper by Berger and Beyr is focused on safety of private health data, and discusses e.g indirect disclosure of sensitive data from fully anonymized databases. Blobel, Ruotsalainen, Lopez and Gonzalez analyse in their paper the HL7 composite security and privacy domain analysis model from the perspectives architecture-centric, ontology-driven perspective. Engelbrecht explores in his opinion paper how we can improve health and develop more for personal health. The research paper by Jimenez-Lozano, Hlavsa, Pelletier, Exertier and de la Torre deals with development of an open bioinformatics platform for personalized treatment of cancer. The platform will enable translation of bioinformatics methods to the clinical environment. The paper by Kalina studies the statistical challenges of big data analysis in medicine with examples from cardiovascular genetics, biometric authentication and brain activity. The paper by Krsicka and Seidl explore the possibilities of big data to find associations between folate deficiency and autism. The opinion paper by Lhotska, Bursa, Huptych and Hrachovina discusses the use and applicability of big data to rare cases of medicine. The research paper by Pecen, Jirina and Novak presents the development of diagnostic decision support software to support detection and interpretation of tumor markers. The developed software has been in successfully used in a number of health care units. The research paper by Pokorny describes the big data storage and processing in personalized medicine. The paper concludes that big data makes it possible to develop new types of applications for personalized medicine, but the success requires that natural language processing is managed, pattern recognition algorithms are efficient for image and video sources and predictive modelling and effective statistical analysis tools are in use. Finally the research paper by Schlenker and Bohuncak tackle the important issue of security enhancement in hospital information systems. The paper proposes that security can be improved with multifactor authentication and keystroke dynamics.

The conference papers demonstrate well the challenges big data offer today for biomedical informatics - to connect molecular and cellular biology to the clinical world allowing us to consider individual variations and not simply population averages. But, with big data we need to take into account the security and privacy requirements for personal health data which means that data users should be accountable for the custodianship of personal medical information. This will be a challenge as we are dealing with both regulated and non-regulated healthcare environments and with reuse of data and secondary use of health data. 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.

Safety of Private Data in Big Data and Biomedicine

Jiří Berger[1], **Karel Beyr**

[1] Institute of Pathological Physiology, 1st Faculty of Medicine, Charles Universty in Prague, Prague, Czech Republic

Correspondence to:

Jiří Berger
Institute of Pathological Physiology
1st Faculty of Medicine, Charles Universty in Prague
Address: U nemocnice 5, 128 53 Prague 2
E–mail: jiri.berger@e-fractal.cz

IJBH 2015; 3(1):2–5
received: April 30, 2015
accepted: May 7, 2015
published: June 15, 2015

1 Aims of Research

Big Data have a great potential for research in biomedicine in many areas

- Analysis of patient segmentation, treatment price and result helps determine the medically and economically most efficient course of treatment for a given patient;
- Proactive identification of patients who would benefit from preventive health care;
- Analysis of disease incidence can provide epidemiological findings and suggest preventive measures;
- Assisting in detecting and minimizing fraud attempts in health care;
- Cooperation with pharmaceutical companies, so that it will be easier for them to identify group of relevant patients for clinical trials (assuming the patients' prior consent).

Nowadays, the trend of digitization of medical and related documents marks a time for engagement of technology called Big Data for biomedical informatics. This technology provides faster and more efficient processing and sharing of huge amount of data. Since health care involves sensitive data, the main concern is a protection of patients' private data. Many countries are implementing computerization of health care. For example, in USA a "Health Information Technology for Economic and Clinical Health Act", (HITECH) is being employed. The aim of this research is to design and define the rules that prevent the abuse and fraud concerning sensitive biomedical data, but which would not limit its efficiency and quality of output data at the same time.

2 State of the Art

The larger the quantity of heterogeneous biomedical data grouped in Big Data so that they contain as complex data set as possible, the higher benefit will they have for future processing of various analyses based on health documentation of whole population and concerning biomedical information.

Once such a project starts – and as authors of this article assume, it is not going to be a technological, but rather an organizationally ethical problem – the most effective means of processing such a large amount of data will be to provide it to the professional public as a source of research. It is quite common abroad, that when a project is fully or even only partially funded from public resources, the conditions are set so that the information is available for non-commercial activities with minimal limitations. In spite of the fact that it will never be possible to fully disclose biomedical information (with regard to sensitivity of stored data), there still exists a large array of uses, implementations and applications that would benefit from having access to them.

Due to the use of population data, there exists a risk of indirect identification of patients' information. The slightest sign of abuse brings ethical problems and may even stop the entire research.

To prevent possible issues with privacy, it is necessary to enforce very strong and efficient rules that facilitate maximum data yield, together with a strict conservation of anonymity and protection of private data. It is necessary to limit data mining so that it would not be possible to abuse or disclose sensitive data by any, even theoretical, means. Abroad, a relevant legislative is already in place, e.g. the current version of "Health Insurance Portability and Accountability Act" (HIPAA) in the USA specifies the standards concerning healthcare records transactions. Similarly, the EU legislation called Data Protection Directive 95/46/EC defines the necessity of patient consent concerning processing of his private data and portability of medical data. However, EU still does not have a unified approach to protection of private data [1].

3 Application in Biomedicine and Healthcare

The usage of Big Data in biomedicine and healthcare will always have its specifics. The amount of anonymization used on data will always be inversely proportional to the quality of output data [2]. It means, that one of the key elements of successfully using Big Data will be setting of the ratio between anonymization and the quality of mined data. Basic anonymization would be required for efficient usage. By basic anonymization we consider removal of (or in any way denying access to) private information like name and national identification number and their replacement with anonymous identifier which would identify subject across the data set. Unfortunately, data altered in such a way will still be vulnerable. Therefore it is necessary to prevent various types of possible privacy attacks.

For example, a query returning prescribed medicaments and their dosage for concrete patient contains sensitive data. From the knowledge of medicine prescribed, it is possible to infer patient diagnosis. If the private data (name, national identification number) are anonymized, then we can assume that returned data will not contain sensitive information.

On the other hand there exists a plethora of queries that do not return sensitive data. For example: Query about the amount of practitioner's patients, query about prescribed medicine in certain region, or query about specific diagnosis across the population.

3.1 Securing the Whole Database

One possibility to increase security in biomedicine and health care is to encrypt the underlying data. It adds another safety layer and thus decreases the risk of sensitive data leak or abuse [3].

There are various advanced algorithms [4] which can encrypt medical records so that only personnel with relevant authorization can decode them. Those algorithms have advantages over classical encrypting methods (symmetrical and asymmetrical ciphers) – they are faster and cheaper than traditional RSA concept, and they provide better security in case of stolen password. Authors of [5] describe querying of medical data encrypted using these algorithms. Their findings are perfectly suitable for Big Data concept.

As encryption will inevitably bring slower querying, it should not be recommended for the whole data set, but only for structured patient data.

3.2 Indirect Disclosure of Sensitive Data from Partially Anonymized Database

If there are completely non-anonymized data, or data with basic anonymization in which the base patient data (name, national identification number) were replaced, it is necessary to manage access restrictions. In such case, it is often possible to gain specific data patient, or at least data which can be inferred with a high degree of probability.

For aforementioned reasons, it is necessary to employ a solution that can limit queries, combination of which can reveal sensitive data, or only enable those combinations to personnel with higher access rights while ensuring feedback control and risk analysis of searched queries and their results.

Another query which can lead to sensitive data breach is the one that returns significantly small set of result entities.

Queries can contain combination of various factors. However, if a query is "over combined", it can, in extreme case, lead to a scenario in which only one patient is in the result set. Even if the patient's name is not stored in database, it can sometimes be inferred.

For example, if we know a fraction of health record of a given person, then using relevant information (age, sex and address) we can indirectly gain his sensitive information. Such a risk can be eliminated to some extent by employing heuristic rules and their gradual improvement. Those rules would block answers that could contain risky set of information.

3.3 Indirect Disclosure of Sensitive Data from Fully Anonymized Database

In case of anonymized data set, full access to database can be provided under specified conditions.

Before the medical records are saved to database, it is possible (or sometimes even required by relevant legislation) to anonymize the records (remove name and national identification number) and also generalize them. We call this process full anonymization. By generalization we mean making identification of a person by quasi-identifiers (eg. date of birth, address, sex) difficult or impossible. By matching quasi-identifiers, medical records the governor of Massachusetts were leaked using quasi-identifiers that were accessible in anonymized mode and matched with electoral data containing quasi-identifiers that were published together with names.

Author of [6] describe generalization algorithm as a principle that patients form groups based on their quasi-identifiers, and each group must contain at least K patients.

This approach makes it difficult to identify people, but in some cases identification is still possible in contrary to K-anonymity. Authors of [7] improve K-anonymization by proposing improvement, L-anonymization. It requires patients in the same quasi-identifier group to have heterogeneous sensitive data.

Another approach to generalization lies in rounding quasi-identifiers. Data can be stored in database in more forms, each time with different level of precision. The higher authorization level of the personnel reading data, the more precise quasi-identifiers can be accessed. For

example, instead of a birth date, only a birth year or a decade is stored. Instead of a whole address, only a city or region name is stored.

Another possibility of generalization lies in not storing certain quasi-identifiers at all.

ICD codes (International Statistical Classification of Diseases and Related Health Problems), are used abroad. They are maintained by WHO. These codes have hierarchical structure, therefore they are perfectly suitable for generalization. Authors of [8] specify probability of patient identification based on frequency of rare ICD codes. They recommend a removal of 5% to 25% of the rarest codes, and their replacement by more generic ones. This leads to significantly lower probability of patient identification with relatively low precision lost.

3.4 Non-triviality of Querying

A big problem concerning research on Big Data in biomedicine informatics concerns the creating of queries. It is not expected that a majority of researches in biomedical field would be able and willing to design their own Map/Reduce parallel algorithms to solve queries in Big Data medical database. It is more probable that there will be a cooperation from IT technicians, analysts and programmers who will create tools that could be parametrized, run on demand etc.

Both issues (safety and non-triviality) could be solved by query tool. The tool would contain query "templates", programs and algorithms that could be parametrized via user interface. The Big Data database could then be simply queried by such a tool without complex training.

We advocate using templates primarily as a means of simplification for broad research community and to increase availability of relevant research and its results for wide range of applications. The question of security would also be settled.

Template examples:

- Prescription of concrete active substance according to medical specialization;
- Frequency analysis according to place of residence;
- Demographic composition of patients;
- Volumes of medical actions of a facility by period;
- Correlation between diseases and patient's type of profession.

Multilayer architecture can be built for this need. It will be an extension of classical Big Data technologies in specific biomedical implementation. It will be split at least into those layers:

1. Professional public will be granted permission to use only prepared templates into which it would be possible to insert custom parameters, but it would not be possible to change the nature of a query. Query could utilize only one template, it would not be possible to combine templates. This way it is guaranteed that no sensitive data could leak. This access will be primarily for postgraduate students for their basic research.
2. Specialized workplaces would be allowed to combine templates and will have greater freedom in their parametrization. A set of heuristic rules will oversee the queries and will report or even block the combinations that could lead to a disclosure of sensitive information.
3. Team of analysts with security credentials will prepare templates including heuristic rules that will watch over their usage. Under standard security checks they could also perform Big Data queries. This practice could be used in cases where there will be a probability of work with sensitive data. As part of their workload, they could handle complex tasks and queries according to requests of individual workplaces in cases when it would not be efficient to use classic templates or when there would be a risk of leak of sensitive data. Resulting data would be checked and possibly anonymized before their return to requesting workplace.
4. Narrow specialized team of auditors will define and configure advanced heuristic rules and approve templates before release.
5. The last level could be based on a system with elements of artificial intelligence. It would be based on advanced pattern recognition algorithms, neural networks and learning process. It could automatically scan and detect unhandled possibilities of data abuse in real time.

4 Discussion

Research will continue in three main fields.

Firstly, it will be aimed on standard safety rules and relevant security technologies and their utilization in biomedical data. Their specifics and deviations from standard approach in data security methods will be defined. This part will concentrate specifically on analysis of encryption algorithms with regard to granted permissions, their benefits and disadvantages and also their influence on efficient data analysis.

Second research area will focus on anonymization algorithms and their influence on data yield and efficiency of processing. It will define the principles and the influence of anonymization on biomedical data with regard to theoretical possibilities of retrieving sensitive data using various query combinations. This area will not only be focused on theoretical field, but it will also try to generalize results received by combinatorial methods from representative sample of anonymized data and to compare them with real data and evaluate their similarity.

The third and largest area is to define an interface between low-level Big Data querying mechanism and query templates, where the main objective will be a balance of safety mechanisms and usability for wide professional public. Target state would be to find an interface which would be comparable in usability to MS Excel or MS Access, or their alternatives. This area would elaborate on distribution of rights and responsibilities among four defined roles, their detailed description and process mapping of their relation to data security. Each role will be analyzed for potential risks and threats including relevant countermeasures. Basic analysis of heuristic functions and elements of artificial intelligence as means of improving safety of biomedical data will be performed.

Acknowledgements

This paper has been partially supported by the SVV-2015-260158 project of Charles University in Prague.

References

[1] Boussi Rahmouni H, Solomonides T, Casassa Mont M, Shiu S. Modelling and Enforcing Privacy for Medical Data Disclosure across Europe. In Adlassnig KP, editor. Medical Informatics in a United and Healthy Europe – Proceedings of. Sarajevo: IOS Press; 2009. p. 695–699.

[2] Duncan et al. Disclosure Risk vs. Data Utility: The R–U Confidentiality map: Los Alamos National Library; 2001.

[3] Amazon Web Services. Creating Healthcare Data Applications to Promote HIPAA and HITECH Compliance. 2012.

[4] Alshehri, Radziszowski, Raj K. Designing a Secure Cloud-Based EHR System using Ciphertext-Policy Attribute-Based Encryption.

[5] Narayan S, Gagné M, Reihaneh SN. Privacy preserving EHR system using attribute-based infrastructure.

[6] Sweeney L. *k*-anonymity: a model for protecting privacy. International Journal on Uncertainty. 2002; 10(5): p. 557–570.

[7] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. *L*-diversity: Privacy beyond *k*-anonymity. ACM Transactions on Knowledge Discovery from Data. 2007 March; 1(1).

[8] Vinterbo S, L OM, S D. Hiding information by cell suppression. In Proc AMIA Symp; 2001. p. 726–730.

Paradigm Changes of Health Systems Towards Ubiquitous, Personalized Health Leads to Paradigm Changes of Security and Privacy Ecosystems

Bernd Blobel[1]

[1] Medical Faculty, University of Regensburg, Germany

Abstract

Paradigm changes regarding organizational, methodological, and technological aspects of health systems lead to paradigm changes for security, privacy, trustworthiness requirements and solutions. This is especially true for personalized, preventive, predictive, and participative health services based on Big Data and Analytics. The paper roughly defines the concepts of Big Data, Analytics, security, privacy and trust, and describes the challenges for security and privacy ecosystems when collecting and deploying massive data volumes from multiple sources in multiple formats for data-driven decision support. Traditional concepts for security and privacy are too rigid for meeting the requirements of an extremely complex, unpredictable and flexible Big Data ecosystem. Therefore, the consideration of the context of those data and the deployment scenarios as well as the establishment of ethical and fair information principles is inevitable. Context and conditions, expectations and preferences, rules and regulations must be formalized in proper policies. The paper highlights the need for appropriate architectures, infrastructures, and tools, and refers to another contribution in this volume about policy design and representation.

Keywords

Big Data, Analytics, personal data, personal health information, privacy, security, policies

Correspondence to:

Bernd Blobel
Medical Faculty, University of Regensburg
Address: c/o HL7 Deutschland e. V., An der Schanz 1, 50735 Köln, Germany
E–mail: bernd.blobel@klinik.uni-regensburg.de

IJBH 2015; 3(1):6–11
received: April 30, 2015
accepted: May 6, 2015
published: June 15, 2015

1 Introduction

Health systems around the globe are undergoing paradigm changes for improving quality and safety of patient's care and enhancing the care process efficiency and efficacy. Those paradigm changes address organizational, methodological, and technological issues. Organizationally, health systems turn from organization-centric through process-controlled to person-centric care.

Regarding the methodology applied, traditional health settings realize a phenomenological approach by addressing health problems generally. Stratifying the population for specific clinically relevant conditions, we move to evidence-based medicine for dedicated care. In the next step, systems medicine enables the multi-disciplinary understanding of the mechanisms of diseases and their therapy from elementary particle to society. Considering the individual health status of citizens, conditions and context, care gets personalized, preventive, predictive, and participative (P4), so best meeting the aforementioned objectives of improving quality and safety of patients' care and enhancing the care process efficiency and efficacy. Care delivery is provided ubiquitously, i.e. independent of time and of the location of actors involved.

Regarding the technology in hardware and software applied, mainframes and client-server architectures are replaced by distributed systems using the Internet. Mobile technologies, nano and bio technologies, knowledge representation and management, Artificial Intelligence, Big Data and Business Analytics, Cloud Computing, and social business are coming to play with increasing data capacity from KB through PB up to YB, and complex as well as highly parallel real time computing. More details on the aforementioned paradigm changes can be found, e.g., in [1].

Cooperative care also impacts the way interoperability is defined and implemented [2]. The paper investigates how the described paradigm changes and the advanced interoperability challenges modify the way, health services are provided medically, socially, legally, and ethically, especially considering security, privacy, and trust issues.

2 Methods

After defining Big Data and Business Analytics as well as security, privacy, and trust, the security and privacy challenges of Big Data and Analytics are highlighted. In that context, the special role of dynamic policies for ruling the new types of care service delivery is emphasized.

As medicine turns towards systems medicine, also health service interoperability solutions have to be analyzed, designed, and implemented using systems theory, an architecture-centric approach including the ontological representation of all domains realizing a perspective on that system and their harmonization for achieving comprehensive interoperability. As reference architectural model and framework, the Generic Component Model (GCM) is used, developed in the early nineties by the German Object Management Group (OMG) chapter and advanced at the Medical Informatics Department of the University of Magdeburg [3, 4].

3 Results

3.1 Definitions

Personalized ubiquitous health services include the individualization of diagnosis and therapy, realized independent of time and location by any type of principals (person, organization, device, application, component, object) [5]. Such services require cooperation of many different and sovereign stakeholders in a multi-disciplinary approach including comprehensive medicine, natural sciences, engineering, but also social and legal sciences. Therefore, they cannot be ruled in a centralized and predefined way, but have to be highly flexible and dynamic, adaptive, and autonomous (self-organizing). For analyzing, designing, and implementing such very complex systems, selection and abstraction based on the entire systems sciences world (systems medicine, systems biology, systems pathology, etc.) must be applied [1, 2].

Big Data are comprised of large amounts of data of a variety of media types (images, audio, video, text, parameter, measurements, etc.) and structure (structured, semi-structured, unstructured), unpredictably coming in (almost) in real time from many different sources (traditional sources, Web server logs and Internet clickstream data, social media activity reports, mobile-phone call detail records, RFID tags, smart metering, and information captured by sensors) to be linked, matched, cleansed, and transformed across systems [6]. As Big Data, characterized by its volume, variety, velocity, variability, and complexity, also contains personal information or enterprise data, which are especially vulnerable, they require specific protection [7, 8].

Big Data Analytics is the process of examining Big Data as mentioned before to uncover hidden patterns, unknown correlations, and other useful information [9]. This requires establishing a robust data infrastructure for integration, determining upfront which data are relevant, and solving security, privacy, and trust problems.

In the EuroMISE Mentor Association Mentoring Course realized by the author in the context of the presentation of this paper, more details on Big Data management, deployment, related tooling and applications, but also on values extracted from Big Data for optimizing the outcome of the healthcare ecosystem are offered [10].

Security covers concepts, services, mechanisms and data to guarantee integrity, availability, authenticity, accountability, and confidentiality of information.

Privacy is a human right for self-determination, respecting wishes and demands regarding collection, processing, communication, and use of personal information, thereby preventing harm from disclosure.

Trust defines the individual expectations in the context of the collection, processing, communication, and use of personal information. It allows acceptance of risk and balancing privacy needs against benefits. Trust can be based on a) knowledge and experiences of an entity about actors and processes involved in (personal) data management; b) regulations transparently established for ruling related actors' behavior and processes; c) legislation binding actors and enforcing processes (law enforcement), and d) services technically enforcing the processes and policies. For more information see, e.g., [11, 12]

3.2 Big Data Security, Privacy, and Trust Challenges

Big Data and Analytics require usability and trustworthiness of data. For being exploited, also the systems deployed have to be usable and trustworthy [13].

In its Study Report on Ethics for Big Data and Analytics, Mandy Chessell from IBM highlighted the importance of context, consent and choice, reasonableness, substance, ownership, fairness, consequences, access control, and accountability for legality and ethical correctness of Big Data and Analytics deployment [14].

Inconsistency between user preferences and expectations regarding the use of personal information on the one hand, and the realized use of that information on the other hand, damages the trust needed for Big Data deployment. Context-aware use of personal information, based on context-sensitive governance can restore the lost trust, so facilitating development and use of the personal data ecosystem [15]. This also includes the right level of transparency, accountability, and individuals' empowerment related to the current context [16]. Rigid policies are unable to meet changing context. Therefore, flexi-

ble and dynamic policies are necessary. In [17] published in this volume, the author demonstrates how this can be done based on international standards.

No data is more personal than individuals' medical data. Therefore, appropriate security and privacy controls have to be put in place to protect those data appropriately [18].

As Big Data architectures store and analyze very large data volumes from multiple sources usually without having security controls in place appropriate for such architectures, they are very vulnerable in an increasingly sophisticated cybercrime environment [19].

However, the risks Big Data projects are exposed to are not just technical by nature. There are also relevant methodological and behavioral aspects to be considered, as highlighted, e.g., by Buytendijk and Heiser [20]. Here, the impossibility of anonymization and data masking in Big Data environments, the careless behavior of individuals (especially in social media), the wrong interpretation of patterns as reality, the danger that data are taken for reality, and finally the ignorance of people have to be mentioned. The authors recommend a debate on that dilemma and the development of a code of conduct.

3.3 Big Data Security, Privacy, and Trust Risk Mitigation

The traditional privacy controls such as data minimization and purpose of use limitation as well as the concepts of personally identifiable information and consent on data and its use do not work for Big Data Analytics. Therefore, authors like Tene and Polonetsky (e.g. [21]) suggest adjusting the privacy frameworks properly by implementing the Fair Information Practice Principles (FIPPs) Transparency, Individual Control, Respect of Context, Security, Access, Accuracy, Focused Collection, and Accountability to compensate problems in data minimization and individual control by stronger focusing on transparency, access, and accuracy. Instead of predefining rigid policies, individuals should be encouraged to express their expectations and preferences depending on current context and environmental conditions in personal, at best formal and therefore machine-processable, policies. The process of establishing personal policies and mapping them with the legal and organizational policies of service providers should be done in an automated way. In cases of open, i.e. not individually authorized access to data, de-identification to a reasonable extent as well as robust mechanisms for ID management and authentication, and secure communication channels must be provided. User-centric or federated identity management schemes including single sign-on (SSO) capabilities are possible solutions [22]. At the same time, misusing those mechanisms for universal surveillance has to be excluded technically, legally, and culturally.

Kessler proposed the following Big Data security and privacy measures [23]:

- Secure computations in distributed programming frameworks;
- Security best practices for non-relational data stores;
- Secure data storage and transactions logs;
- End-point input validation/filtering;
- Real-time security/compliance monitoring;
- Scalable and composable privacy-preserving data mining and analytics;
- Cryptographically enforced access control and secure communication;
- Granular access control;
- Granular audits;
- Data provenance.

Green summarized the guidelines offered by the Federal Trade Commission in their Report "Protecting Consumer Privacy in an Era of Rapid Change" (preliminary report in 2010) [24] for Big Data best privacy practices to [25]:

- Privacy by design based on proper risk analysis and advanced modeling and design principles;
- Simple and easy-to-use consumer choice and opt-in based on the knowledge about who has collected which data for which use;
- Transparency;
- De-identification as a basic risk mitigation technique.

The latter is a gold standard measure despite the aforementioned problems regarding anonymization and de-identification in Big Data environments.

3.4 Health System Paradigm Change Effect on Security and Privacy Paradigms

The organizational, methodological, and technological paradigm changes for health systems both in industrialized and low- and medium income (formerly named developing) countries result in the following adoptions of business processes:

- Decentralization of business processes resulting in service orientation;
- Decentralization of development processes;
- Decentralization of security and privacy management;
- Middleware, moderator, mediator services (also for security, e.g., key splitting and sharing);
- Business intelligence;

- Automated data-driven decision support.

It is not surprising that those paradigm changes massively affect the security and privacy paradigms as follows:

- Build-in security and privacy, resulting in security and privacy by design;
- Patient empowerment;
- Separation of security and privacy management and services on the one hand, and applications on the other hand;
- Intelligent security and privacy management by security analytics, and security and privacy intelligence.

Those aspects are generically summarized with a focus on privacy and trust in [11, 12].

3.5 Advanced Security and Privacy Solutions

For realizing security and privacy by design, security and privacy must turn from an add-on aspect towards a basic perspective of the business system and its processes. To achieve this goal, the definition of the system with its components, their functions and relations, representing the system's architecture, and the description of the behavior of that system for meeting the system's business objectives and related use cases have to model from the beginning the security and privacy aspects as relevant – or when considering the acceptance of the system's solution – as foremost perspective on the system to be considered. Examples for correctly modeling and representing security and privacy architectures can be found, e.g., in [26]. To empower patients, education is one specific challenge. Another one is the aforementioned appropriate design of security and privacy services as part of the intended business solution, thereby separating the system's perspective security and privacy management and services from applications. This also includes architecture-centric, ontology-driven design and implementation of intended policies and their harmonization [27, 28]. This is both a research and development, and also a standardization challenge. In consequence, the author was also engaged in related standards such as ISO 22600 Privilege management and access control [29], or HL7 Healthcare privacy and security classification system (HCS) – Release 3 [30]. The latter includes security labeling services, privacy and protection services, and the related trust services [28].

Big players in ICT scene such as IBM, HP, SAP, Microsoft, or SAS are leadingly engaged in the final challenge of security analytics, and security and privacy intelligence. Security intelligence enables consolidation of data silos, threat detection, fraud discovery, risk assessment and management, using methods and tools of business intelligence such as [31, 32]

- distilling large amounts of information into an efficient decision-making process by reducing billions of pieces of data to a handful of action items;
- operationalizing data collection and analysis through automation and ease of use;
- delivering high-value applications that help organizations derive the most benefit from their data to understand and control risk, detect problems and prioritize remediation;
- validating that the organization has the right policies in place;
- assuring that the controls the organization has implemented are effectively enforcing those policies.

Among others, IBM has defined the outcomes of security intelligence solutions as [32]:

- Reduce costs associated with deployment and operations of security and privacy solutions by qualifying the existing staff for making security relevant to the business;
- Support access to appropriate products (including open source products), thereby enabling all organizations instead of just a few highly sophisticated organizations security capabilities;
- Facilitate deployment of a unified platform to approach acceptable security intelligence instead of multiple products requiring cumbersome and costly integration processes;
- Automate the collection, normalization and analysis of massive amounts of security data from technical and organizational silos, so providing rich context to every analysis;
- Enhance threat detection, applying context to detect possible attacks that might go unnoticed by a particular security technology;
- Improve incident response through accurate and quick detection by considering massive data from multiple resources, formats, etc.;
- Realize staffing ROI and the implementation of new security solutions by offering global services for thread monitoring, incidence reporting, knowledge sharing, etc.;
- Empower enterprises to run highly robust security programs, processing billions of records daily and producing a score of high-priority action items every 24 hours.

4 Discussion

Big Data is revolutionizing healthcare, supporting the integration of acute care, prevention, public health, elderly care, Ambient Assisted Living, etc. It supports patients in taking choice to optimize their lifestyle, in finding optimized resources, but also as tool for eLearning and better understanding. On the other hand, the subject of care is also information originator/provider [33]. In Big Data environments, this role is not limited to the patient, but frequently also includes relatives or third parties.

There are six issues relevant for protecting Big Data: a) monitoring, b) analyzing and auditing for unauthorized activity, c) analyzing and auditing for sensitive information, d) identity management, e) data masking, and f) application security [34]. From a user's perspective, privacy is frequently a priority, but cannot be provided without security [35]. Therefore, the tendency of installing in healthcare establishments a Chief Information Officer (CIO), a Chief Information Security Officers (CISO) and a Chief Privacy Officers (CPO) turns meanwhile back by combining the function of CISO and CPO, as announced, e.g., at Intel [36].

Despite all the advancements mentioned in this paper or presented in other sources, there are still open security and privacy issues to be addressed in a connected world of the Internet of Things and the aforementioned P4 care paradigm such as:

- Bring Your Own Device (BYOD) policies;
- Responsibility/accountability issues including the role of the patient and relations between health professionals, health service providers, and patients (see here also [11, 12]);
- Liability problems;
- Safety risks of mobile Health (mHealth) including the regulations for mobile devices;
- Policies.

As mentioned before several times already, Big Data privacy protection is not just a technical category. Therefore, the U.S. President's Council of Advisors on Science and Technology (PCAST) defined in its Report to the President on Big Data and Privacy: A Technological Perspective [37] the following recommendations:

- Policy attention should focus more on the actual uses of Big Data and less on its collection and analysis.
- Policies and regulation, at all levels of government, should not embed particular technological solutions, but rather should be stated in terms of intended outcomes.
- Agencies should strengthen U.S. research in privacy-related technologies and in relevant areas of social science that inform the successful application of those technologies.
- Related agencies, educational institutions, and professional societies should encourage increased education and training opportunities concerning privacy protection, including career paths for professionals.
- The U.S. should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today.

At least the latter statement does not really justify when looking at the European developments the author is actively involved in, but also when reflecting the series of security and privacy breaches and even high level scandals happening in the U.S. as well as globally.

References

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

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

[3] Blobel B. Assessment of Middleware Concepts Using a Generic Component Model. Proceedings of the Conference "Toward An Electronic Health Record Europe '97", pp. 221-228. London; 1997.

[4] Blobel B, Holena M. Comparison, Evaluation, and Possible Harmonisation of the HL7, DHE, and CORBA Middleware. Stud Health Technol Inform. 1997;45:40-47.

[5] Object Management Group. Security Service Specification V 1.7. Needham: OMG; 2001.

[6] Zikopoulos PC, Eaton C, deRoos D, Deutsch T, Lapis G. Understanding Big Data. New York: McGraw-Hill; 2012.

[7] ISACA. Privacy Big Data. Aug 2013.

[8] NetApp Inc. Big Data in Healthcare: Tackling Challenges, Pursuing Opportunities. Produced in cooperation with HIMSS media. Sunnyvale, CA: NetApp Inc.; 2013.

[9] SAS. Big Data Analytics. What it is why it matters. Cary, NC: SAS Institute Inc. http://www.sas.com/en_us/insights/analytics/big-data-analytics.html

[10] Blobel B. Big Data Modeling and Management. International Joint Meeting EuroMISE 2015, Mentoring Course BIG DATA – ANALYSIS AND MODELING CHALLENGES, June 16th, 2015, Prague, Czech Republic.

[11] 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/

[12] Ruotsalainen PS, Blobel B, Seppälä A, Nykänen P. Trust Information-Based Privacy Architecture for Ubiquitous Health. JMIR Mhealth Uhealth 2013;1(2):e23, URL: http://mhealth.jmir.org/2013/2/e23/

[13] Institute for Health Technology Transformation. Transforming Health Care Through Big Data. New York: Institute for Health Technology Transformation; 2013.

[14] Chessell M. Ethics for Big Data and Analytics. IBM Corporation; 2014.

[15] World Economic Forum. Rethinking Personal Data: Trust and Context in User-Centred Data. Geneva: World Economic Forum; May 2014.

[16] World Economic Forum. Rethinking Personal Data: A New Lense for Strengthening Trust. Geneva: World Economic Forum; May 2014.

[17] Blobel B, Ruotsalainen P, López DM, Gonzalez C. How to Use the HL7 Composite Security and Privacy Domain Analysis Model. (in this volume)

[18] Marr B. How Big Data is Changing Healthcare. Forbes April 21, 2015.

[19] Kessler A. Big Data: A Big Problem or a Big Opportunity? Vormetric Data Security Blog. http://blog.vormetric.com/2013/05/02/big-data-a-big-problem-or-big-opportunity/

[20] Buytendijk F, Heiser J. Confronting the privacy and ethical risk of Big Data. Financial Times, September 24, 2013.

[21] Tene O, Polonetsky J. Big Data for All: Privacy and User Control in the Age of Analytics. NW J Tech Intell Prop. 2013 April;11,5:240-273.

[22] Hudson S, Frazier T. ICD-MarketScape: Worldwide Federated Identity Management and Single Sign-On 2014 Vendor Assessment. Framingham, MA: IDC Corporate USA; 2014.

[23] Cloud Security Alliance. Top Ten Big Data Security and Privacy Challenges. November 2012

[24] Federal Trade Commission. Protecting Consumer Privacy in an Era of Rapid Change. March 2012. www.ftc.gov

[25] Green A. Big Data Best Privacy Practices, FTC-Style. Varonis.Blog, Sept. 5, 2013 http://blog.varonis.com/big-data-best-privacy-practices-ftc-style/

[26] Blobel B. Ontology driven health information systems architectures enable pHealth for empowered patients. Int J Med Inform. 2011;80,2:e17-e25.

[27] Blobel B, Ruotsalainen P, González C and López D. Policy-Driven Management of Personal Health Information for Enhancing Interoperability. Stud Health Technol Inform. 2014;205:463-467.

[28] Blobel B, Davis M, Ruotsalainen P. Policy Management Standards Enabling Trustworthy pHealth. Stud Health Technol Inform. 2014;200:8-21.

[29] International Organization for Standardization. ISO 22600 Health informatics – Privilege management and access control. Geneva: ISO; 2006.

[30] HL7 International Inc. HL7 Healthcare Privacy and Security Classification System (HCS) – Release 3. Ann Arbor: HL7 International; May 2013

[31] Q1 Labs. IT Executive Guide. Waltham, USA: Q1 Labs; 2011.

[32] IBM Software. IT executive guide to security intelligence. Somers: IBM Corporation; Jan 2013.

[33] Yamamoto K, Ishikawa K, Miyaji M, Nakamura Y, Nishi S, Sasaki T, Tsuji K, and Watanabe R. The Awareness of Security Issues among Hospitals in Japan. Caring for Health Information Safety, Security and Secrecy. Heemskerk, The Netherlands, November 13-16, 1993.

[34] Masters Emison J. 6 Tools to Protect Big Data. InformationWeek::reports; June 2014. (Report ID: S7940614)

[35] IBM Software. Leverage security intelligence to protect sensitive healthcare data. Dec 2013.

[36] Chabrow E. CISO as Chief Privacy Officer. BankInfoSecurity, Apr 2, 2013. http://www.bankinfosecurity.com/interviews/ciso-as-chief-privacy-officer-i-1875

[37] Executive Office of the President. President's Council of Advisors on Science and Technology (PCAST) defines in its Report to the President on Big Data and Privacy: A Technological Perspective. Washington, May 2014.

How to Use the HL7 Composite Security and Privacy Domain Analysis Model

Bernd Blobel[1,2], **Pekka Ruotsalainen**[3], **Diego M. Lopez**[2,4], **Carolina Gonzalez**[2,4]

[1] Medical Faculty, University of Regensburg, Germany

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

[3] National Institute for Health and Welfare (THL), Helsinki, Finland

[4] Electronics and Telecommunications Faculty, University of Cauca, Colombia

Abstract

For facilitating its understanding and use, the HL7 Composite Security and Privacy Domain Analysis Model (CSPDAM) has been considered from an architecture-centric, ontology-driven perspective. The decomposition of the CSPDAM system into domain-specific subsystems and the decomposition of those subsystems into components at different granularity level have been performed, based on the Generic Component Model (GCM), and represented using standardized ontologies. Using GCM principles, the approach enables the aggregation of components at the right granularity level, thereby qualifying the associations between components within and across domains, finally also characterizing the transformation between the Reference Model of Open Distributed Processing (RM-ODP) views. The resulting outcome itself is meanwhile documented in international standards authored by team members.

Keywords

Clinical informatics, privacy, security, modeling, ontologies, HL7

Correspondence to:

Bernd Blobel
Medical Faculty, University of Regensburg
Address: c/o HL7 Deutschland e. V., An der Schanz 1, 50735 Köln, Germany
E–mail: bernd.blobel@klinik.uni-regensburg.de

IJBH 2015; 3(1):12–17
received: April 30, 2015
accepted: May 6, 2015
published: June 15, 2015

1 Introduction

Health Level 7 Inc. was established in the United States in 1987 in order to develop standards and protocols to facilitate communication between different components of hospital information systems. Meanwhile, mission and scope have moved beyond hospitals (including now, e.g., primary care) and even beyond direct care to cover also public health, clinical studies, knowledge representation and management, biomedical devices, and recently also mobile health (mHealth, i.e. the practice of medicine and public health supported by mobile technologies) without defining a final limitation. Furthermore, the organization turned from a national body to an international institution with currently 38 affiliate countries connected to HL7 International Inc. [1]. In order to provide both a consistent set of requirements and principles for the evolving domains covered by HL7, and to enable interoperability between different domains, HL7 Domain Analysis Models are under development. Domains are areas of activities and/or interests (e.g. disciplines dealing with specific aspects of the healthcare system) usually defined and managed by domain experts using their own terminologies derived from specific domain ontologies, and deploying their specific skills and methodologies. This paper considers the HL7 Composite Security and Privacy Domain Analysis Model (CSPDAM) [2].

In general, there are two ways to achieve semantic interoperability: a) the alignment of domains and solutions based on enforced basic concept reference models, and b) the harmonization between different concepts and their representation style, i.e. ontology mapping. If the models belong to different phases of the development process expressed by different views of ISO 10746 “Information technology - Reference Model Open Distributed Processing (RM-ODP)” [3], transformation between them is inevitable. For inter-domain interoperability reasons, HL7 binds his developers to the use of the HL7 Reference Information Model (RIM). It represents the information view on an information and communication technology (ICT)

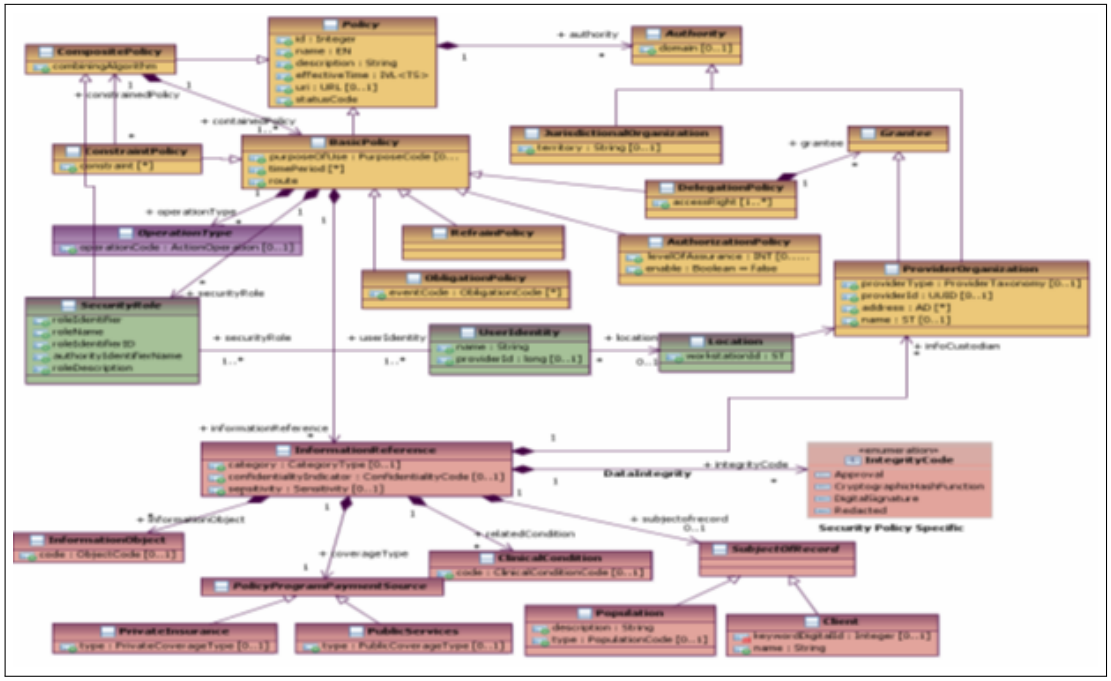


Figure 1: HL7 Composite Security and Privacy Domain Analysis Model [2].

centric system. Despite being specific for HL7, the RIM as an ICT ontology is usually inappropriate to derive ICT independent real world domain models. The technology dependency of different approaches does not guarantee interoperability even when they rely on the RIM [4]. On the other hand, justified real world domain models as general and technology independent reference can be easily transformed into different information models. As important legal, ethical, social, organizational and resulting technical aspects of privacy (and partially also security) – comprehensively defined as policy [5] - are defined and managed outside the healthcare domain, the CSPDAM should be based on those domains' ontologies instead of being bound to the healthcare-specific, ICT-centric RIM. However, this statement holds generally for real world domains, as demonstrated for clinical models [6].

2 Methods

The CSPDAM aims at providing "a harmonized analysis of security and privacy system requirements of healthcare organizations ... by identifying the information and the system behaviors required to implement technological controls enforcing healthcare security and privacy policies" [2]. Regarding the two basic concepts communication security/privacy and application security/privacy [5], tackling the application security/privacy concept is specific to the healthcare domain and therefore the major challenge. Contrary, the communication security/privacy concept is quite domain-independent and can therefore be managed by existing solutions borrowed from other domains. To guarantee that the provided model conforms to the real world system, the CSPDAM (see Figure 1) has to follow both a framework compliant with that system's architecture and established good modeling principles [7]. As architectural framework, the Generic Component Model (GCM), widely and successfully deployed in international standards and projects, has been applied, summarized, e.g., in [8]. The good modeling principles have been exemplified in the context of an architecture based and ontology driven approach to interoperability using the GCM, e.g., in [9].

The GCM simplifies the representation of a system by separating the domains (thereby forming subsystems) which tackle specific aspects of that system. It provides the architectural composition/decomposition of the system's components. Finally, it formalizes the developments process according to the RM-ODP [3]. All models are represented using the Unified Modeling Language UML [10]. As domains are represented by domain ontologies, the latter can be used to distinguish domains. The business view representation of the GCM is provided by the Basic Formal Ontology (BFO) [11, 12], while the domain-specific instantiation is realized through BFO-compliant, reused and partially adapted domain ontologies. For developing and exploiting the HL7 CSPDAM, good modeling principles have been and have to be applied, such as: a) For reflecting the stakeholders view on the system in question, the key concepts and key relations have to be captured first, considering a high level of abstraction; b) Thereafter, different levels of abstraction should be used iteratively, where the first iteration is performed in a top-down manner to provide the model's conceptual integrity [7].

The challenge to be met is the architectural description of the system in question regarding its structure and be-

havior, i.e. the system's components, their functions and interrelationships, comprehensively. Thereby, the representation of the dynamic behavior describing the system's function is more challenging than that of the static one describing its structure. The architectural representation of that system is developed for each domain and thereafter combined mapping those domains. Two basic principles have to be followed:

- Reduction of the complexity of the system by including only components relevant in the considered business case. This implies to refer to other components as environment interacting with the system in question as well as limiting the granularity considered to the necessary level.
- Simplification of the system by including only the domains relevant in the considered business case.

3 Results

First, the domains relevant in the security and privacy business case have been defined (Figure 2). The architectural decomposition of the information security management system according to the GCM generic granularity levels Relations Networks, Aggregation, and Details is shown in Table 1.

Table 1: Security Concepts represented at GCM.

GCM Granularity Level	Security Concepts
Business Concepts	Information security management
Relations Networks	Directory services; ID management; certification management; policy management; naming services; key management; ...
Aggregations (Services)	Identification; authentication; integrity check; non-repudiation; security logging; signature service; ...
Details	Enabling/disenabling access, delegation; encoding/decoding; signing; ...

From a technology-independent perspective, the security system can be specialized into two subsystems: security services management and policy representation, exemplifying the administration and policy domain, respectively, and represented by the related domain ontologies. The policy system, administered by the administration system, rules the connection (access) to the medical documentation system (e.g.an EHR system). The involved domains are shown in Figure 2. With the GCM, some basic principles have been defined to be strictly followed:

- The architectural dimension of system decomposition represents components specializations. In other words: The vertical relations represent specializations, the horizontal ones within a view represent aggregations.
- The domain dimension separates system perspectives/aspects. Only neighbored components can get connected. In other words: Relations can only be established at the same granularity level. There are only horizontal relations between domains performed as mapping between concepts representing the components in question.

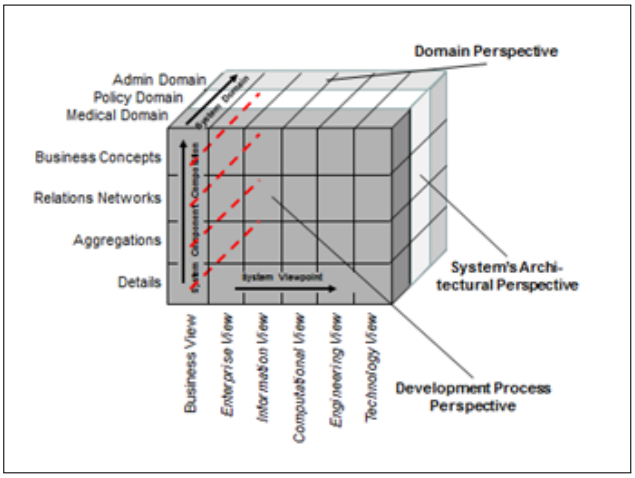


Figure 2: The CSPDAM representation at GCM.

For representing the medical domain through comprehensive documentation established in Electronic Health Records (EHRs), several approaches have been developed such as ISO EN 13606-1 Electronic Health Record Communication, Part 1 – Reference Model [13]. In this context it should be mentioned that such reference model is suited to represent EHR systems, but not medical knowledge as it is represented by the SNOMED-CT ontology. Here, we would like to refer to [6].

Due to legal, cultural and social impacts as well as jurisdictional dependencies, it is almost impossible to define a globally valid reference model for the administrative subsystem.

Regarding the policy domain, there is a reference model and high level ontology representing the policy subsystem: ISO 22600 "Privilege management and access control Part 2 – Formal models" [14]. Figure 3 shows the Policy Base Class Diagram of that standard, which has been substantially borrowed from the PONDER policy language specification [15]. The ontology stuff is explained later on in some more details.

The Policy Base Class Diagram contains three major parts of the policy model: BasicPolicy, MetaPolicy, and Composite-Policy. MetaPolicy is needed when describing systems from a meta-model perspective, while the CompositePolicy branch supports a specific mapping of

policies. The BasicPolicy is used as parent class for deriving any type of policy instance (in the sense of individuals of the aforementioned BFO specification). The use cases considered in the CSPDAM context resulted in established or currently developed security and privacy specifications such as the HL7 RBAC Catalog (Role Based Access Control Catalog), eConsent Management, HL7 Healthcare Privacy Security Classification System, and HL7 Consent Directive.

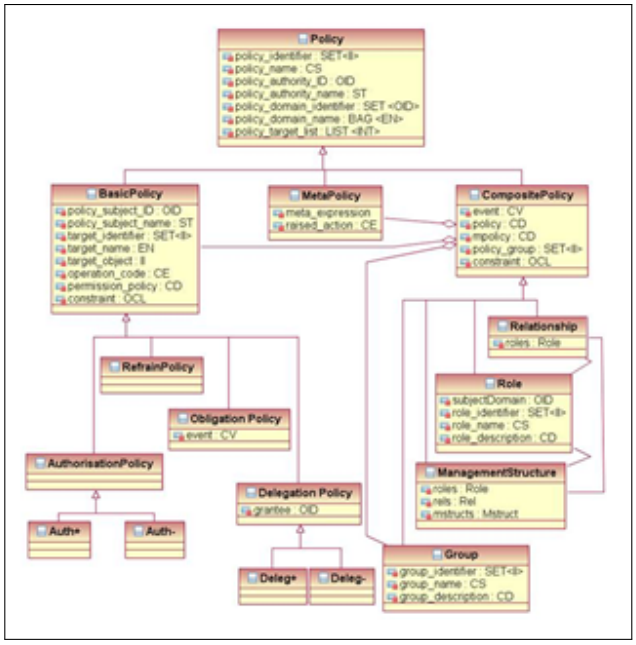


Figure 3: ISO 22600 policy model [14].

As security and privacy management aims at controlling the structurally and functionally binding of entities (principals according to the Object Management Group (OMG) definition [16]) to targets, i.e. the relations between components of the administrative and the medical domain, policy components act in the simplest case as association classes between those components constraining those interrelations. However, policy can also constrain the entities themselves (Figure 4).

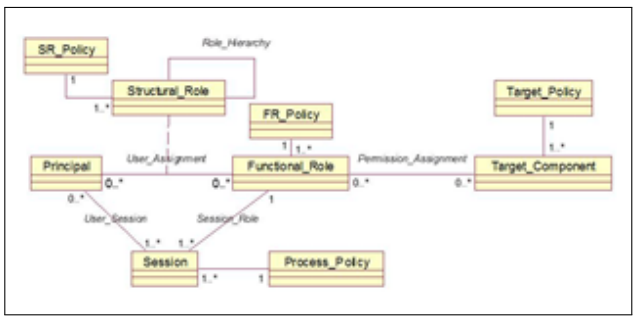


Figure 4: Policy-driven RBAC [14].

Therefore, the policy subsystem forms the core of the HL7 CSPDAM. The architectural refinement of policies, thereby deriving special instances, must be performed within the related domain. Policies can be expressed explicitly or implicitly. In the first case, detailed representation of all policies represented in the ISO 22600 policy ontology base classes' model is applied using (frequently XML-based) policy languages. In the latter case, attributes and classification codes are deployed, resulting in a more coarse-grained specification of policies as realized in the HL7 RBAC approach. The HL7 RBAC specification deploys the CompositePolicy class which defines a group of policies (authorization, obligation, delegation, and refrain policies) to derive the abstract SecurityRole class. The SecurityRole class is specialized into functional and structural roles according to ISO 21298 [17], representing process-related and institutional policies. Those roles, but also processes and targets, can be mapped to subject of care policies as well as legal and institutional regulations. Specializing the SecurityRole to FunctionalRole and StructuralRole and binding the SecurityRole specific policy to permissions instantiated in the HL7 PermissionCatalog, the HL7 RBAC Catalog use case can be easily derived from the CSPDAM.

Despite being quite granular, the HL7 RBAC Catalog practically exemplifies the implicit policy representation. While such solution is acceptable in predefined environments offering an intermediate solution on the way to fully implement ISO 22600, the move to personalized pervasive health services requires an open, adaptive and ontology-driven approach [18]. In that context it should be mentioned that the domain-specific representation of a real-world system's architecture, i.e. its components, their functions und relations, establishes the domains ontology at the considered level of granularity as well as within the chosen focus forming a partial ontology. So, the GCM-compliant ISO 22600 Base Class Diagram represents the basic concepts of a policy ontology.

Within the GCM policy domain, the HL7 ConsentDirective can be derived similarly by specializing Policy to PrivacyPolicy and interpreting BasicPolicy, which just reflects RefrainPolicy and ObligationPolicy, as PrivacyRule. The PrivacyPolicy is constrained to a PrivacyDirective. Instead of applying the explicit policy representation and binding offered by the CSPDAM when deploying an architectural framework like the GCM, implicit policies based on coarse-grained attributes and classifiers have been used here, thereby also problematically merging with the ICT ontology of the RIM.

Other use cases specified at HL7 in the same way are, e.g., ePrivacy and eConsent [19]. Also here, the only partially expressed CSPDAM ontology is merged with the ICT ontology of the RIM.

4 Discussion

The CSPDAM overview model presents a structural diagram, which has been use-case-specifically refined by attributing concepts instead of refining them from ontology models such as the ISO 22600 policy ontology [14].

Also in the RIM-based information view developed to implement ePrivacy and eConsent by representing these concepts derived from the HL7 Act class, some basic actions have been attributed. However, this approach bears the problem that information reference models such as the HL7 RIM or the openEHR Object and Archetype Models do not properly reflect real world systems ontologies. So, CSPDAM is directly applicable in static environments where all constraints have been predefined and expressed in single attributes. Because security and privacy management is especially determined by contextual conditions, behavioral aspects have to be added due to impacts on activity nodes and action modifications.

Contrary, the system-oriented framework according to Figure 2 presents the architecture of the system in question, i.e. the system's components, their functions and relations for all subsystems needed to include all relevant domains. Additionally to the structural aspect, this consideration addresses behavioral aspects at components as well as at relations level, thereby formally and explicitly representing the applicable rules, so verifying Aristotle's statement that a system is more than the sum of its parts. So, the presented approach enables to qualify all elements of the CSPDAM gaining the maximal benefit for the derived use-case-specific models. By that way, simplified intermediate solutions as well as matured specifications considering all contextual information in detailed manner up to the level of intelligent, i.e. adaptive and self-organizing systems can be designed and implemented. Alternatively to the CompositePolicy harmonization approach, general harmonization algorithms enabling merging, aligning, matching or mapping ontologies can be deployed, as described, e.g., by Uribe et al. [20].

The Security and Privacy Ontology currently developed at HL7 for representing the referenced and future use cases and solutions for security and privacy systems borrows from GCM-compliant policy ontology of ISO 22600 and the principles presented in this paper. However, regarding the other domains as well as the inevitable BFO principles, there is still a lot of work to be done.

5 Conclusion

The HL7 Composite Security and Privacy Domain Analysis Model has been designed on the basis of a system-theoretical, architecture-centric framework, the Generic Component Model, and its Security and Privacy related instantiation, the ISO 22600 "Privilege management and access control" formal model. The representation of such real world system architecture model for an entire domain additionally provides a related domain ontology – here the policy domain ontology.

Contrary to several other HL7 Domain Analysis Models, the CSPDAM provides an analysis of security and privacy system requirements of healthcare organizations by identifying information and system behaviors required to implement technological controls enforcing healthcare security and privacy policies independent of the HL7 RIM. By that way, it enables to integrate real world systems from the non-HL7 world which dominates security and privacy specifications and solutions. Furthermore, it supports interoperability between related systems and solutions from other domains such as eGovernment, eLearning, eCommerce, eProcurement, etc. The HL7 CSPDAM allows to derive and to harmonize existing and emerging HL7 specifications on security and privacy related areas. Especially in the context of pre-defined, static conditions, the development of information and computational models as well as their further transformation into the RM-ODP engineering view has been successfully demonstrated. To take full advantage from this Draft Standard for Trial Use (DSTU) and to respond to analysis and design challenges in a future-proof way, an architectural framework such as the Generic Component Model must be deployed. Going back to the GCM or deploying UML 2.1 and higher also enables the dynamic representation of policies and policy management, thereby reflecting contextual changes usually happening.

Based on the GCM, the separation of concerns and the compositional formalization eases the development of appropriate models as well as further evolution of, and harmonization with, related specifications and their underlying models.

Acknowledgement

The authors are indebted to thank their colleagues from HL7 International, ISO TC 215 "Health informatics", CEN TC 251 "Health informatics", and IHTSDO for kind support and intensive cooperation. Especially, they thank Mathias Brochhausen, University of Arkansas of Medical Sciences, Little Rock, as well as their US colleagues Mike Davis, Veterans Administration, San Diego, John Moerke, General Electric Company, Menomonee Falls, Ioana Singureanu, Eversolve LLC, Windham, and Kathleen Corner, Veterans Administration, San Diego.

Conflicts of interest

There are no conflicts of interests.

References

[1] Health Level 7 International Inc. www.hl7.org

[2] Health Level 7 International Inc. HL7 Composite Security and Privacy Domain Analysis Model. Ann Arbor, MI, USA: HL7; 2009. (DSTU: 2012) www.hl7.org

[3] International Standards Organization. ISO IEC 10746 Information technology – Reference Model – Open Distributed Processing. Geneva, Switzerland: ISO; 1996. www.iso.org

[4] Smits M, Kramer E, Harthoorn M, Cornet R. A comparison of two Detailed Clinical Model representations: FHIR and CDA. European Journal for Biomedical Informatics 2015;11,2:en7-en17.

[5] Blobel B and Roger-France F. A systematic approach for analysis and design of secure health information systems. Int J Med Inf. 2000;62:51-78.

[6] Blobel B, Goossen W and Brochhausen M. Clinical Modeling – A Critical Analysis. Int J Med Inf. 2014;83,1:57-69.

[7] Lankhorst M et al. Enterprise Architecture at Work. The Enterprise Engineering Series, pp 121 ff. Berlin, Heidelberg: Springer Verlag, 2009.

[8] Blobel B. Architectural Approach to eHealth for Enabling Paradigm Changes in Health. Methods Inf Med. 2010;49:123-134.

[9] Blobel B. Standards and Solutions for Architecture Based, Ontology Driven and Individualized Pervasive Health. Stud Health Technol Inform. 2012;177:147-157.

[10] Object Management Group Inc. Unified Modeling Language (UML) 2.1. Framingham, USA: OMG; 2007. www.omg.org

[11] Arp R and Smith B. Function, Role and Disposition in Basic Formal Ontology. Nature Precedings: hdl:10101/npre.2008.1941.1 : Posted 2 Jun 2008

[12] Basic Formal Ontology 2 (BFO2). http://code.google.com/p/bfo/

[13] International Standards Organization. ISO EN 13606 Health informatics – Electronic Health Record Communication, Part 1, Reference Model. Geneva, Switzerland: ISO; 2008. www.iso.org

[14] International Standards Organization. ISO EN 22600 Health informatics – Privilege management and access control, Part 2, Formal Models. Geneva, Switzerland: ISO; 2006. www.iso.org

[15] Damianou N, Dulay N, Lupu E, Sloman M (2000) Ponder: A Language for Specifying Security and Management Policies for Distributed Systems. The Language Specification, Version 2.3. Imperial College Research Report DoC 2000/1. 20 October, 2000.

[16] Object Management Group Inc. The CORBA Security Specification. Framingham, USA: OMG; 1995, 1997.

[17] International Standards Organization. ISO EN 21298 Health informatics – Structural and functional roles. Geneva, Switzerland: ISO; 2008. www.iso.org

[18] Blobel B. Ontology driven health information systems architectures enable pHealth for empowered patients. Int J Med Inf. 2011;80:e17-e25.

[19] Health Level 7 International Inc., Ann Arbor, MI, USA. www.hl7.org

[20] Uribe GA, López DM, Blobel B. Architectural analysis of clinical ontologies for pHealth interoperability. Stud Health Technol Inform. 2012;177:176-82.

Opportunities and Challenges of Big Data

William Edward Hammond[1]

[1] Duke University, Duke Center for Health Informatics, Durham, USA

Correspondence to:

William Edward Hammond
Duke University, Duke Center for Health Informatics
Address: Hock Plaza, 2424 Erwin Road, Durham NC 27205, USA
E–mail: william.hammond@duke.edu

IJBH 2015; 3(1):18
received: April 24, 2015
accepted: May 6, 2015
published: June 15, 2015

Big Data is the new phenomenon enable by technology. This topic has generated many opportunities to acquire new knowledge and increase our ability to discover more precise information about disease and sub-disease and outcomes as a function of details of an individual. The timing of Big Data leads the announcement of U.S. President Obama on the Precision Medicine Initiative. Big Data provides the detailed and individualized data to learn individualized outcomes. Pragmatic clinical trials across large amounts of data far expand the numbers of individuals contributing to clinical trials. Patients at the beginning of a disease can be matched with groups of similar patients who are further along the disease path to help understand and predict the course of the disease.

Big Data has been defined as having 3 major characteristics: volume, velocity, and variety. The inclusion of behavioral, genomic, environmental, societal, and economic data with clinical data result in increased volume as well as variety. Variety also includes many new forms of data such as images, videos, geospatial, patient reported, and others. Wearable sensors overwhelmingly create big data with great velocity and volume. The challenge becomes how to analyze this data in acceptable time frames as well as adequately filter the data for human consumption.

Challenges include how to establish quality and noise reduction sufficient to validate studies derived from big data. Completeness and consistency of data are challenges. What will be the impact of the lack of a global common data model with consistent meaning, units and other attributes? How do you discover new knowledge? Clearly, spreadsheets are inadequate to see patterns. Finally, misinterpreting correlations of data across large data creates some interesting spurious correlations.

Structural Equation Modelling (SEM) in Connection to Big Data

Arie Hasman[1]

[1] Department of Medical Informatics, AMC-University of Amsterdam, Amsterdam, The Netherlands

Correspondence to:

Arie Hasman
Department of Medical Informatics,
AMC-University of Amsterdam
Address: Amsterdam, The Netherlands
E–mail: a.hasman@amc.uva.nl

IJBH 2015; 3(1):19
received: May 6, 2015
accepted: May 18, 2015
published: June 15, 2015

Structural Equation Modelling (SEM) is a statistical method for the quantification and testing of theories. SEM is used in many disciplines.

On the basis of a theory a model is described in terms of interrelationships between dependent and independent variables, some of which may be latent. A latent variable is a theoretical or hypothetical construct of major importance in many sciences. These latent variables are also used in Technology Acceptance models (TAM) [1, 2]. Latent variables cannot be directly measured. Examples of latent variables are intelligence, anxiety, math ability and in TAM for example Perceived Usefulness and Perceived Ease of Use. Besides the latent variables the model also contains observed variables which are the variables that are actually measured or recorded on a sample of subjects, such as manifested performance on a particular test or the answers to items or questions in a questionnaire.

Latent variables can only be estimated via proxies, using specifically developed measuring instruments, such as tests, questionnaires, scales, etc. For example, the latent variable socioeconomic status may be considered to be measured in terms of income level, years of education, bank savings, type of occupation.

With the help of a SEM analysis the extent to which a theoretical model is supported by sample data can be determined [3]. The elements of a covariance or correlation matrix of all pairs of observed variables are expressed in terms of the model parameters. These parameters are then incrementally updated in such a way that the covariance matrix fits as closely as possible the covariance matrix as determined directly from the values of the observed variables. The closer the fit between the two matrices, the better the model.

SEM can be used to quantify and test plausibility of hypothetical assertions about potential interrelationships among the constructs as well as their relationships to measures assessing them. Because the models can be quite complex, the sample should be quite large. Therefore SEM can be fruitfully used when enough data is available. Although an available theory can directly lead to the construction of a model, SEM can also be used for theory development purposes. In theory development, repeated applications of SEM are carried out in order to explore potential relationships between variables of interest. Theory development assumes that no prior theory exists—or that one is available only in a rudimentary form—about a phenomenon under investigation.

Due to the mathematical complexities of estimating and testing these relationships and assertions, computer software is a must in applications of SEM. To date, numerous programs are available for conducting SEM analyses.

In this presentation SEM will be explained at an introductory level. The idea is to give the listener an idea of the notions behind SEM. No attention will be given to the relevant computer software.

References

[1] Hasman A. Introduction to technology acceptance, IJBH 2014, 2(2):19-27

[2] Holden, RJ and Karsh, B-T. The technology acceptance model: its past and its future in health care. J. Biomed Inform 2009, 43(1):159-72

[3] Ulmann JB. Structural Equation Modeling: Reviewing the Basics and Moving Forward, Journal of Personality Assessment 2006, 87(1):35–50

An Open Bioinformatics Platform for Personalized Treatment of Cancer

Natalia Jiménez-Lozano[1], Tomás Hlavsa[2], Benoît Pelletier[1], François Exertier[1], Víctor de la Torre[3]

[1] Big Data and Security, Atos France

[2] Big Data and Security, Atos IT solutions and services, Prague, Czech Republic

[3] Fundación Centro Nacional de Investigaciones Oncológicas (CNIO), Spain

Abstract

Currently there is an explosion of data and bioinformatics methods that is mostly enclosed in the research environment. Therefore, there is a need of transparent and trustworthy predictive tools whose exploitation in the clinic can be supported by sustainable business models and validated in multiple conditions. In this paper we describe the development and further testing of a system that will catalyse of the translation of bioinformatics methods to the clinical environment to improve the efficacy of treatment in patients with cancer.

This aim will be achieved by building an Open Bioinformatics Platform based on existing solid industrial-grade computational infrastructure able to incorporate bioinformatics methods in a flexible way. The overarching strategy of this Platform is to enable the exploitation of the vast amount of –omics, phenotypic and treatments data together with the most advanced bioinformatics methods by developing an innovative software platform to support the clinicians at work.

Correspondence to:

Natalia Jiménez-Lozano
Big Data and Security, Atos France
Address: Les Clayes-Sous-Bois, France
E–mail: natalia.jimenez@atos.net

IJBH 2015; 3(1):20–23
received: April 30, 2015
accepted: May 18, 2015
published: June 15, 2015

1 Introduction

Cancer represents one of the main causes of death in our society. There is an obvious need for more specific treatments in the market. Recent discoveries emphasize on the increase in effectiveness of drugs that target distinct types of cancer, and are in the process of being tested in ongoing medical trials [1]. However discovering the right treatment for each patient holds much room for improvement [2]. Nowadays, patients are classified based on their clinical history and tissue characteristics, while the molecular aberrations presented in the tumour are not often considered in the routine.

To change this situation several hospitals, research institutions and companies have turned their focus towards personalized cancer treatments and the results are promising [3]. Turning such a strategy into routine requires the assembly of an analysis platform that consumes patient´s and public data, transforming this information into a set of candidate treatments [4]. In hospitals clinicians have started to perform exome sequencing even if, for practical reasons, only the results for a small set of well characterized cancer genes and mutations are reported to the clinicians. While the technical aspects of genome analysis (sample handling, sequencing, etc.) are already part of the standard health system, the big challenge is the bioinformatics analysis of the genomic data. Right now the analysis of genomic information is performed with academic/experimental/prototypes systems that are highly heterogeneous and can only be implemented in large hospitals that can afford specialized personnel and technical support. According to a report created by the EU-funded BioMedBridges consortium: "One of the problems hampering the broader use of the personalised medicine data is the lack of suitable and interoperable IT solutions for structuring, organising, managing, using, and querying/analysing data." (`http://goo.gl/6E2sQi`).

The history of the technologies used in medical environments, for example with the introduction of all the imaging systems including equipment, software and support engineers, demonstrates that the full deployment of new systems can only be achieved by companies able to sell and support complete solutions to the health systems. The reasons are in part technical, related with the need

of implementing industrial software with industrial standards but also with the need of providing technical support to the system and training to the users. Large IT providers also have the real capacity to install complex software systems in hospitals including the introduction in the public health purchasing systems.

2 Gap in the Application of Bioinformatics New Discovered Knowledge into the Clinic

Bioinformatics has dramatically changed the way healthcare is operating. The transition between the "one problem one gene" to the "-omic era" has opened a wide range of new opportunities to understand, modify and improve biological systems, including derailed systems leading to human diseases.

The so-called "data deluge" has raised major challenges that should be overcome in order to provide real value to the patients. The recent trend is a change in the character of computational tools, which lately is moving from 'descriptive' to 'predictive'.

The bioinformatics community is producing every year a substantial amount of new cancer analysis methods and databases [5]. Many of those can be potentially applied to improve the diagnosis and to propose personalised treatments to cancer patients. The FDA does have a "compassionate use" exemption that permits access to experimental therapies for cancer patients—provided they have first failed so-called "proven" therapies. However, regrettably, there are still many factors delaying the application of this knowledge in the clinic [3].

One of these factors is the disconnection between the hospitals' environment where complex commercial IT solutions predominate, and the research environment where software and methods are developed by bioinformaticians. The inherent innovative and exploratory nature of these methods delivers software at the prototype level that cannot yet be used in a "production" environment. Moreover, for many of the SMEs that are offering recommendations for alternative treatments the major challenge is to get access to real data coming from patients. Confidentiality and data security issues also impose a barrier that is hard to overcome in many cases.

For genomic information to be widely used in hospitals in the future, it is essential to provide a solution that speeds up the transfer of the academic technology currently in use (e.g. the European Research Infrastructure ELIXIR is creating a wide range of services for genomic analysis), to solid, robust and standardized commercial platforms.

3 Filling the Gap: Open Bioinformatics Platform

The challenges described in the previous section highlight the need of a new business model to guarantee and accelerate the translation of bioinformatics solutions into the clinic. We propose an innovative solution to this problem, by developing an Open Bioinformatics Platform that will harmonize the way bioinformatics reaches the hospitals. This effort will be aligned with those initiatives emerging from the scientific domain. New research infrastructures such as ELIXIR will definitely contribute to make all the available public data more interoperable and will help to increase the quality of the methods developed. Our Platform is adopting ELIXIR's standards and best practices.

This paper describes a project where academic institutions and the industry have joined forces to develop new therapies based on the knowledge extracted from the vast amount of data that is now available.

The Open Bioinformatics Platform will meet the industrial standards that will support clinicians in choosing personalized treatments for patients with different types of cancer. The software solution will gather standard genomics pipelines to analyse the patients' data and will offer an innovative mechanism to plug-and-play different bioinformatics methods to model and propose personalized drug treatments as they emerge from the academic and the SME research.

Our solution will be the platform of choice to plug-in new tools for new approaches as they are developed because we will provide a "knowledge commons" through a one-size-fits-all data and knowledge resource.

We will combine the expertise of two of the worldwide leaders in IT for the development of the interoperable data layer of the platform. On top of this layer some of the best groups in the field of cancer research and modelling will develop and deploy a competitive set of bioinformatics methods to discover personalized cancer treatments as a first proof of concept of the Platform. However, any scientifically tested pipeline will be suitable to enlarge the set of services provided with this Platform. The predictions will be validated in animal models and a group of clinicians will continuously provide feedback to the software development and the scientific teams to guarantee the usability of the final solution.

4 Open Bioinformatics Platform inside

The Open Bioinformatics Platform will be used to design personalized treatments for patients diagnosed with different types of cancers. The platform will be formed by three main layers: 1) the interoperable data layer to be developed following the most strict industrial requirements and that will be compatible with existing IT in-

frastructures available today at different hospitals; 2) the modelling layer that will allow the execution of multiple methods developed by the academic and SME partners in the consortium and; 3) a visualization layer designed to inform clinicians and patients about the treatments proposed and their implications.

The bioinformatics methods developed will use the vast amount of data handled by the data layer.

This approach and computational predictions will be assessed and validated on animal models and the platform's prototype will be further tested with the aim of commercialisation through a roadmap.

Our platform will allow incorporating future bioinformatics tools developed by academy or by SMEs and will build on the interoperability and international standards (in particular those developed through the pan-European project ELIXIR) with the objective to maintain a global and sustainable platform.

Three initial models will be considered to exploit the Platform: 1) To sell the platform together with the hardware required to run the software as a single package, together with a service/support contract model, 2) To deliver the platform via a software as a service model (SaaS), charging a flat "membership" together with a per patient fee, and 3) a mixed/consulting model, where the platform is custom-tailored to the requirements of the client hospital, and deployed and managed there by dedicated specialists.

In any case, this effort will be aligned with those initiatives emerging from the scientific domain. New research infrastructures such as ELIXIR will definitely contribute to make all the available public data more interoperable and will help to increase the quality of the methods developed adopting ELIXIR's standards and best practices

The objectives of Platform are to:

1. Develop an open, fully interoperable platform to integrate patient´s data, including genomics, epigenetic, structural, and clinical records.

2. Develop a new generation of bioinformatics methods to integrate these areas of -omics/molecular information and to predict personalized cancer treatments.

3. Validate the platform's prototype by testing the treatments proposed in animal models and in retrospective cohorts of cancer patients.

4. Develop and further test a business model for the operation of the platform and its sustainability.

5 Discussion

The Platform described in this paper tackles the specific challenge of facilitating the access to the vast amount of -omics data that is currently available worldwide but distributed across several providers. We also address the need for a new generation of bioinformatics methods that take advantage of this multi-layered data proposing new levels of integration and combined predictions.

According to a report produced by the Personalized Medicine Coalition (`http://goo.gl/BhRVuw`) there are 41 cancer drugs in which the effectiveness varies depending on a known variation in a gene. These drugs as a whole are indicated to treat 10 different types of cancers. The portal cancer.net advises patients how to ask doctors for a personalized cancer treatment (`http://goo.gl/9z92hp`). Several SMEs have started to recommend those treatments and to execute the genetic tests; however in Europe there are not yet clear guidelines about how to make the experiments or to analyse the results. This situation creates an immense opportunity to standardize the procedures and hence produce better outcomes for the patients. The work presented in this paper promises to improve the current state of the art by providing an Open Bioinformatics Platform that, rooted in the hospitals, will enable SMEs and academia to use standards protocols to deliver their results to the patients.

The ambition of this project is to accelerate the design, the development and the implementation of new bioinformatics methods to produce personalized cancer treatments for the patients. The Platform will contribute to reduce the time to market for those new methodologies that continuously emerge from the myriad of research projects that are currently active in the field. To team involved in this work is made of leaders in the field of management of IT infrastructures in the hospital environment, development of bioinformatics methods for cancer research and coordination of large-scale cohort analyses in patients with different types of cancer. It is a rare opportunity to bring together several big industry players and excellent academic partners to work on cancer. The coordinated synergy between these expertise will yield to the production of the first open platform delivering personalized medicine in Europe, leveraging significant innovation potential.

In 2008 Declan Butler, journalist at Nature, wrote an article entitled "Translational research: Crossing the valley of death". The first sentence of this article was: "A chasm has opened up between biomedical researchers and the patients who need their discoveries". The project described in this paper's primary goal is to reduce this chasm by providing a smooth path from the development of the bioinformatics methods to the recommendation of personalized treatments to the patients. The industrial partners in this consortium will lead the definition of an exploitation strategy that will allow us to install as early as possible the platform in different hospitals and to start analysing real cases in a "production" environment. SMEs and academic partners will work together to develop a licensing model in which part of the methods available through the platform could be sold to the end users. This project has the potential to break the barrier that usually divides basic and medical research.

6 Conclusion

After decades of continuous development, bioinformatics has reached the maturity with Petabytes of data that are currently available and thousands of methods produced by the bioinformatics community. One of the most active areas within the biomedical field is cancer research. Now it is time to capitalize on this knowledge to convert all this data into actions that can improve the quality of life of patients diagnosed with different cancer types.

This consortium will deliver a very innovative and open bioinformatics platform that will be a strong bridge between the science produced and the patients that ultimately need the treatments.

It will ultimately transform the practice of cancer treatment by providing support to the clinicians when standard treatments protocols have proven unsuccessful.

The development of the platform will be led by two of the largest IT companies in Europe. The software developed will meet all the requirements for software designed to run in the hospital environment. On top of this platform a modelling layer will be fed with robust methods developed by academic and SME partners that have demonstrated their capability in numerous publications and previous consortia. The Platform will regularly add new data and methods to guarantee that the patients are always receiving the most updated and accurate procedures.

This project represents a timely opportunity to improve and maybe save patients' lives using an innovative and cost-effective solution that will have significant impact on the treatment of cancer through:

- Making accessible to the hospitals the public data available and the patient data within an innovative and open bioinformatics platform operating with industrial standards.
- Developing and deploying a new generation of integrative bioinformatics methods through a modelling layer to propose personal cancer treatments.

Acknowledgements

The work described in this paper has been submitted to H2020-LEIT-BIO-2015-1 call under the title: CancerCureAdvisor - An open bioinformatics platform for personalized treatment of Cancer. If awarded, the work will be carried out by a consortium made of by the following institutions and industries: National Center Oncological Investigations, GeneStack, BioSolveIT GmBH, The University of Maryland Center for Bioinformatics and Computational Biology, Scientific Network Management SL, Norwegian University of Science and Technology, Université Joseph Fourier, University of Hamburg, Center for Bioinformatics, Consorci Mar Parc de Salut de Barcelona, Institut hospital del Mar d'Investigacions Mèdiques, Bull and Atos.

References

[1] Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. Science 331, 1653–1558, 2011.

[2] Von Hoff, D. D. et al. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. Journal of Clinical Oncology 28, 4877–4883 (2010).

[3] Miranda, K. et al. Integrated Next-Generation Sequencing and Avatar Mouse Models for Personalized Cancer Treatment. arXiv 76, 1358–1375, 2011.

[4] Dancey, J. E., Bedard, P. L., Onetto, N. Hudson, T. J. The genetic basis for cancer treatment decisions. Cell 148, 409–420, 2012.

[5] Yang, Y. et al. Databases and Web Tools for Cancer Genomics Study. Genomics, Proteomics & Bioinformatics, 2015. doi:10.1016/j.gpb.2015.01.005

Statistical Challenges of Big Data Analysis in Medicine

Jan Kalina[1]

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

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 2015; 3(1):24–27
received: April 30, 2015
accepted: May 6, 2015
published: June 15, 2015

1 Big Data in Medicine

The amount of data produced by contemporary medical research as well as healthcare organizations is enormous. Their analysis contributes to improving the efficiency of clinical decision making as well as patient safety. Medical data of various types and formats, which can be characterized as big medical data, have an underutilized potential for a dramatic change of current practices of healthcare. So far, intensive attention has paid to technological aspects concerning the storage of big medical data in large databases and their transfer or sharing [1], but not so much to the question how to analyze big medical data reliably remains to contain numerous open problems.

The statistical analysis of big data requires to carefully combine information of various types, including numerical data (continuous or categorical), often with a large number of variables (features), together with other data types (text, images, videos, graphs) [1]. Exactly, the analysis of such data may yield very practical tools of e-health including decision support systems or other telemedicine tools [11]. In other words, the data have a potential to bring the process of decision making closer to an individual patient.

The aim of this paper is to discuss the perspectives, challenges, and limitations of big data analysis and illustrate them on real research examples. We discuss the suitability of particular methods of multivariate statistics and data mining for the analysis of big data in Section 2. The role of statistics in medical research and healthcare practice increases and the everyday analysis of big data from various sources contributes gradually to the shift of current medicine towards a concept, which we call information-based medicine. We discuss this paradigm in Section 3. Section 4 describes our experience with analyzing real data within a broader interdisciplinary research. These include a cardiovascular genetic study (Section 4.1, a study of biometric authentication by means of keystroke dynamics (Section 4.2), and a study of brain activity within a psychiatric research (Section 4.3).

2 Principles of Statistical Analysis

2.1 General Principles

In the first place, we must mention the importance of the quality of big data. If the data are e.g. contaminated by severe noise or gross measurement errors, the result of the statistical analysis can be hardly reliable and we speak of the so-called garbage in–garbage out (GIGO) problem. The poor quality spoils for example large databases of molecular genetic data (e.g. RefSeq), which are publicly accessible also for (uncontrolled) input of new data.

Pre-processing the data is claimed to be the most time consuming part of the analysis of big data. The exploratory analysis includes primarily visualization tools (histograms, box plots or QQ-plots for individual variables, scatter plots for pairs of variables). If individual variables are measured in different groups, then descriptive statistics (quantiles, means and variances) or test statistics (of the F-test or χ^2-test) are useful to evaluate the difference among the groups.

Clinical decision making has the aim to a certain medical conclusion, integrating uncertainty as one of the aspects with an influence on the outcome. The physician solves the task of medical decision making based on data and knowledge connected to the cognition and determination of diagnosis, therapy and prognosis [11]. Although it is possible to reduce the complexity of the data by a prior variable selection, some authors pointed out at the suboptimal results [6]. Preferable approaches seem to employ a suitable regularization to obtain modified (regularized) versions of standard classification or clustering methods. Nevertheless, powerful and perspective tools for a reliable analysis of big data are searched for [8] and properties of available methods of multivariate statistics and data mining for the big data context are currently investigated.

The methods suitable for the analysis of big data (with a large number of observations n) often require efficient algorithms for a fast computation. Some of statistical methods, which are theoretically suitable for big data, are implemented in commercial software in a way, which is un-

suitable for big data applications. Another important requirement expected from statistical methods for big data is their comprehensibility for the medical problem of interest. In other words, the results of the computations must be clearly interpreted from the point of view of the medical research topic.

2.2 Dimensionality Reduction

Dimensionality reduction is commonly used as a preliminary or assistive prior to the statistical analysis of big data. Its methods may bring several benefits:

- Simplification of consequent computations,
- Comprehensibility (e.g. allowing to divide variables to clusters),
- Reduction or removing correlation among variables.

Numerous variable selection methods with the ability to ignore redundant variables suitable for a large p include (see [5]) e.g.

- Wrappers,
- Filters,
- Embedded methods,
- Minimum redundancy maximum relevance (MRMR),
- Information-theory based methods.

Feature extraction methods reduce dimensionality by replacing the raw data by linear combinations of variables. Their most important examples reliable for data with a large number of variables include

- Principal component analysis (PCA),
- Factor analysis,
- Correspondence analysis,
- Multidimensional scaling,
- Independent component analysis,
- Partial least squares regression.

If the dimensionality reduction is not performed prior to a more sophisticated analysis of the big data, one must resort to computationally demanding methods. However, the behavior of various methods for mining big data have not been systematically compared in various medical applications. Also the situation concerning the suitability of particular statistical methods for particular tasks can be described as rather chaotic. The development and validation of methods for the analysis of big data contains various open problems.

2.3 Classification Analysis

Commonly, the aim of the medical research is to learn a classification rule allowing to assign a new patient to one of several different groups according to the diagnosis or prognosis. However, some simple classification procedures are unsuitable for data with a large number of variables p.

The k-nearest neighbor classifier performs poorly for big data and linear discriminant analysis (LDA) is computationally infeasible if $n < p$. Multilayer perceptrons are commonly claimed to be universal classifiers suitable for high-dimensional data from the theoretical point of view. In practice, however, their implementation is computationally infeasible for $n \ll p$ e.g. in R statistical software. Support vector machines (SVM) are reliable for big data thanks to their suitable regularization. Their disadvantages include the impossibility to find optimal values of the parameters, tendency to overfitting, relying on a too large number of support vectors, and unavailability of a clear interpretation of the results.

There is a good experience with regularized versions of LDA, which contain an intrinsic variable selection and thus represent a more comprehensible tool [7]. Regularized methods, which can be characterized as tailor-made classification methods suitable for high-dimensional data, include also the lasso estimator in logistic regression.

2.4 Clustering

Cluster analysis is a general methodology for extracting knowledge about the multivariate structure of given data. In biomedical informatics, it is claimed to be the most common analytical method. It solves the task of unsupervised learning by dividing the data set to several subsets (clusters) without using a prior knowledge about the group membership of each observation. It is often used as an exploratory technique and can be also interpreted as a technique for a dimensionality reduction.

Clustering contains a wide variety of methods with numerous possibilities for choosing different parameters and adjusting the whole computation process [12]. While hierarchical clustering is not suitable for data with a large number of variables, there have been proposed efficient algorithms for the k-means clustering. A regularized version of the correlation coefficient was proposed within k-means clustering by [15]. A k-means cluster analysis based on the Euclidean distance of each observation from a shrunken mean of a particular cluster was proposed by [3]. However, the covariance matrix was assumed to be regular and therefore it was applied only to low-dimensional data in the numerical illustrations. Some approaches to k-means clustering are sensitive to the initialization of the random algorithm and to the initial selection of means of the clusters. Besides, the results strongly depend on the choice of the particular algorithm.

3 Information-based Medicine

The availability of large data sets contributes to the shift of current clinical practice towards a concept, which can be well characterized as information-based medicine. We understand it to be a new perspective paradigm in medicine, going beyond the current concept of evidence-based medicine (EBM) [4, 13].

The very essence of evidence-based medicine is the clinical evidence, i.e. knowledge acquired by clinical research. The limitations of a clinical study include transferring averaged results obtained by statistical methods to an (averaged) patient, without reflecting his/her individual situation [2]. Besides, clinical studies have a recent tendency to be bigger and more expensive, instead of focusing on small groups of specific patients.

Other types of data including molecular genetic data, medical signals and images, speech records or unstructured text remain outside of the clinical approach to evidence. This brings us to the concept of information-based medicine, which intends to overcome limitations of evidence-based medicine. Its distant aim will be to bring a virtual averaged information towards an individual patient based on his/her genetic and metabolomic parameters. At the same time, the effort to improve the information for clinical decision making brings new challenges for the basic research, e.g. in molecular genetics.

New data as new results of basic (not only clinical) research bring new information applicable to implement new tools of e-health. Thus, modern statistical methodology represents a necessary (but not sufficient) tool endorsing the realization of the ideals of the paradigm of information-based medicine.

4 Examples in Medicine

Three real applications of this section illustrate the diversity of sources and types of the data in current medicine research. The author of this paper was participating in the past or is participating at present on the analysis of these data sets, which most commonly contain a large number of variables p compared to the number of samples n. Such high-dimensional data can be analysed by methods which are non-standard and computationally demanding.

4.1 Cardiovascular Genetics

In a cardiovascular genetic study performed at the Center of Biomedical Informatics in Prague [11], a research of gene expressions was performed to construct a decision support system based on clinical and gene expressions data.

The microarray technology was used to measure average gene expressions of more than 39 thousands gene transcripts across the whole genome. LDA cannot be even computed if the number of observed variables p exceeds the number of observations n. Nevertheless, it is a typical situation in molecular genetics that there are thousands or tens of thousands of variables (gene expressions) measured on a sample of tens or hundreds of patients.

The analysis is complicated by the fact that there seems no remarkable small group of variables responsible for a large portion of variability of the data and the first few principal components seem rather arbitrary. Apart from LDA, also other methods of both statistics and data mining suffer from the so-called curse of dimensionality for such high-dimensional data, which is revealed through numerical instability or computational infeasibility [9]. We used regularized LDA to learn a classification rule allowing to assign a new individual to one of the given categories according to the diagnosis.

The resulting classifications rules for different situations are based on small sets of 10 genes. Their expressions allow to predict the risk of a manifestation of acute myocardial infarction or cerebrovascular stroke in the next 5 years for a particular patient. If a particular patient currently has acute myocardial infarction, then it is possible to predict the risk of a more severe prognosis or a relapse. Patients with a high risk of a future manifestation of a cardiovascular disease can be consequently monitored, which can increase the patient safety and lead to a more effective and safer care for patients with a life-threatening risk.

4.2 Biometric Authentication

The next example is an implemented software application intended to be applied for biometric authentication within a hospital based on writing medical reports. The application captures the keystroke dynamics when entering a username and a fixed password or continuously during the typing of the text [14], collecting the data directly from the operation system of the computer. Particularly, it records the code of every key, its name, time of pressing and releasing the key automatically without any delay.

The data are measured on $K = 32$ probands, who were asked to write a fixed password (*kladruby*) 5-times slowly and 5-times at the habitual speed. There are $p = 15$ variables including 8 keystroke durations and 7 keystroke latencies measured in milliseconds. The authentication is a classification task to 32 groups. Although p is small, the data are high-dimensional, because p still exceeds the number of observations, i.e. the number of repeated typing the password. This complicates the statistical analysis of the data, together with the fact that they are contaminated by severe noise.

Robust regularized versions of LDA turn out to perform better on the noisy data for classification to a small number of groups, but lose their performance for as many as 32 groups. This is a general problem of methods based on the Mahalanobis distance. Thus, our preliminary results show the SVM to be the most suitable among the available classification methods. However, an SVM is sensitive to the presence of outliers and improved results were obtained if outlying measurements were manually discarded prior to learning the classification rule.

4.3 Brain Activity and Schizofrenia

Different parts of the brain are responsible for specific functions. The spontaneous brain activity (i.e. resting-state brain networks) has been a hot topic in current neuroscience. We describe a psychiatric research study inves-

tigating the spontaneous activity of various parts of the brain by means of neuroimaging methods.

In the study, the brain activity of 24 probands is measured by means of fMRI under 7 different situations. One of them can be characterized as a resting state, without any stimulus. Besides, the probands were observing each of 6 different movies while measuring the brain activity in the same way. The data have the form of $p = 4005$ correlation coefficients among fMRI measurements in three-dimensional regions of the brain measured on $n = 24$ probands. Each of the correlation measures evaluates the connectivity between two parts of the brain.

The basic task is to classify the resting state from a movie on the high-dimensional data with $n = 24$ and $p = 4005$. Both SVM and regularized LDA yield the 100 % classification accuracy. Because the comprehensibility of the results is very important, we recommend to use the model based on the regularized LDA, which is based only on 81 variables.

The aim of a future research extending this study is to propose and implement a decision support system allowing to predict the diagnosis of schizofrenia in patients, whose disease is only at the initial stage of its development. Further, a combination with other data of various forms (including fMRI images and gene expression measurements) will reveal genetic predispositions for schizofreny.

Such new research tasks and results acquired in the basic research destroy the current paradigm, which has been called by various names:

- Evidence-based clinical practice in psychiatry,
- Evidence-based mental health,
- Evidence-based medicine in psychiatry,
- Evidence-based practices for psychiatry,
- Evidence-based practices for mental health.

Instead of the evidence-based concept, the development of psychiatry can be described as a way towards information-based psychiatry (IBP), as a specific concept of the information-based medicine for the psychiatric context.

Acknowledgement

The work was financially supported by the Neuron Fund for Support of Science. The data in Section 4.2 come from project 494/2013 of CESNET Development Fund. The data in Section 4.3 come from the Czech Science Foundation project No. 13-23940S.

References

[1] Baesens B. Analytics in a big data world. Hoboken: Wiley; 2014.

[2] Eddy DM. Practice policies: where do they come from? Journal of the American Medical Association 1990; 263: 1265-1275.

[3] Gao J, Hitchcock DB. James-Stein shrinkage to improve k-means cluster analysis. Computational Statistics & Data Analysis 2010; 54: 2113-2127.

[4] Guyatt G, Cairns J, Churchill D et al. Evidence-based medicine. A new approach to teaching the practice of medicine. Journal of the Americal Medical Association 1992; 268 (17): 2420-2425.

[5] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research 2003; 3; 1157-1182.

[6] Harrell FE. Regression modeling strategies with applications to liner models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.

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

[8] Kalina J. Classification methods for high-dimensional genetic data. Biocybernetics and Biomedical Engineering 2014; 34 (1): 10-18.

[9] Kalina J, Duintjer Tebbens J. Algorithms for regularized linear discriminant analysis. Proceedings BIOINFORMATICS 2015, 6th International Conference on Bioinformatics Models, Methods and Algorithms. Lisbon: INSTICC; 2015; 128-133.

[10] Kalina J, Papíková V, Zvárová V. Nové trendy v medicínské informatice a biostatistice. Nové trendy ve zdravotnických vědách II. Ústí nad Labem: UJEP; 2014; p. 3. (In Czech.)

[11] Kalina J, Seidl L, Zvára K, Grünfeldová H, Slovák D, Zvárová J. System for selecting relevant information for decision support. Studies in Health Technology and Informatics 2013; 186: 83-87.

[12] Kogan J. Introduction to clustering large and high-dimensional data. Cambridge: Cambridge University Press; 2007.

[13] Sackett D, Rosenberg WMC, Gray MJA, Haynes BR, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996; 312: 71-72.

[14] Schlenker A, Šárek M. Behavioural biometrics for multi-factor authentication in biomedicine. European Journal for Biomedical Informatics 2012; 8: 19-24.

[15] Yao J, Chang C, Salmi ML, Hung YS, Loraine A, Roux SJ. Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. BMC Bioinformatics 2008; 9 (Article 288): 1-16.

Systematic Exploring of Associations between Folate Deficiency and Autism

Daniel Krsička[1], **Libor Seidl**[1]

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

Correspondence to:

Daniel Krsička
1st Faculty of Medicine, Charles University in Prague
Address: Kateřinská 32, 121 08, Praha, Czech Rep
E–mail: daniel.krsicka@lf1.cuni.cz

IJBH 2015; 3(1):28–30
received: May 1, 2015
accepted: May 13, 2015
published: June 15, 2015

Autism or Autism Spectrum Disorders (ASD) is serious lifelong neurobehavioral impairment, significantly complicating the integration of an individual into normal life. It comprises a group of pervasive developmental disorders characterized by communication difficulties, impaired reciprocal social interaction, imagination insufficiency and stereotypic behavior. The ASD etiology is mostly unknown, genetic syndromes only account for an estimated 15% of autism cases. The ASD has uncertain prognosis and full social integration is currently unachievable for the majority of ASD patients. Prevalence varies between populations and studies, and ranges from 0.1% to 1.5% (our own not yet published survey of the prevalence studies) and todays estimated lifelong costs of supporting an ASD individual is about 2.3 million USD [1]. Thus ASD represents a substantial socioeconomic burden.

Present research consists of 2 major streams. The first stream investigates the polygenic influence of many genes and concentrates on the genome itself. The second stream focuses on epigenetic changes and environmental factors. Both research streams are comprised of Big Data Challenges like high-volume data processing from personalized whole genome sequencing, huge data set comparison, automation of new information derivation and others. For the effective resolution of phenotypes, early diagnostics, screening and biological therapies for ASD it is necessary to better understand synergistic effects of several genetic and epigenetic causes, which together manifest as neurobehavioral syndrome, but each independently represents only a subclinical, not manifest problem.

According to several recent research results it seems that phenotypes could exist for the manifestation of ASD and also for other diseases like schizophrenia or early geriatric dementia. These phenotypes consist of influence of several harmful factors. The genetic component still seems to be important but cannot convincingly explain all the cases and research results. It seems that the environmental factors, forming the epigenetic influences, should be taken into account. These factors can manifest in other way and couldn't be primarily connected to ASD in clinical praxis. One of these manifestations can be the Cerebral Folate Deficiency (CFD). Primary CFD is a neurodegenerative syndrome characterized by reduced levels of 5-MTHF (5-methyl tetrahydrofolate) in the cerebrospinal fluid (CSF) but normal levels of 5-MTHF in serum and erythrocytes. Symptoms and severity are highly variable ranging from heavy disability with progressive microcephaly, psychomotor retardation, sensoric disabilities, cerebellar ataxia and movement disorders, to mild only behavioral, social or cognitive problems like autism, mental retardation, early geriatric amnesia or schizophrenia but without any physical symptoms. Some published papers show that CFD can contribute to the ASD development and progression and that the CFD compensation can suppress the core ASD symptoms, common referred as incurable [2, 3, 4, 5, 6, 7].

In our recent review, we've found a total of 351 published cases of CFD with various ASD reported in 44% of patients [8]. However a number of studies do not indicate whether patients were subjected to differential diagnostic test for the ASD evaluation. Furthermore, in very severe CFD cases the precise diagnostic examination for ASD could not be realized. Many of these studies have reported impaired social interaction and communication in patient's history. Thus it is possible that the overall proportion of ASD was higher than 44%. Variable positive treatment effect on the core ASD symptoms (communication, social interaction ...) was reported in 29% of patients with ASD and CFD. The most common cause of the published CFD cases were the Folate Receptor Auto Antibody (FRAA) positivity, the second largest group consists of an CFD of unknown etiology. The remainder belongs mostly to the genetic defects. Significantly elevated FRAA titers results in a typical CFD, however, the significance of slightly elevated FRAA levels for the ASD pathophysiology is not yet precisely known, although many pa-

tients responded positively to the treatment. Disruption of folate and folate-dependent metabolic pathways localized only in the CNS is in ASD quite hard to diagnose. The only available and completely reliable method is the lumbal puncture with examination of 5-MTHF level in CSF, which is not routinely indicated in ASD for its invasiveness. Less invasive methods, such as MR spectroscopy has not the needed resolution in nmol/L at present. The FRAA assay is available only in a few laboratories in the world. It is therefore possible that mild CFD in ASD escapes the attention in diagnosis and treatment for a long time, and that it may contribute to the development and progression of idiopathic ASD. This disorder is partially or rarely completely compensable. Many publications show the negative correlation between the age of the patient and treatment outcome. Therefore, an early intervention seems to be essential. The potential importance of folate metabolism for ASD pathophysiology also illustrates the fact that there has been published other experimental studies of ASD treatment based on administration of substances, whose synthesis or concentration is directly or indirectly folate-dependent or folate-controlled or they act as cofactors in folate metabolic pathways.

All above mentioned information and the large proportion of unknown CFD etiology leads us to further research. We assume that the ASD etiology may be strongly associated with a synergistic effect of several epigenetic causes, which together manifest as a neurobehavioral syndrome but each independently represents only a subclinical, not manifest problem. Research of this hypothesis is fully appropriate. Potential confirmation would significantly affect the further clinical research of ASD medications according to the rules of evidence based medicine. Heterogeneous causes which can occur simultaneously and synergistically reinforcing each other if they are not examined all together, make practically impossible to correctly select the test group for a clinical trial. Mixed test groups composed of more mutually different ASD phenotypes would affect the outcome of each trial in a random manner and the results of these studies would have always the significant differences. Gradually, it would be possible to identify only the dominant causes, as appears to be in serious CFD based on FRAA. Conversely, if it were possible to identify and classify particular, separately insignificant, but collectively manifest causes, it would contribute significantly to the identification of novel ASD etiologies, and to the respective phenotypes classification. However, the identification and classification of multiple causes and their mutual relations, manifestations, and their relationships across specializations of internal medicine, neurology, genetics and psychiatry, obviously requires a specific approach to capture the issues and the need of a structured information and knowledge modeling using the apparatus of ontologies and formal languages.

In our further work we are focused on systematical research of folate depletion significance and relations. Abnormalities in folate levels could probably contribute to many of epigenetic mechanisms well-connected to etiology and pathophysiology of ASD and other neurodevelopmental and neurodegenerative disorders like abnormal genome methylation, oxidative stress, mitochondrial and neuronal damage or an abnormal immune response. The long term folate income itself is probably not so important. But acting together with other factors like insignificant genetic mutations, subclinical environmental toxic burden, long-term stress or specific autoantibodies like the FRAA, it hypothetically explains and could cause abnormal CNS growth, neuronal differentiation, migration and pruning or activate apoptosis, leading to neurodegeneration. With relation to neural tissue, long-term insult influence is not needed. The influence of short- or medium-term folate depletion has not been systematically studied. To gain the most comprehensive knowledge about the role of folates and their relations to other etiological factors the knowledge from several biomedical databases can be used. According to our previous review we have found specific relations among autism, Folate Receptor Auto Antibodies and local folate disturbances in a brain. Further research of high volume information about metabolomics, signaling, genetic coexpression, protein interactions and other information could reveal some new facts and knowledge about the multifactorial nature of ASD. Such knowledge can be used for defining new clinical hypotheses and also to strongly contribute to the prevention, early diagnosis and treatment of fragile population.

Keywords

Big Data, Autism, Autism Spectrum Disorder, Folate Receptor Auto Antibody, Cerebral Folate Deficiency

Acknowledgements

This paper has been partially supported by the SVV-2015-260158 project of Charles University in Prague.

References

[1] Barrett B. Substantial lifelong cost of autism spectrum disorder. J Pediatr [Internet]. 2014 Nov [cited 2015 May 3];165(5):1068–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25441393

[2] Ramaekers VT, Rothenberg SP, Sequeira JM, Opladen T, Blau N, Quadros E V, et al. Autoantibodies to folate receptors in the cerebral folate deficiency syndrome. N Engl J Med [Internet]. 2005 May 12;352(19):1985–91. Available from: http://dx.doi.org/10.1056/NEJMoa043160

[3] Ramaekers VT, Sequeira JM, Blau N, Quadros E V. A milk-free diet downregulates folate receptor autoimmunity in cerebral folate deficiency syndrome. Dev Med Child Neurol [Internet]. 2008 May [cited 2014 Sep 14];50(5):346–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18355335

[4] Moretti P, Peters SU, Del Gaudio D, Sahoo T, Hyland K, Bottiglieri T, et al. Brief report: autistic symptoms, developmental regression, mental retardation, epilepsy, and dyskinesias in CNS folate deficiency. J Autism Dev Disord [Internet]. 2008 Jul [cited 2014 Sep 6];38(6):1170–7. Available from: http://dx.doi.org/10.1007/s10803-007-0492-z

[5] Steele SU, Cheah SM, Veerapandiyan A, Gallentine W, Smith EC, Mikati MA. Electroencephalographic and seizure manifestations in two patients with folate receptor autoimmune antibody-mediated primary cerebral folate deficiency. Epilepsy Behav [Internet]. 2012 Aug [cited 2014 Sep 14];24(4):507–12. Available from: http://dx.doi.org/10.1016/j.yebeh.2012.05.021

[6] Frye RE, Sequeira JM, Quadros E V, James SJ, Rossignol DA. Cerebral folate receptor autoantibodies in autism spectrum disorder. Mol Psychiatry [Internet]. Nature Publishing Group; 2013 Mar [cited 2014 Sep 6];18(3):369–81. Available from: http://dx.doi.org/10.1038/mp.2011.175

[7] Al-Baradie RS, Chaudhary MW. Diagnosis and management of cerebral folate deficiency. A form of folinic acid-responsive seizures. Neurosciences (Riyadh) [Internet]. 2014 Oct [cited 2014 Oct 19];19(4):312–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25274592

[8] Krsička D, Vlčková M, Havlovicová M. The Significance of Cerebral Folate Deficiency for the Development and Treatment of Autism Spectrum Disorders. Int J Biomed Healthc [Internet]. 2015;1(1). Available from: http://ijbh.org

Big Data Versus Rare Cases

Lenka Lhotská[1], **Miroslav Burša**[1], **Michal Huptych**[1], **Matej Hrachovina**[1]

[1] Czech Technical University in Prague, Prague, Czech Republic

Correspondence to:

Lenka Lhotská
Czech Technical University in Prague
Address: Zikova 4, 166 36 Prague 6
E–mail: lhotska@fel.cvut.cz

IJBH 2015; 3(1):31–32
received: April 30, 2015
accepted: May 20, 2015
published: June 15, 2015

1 Introduction

Big data is a term that has been appearing more and more frequently. Usually it represents data files that cannot be stored, managed and processed using standard software tools in reasonable time due to their volume. It is necessary to mention that these data are very heterogeneous, multimodal, unstructured, and sometimes partially less trustworthy.

Technology development means generation, storage, management and processing of larger and larger volumes of data, which are heterogeneous, multimodal, structured and unstructured, sometimes noisy and incomplete and gathered from different sources. Some data characteristics are common for data from different problem areas. However, data in medicine show certain specific properties. One of them is existence of small data sets described by a high number of features; another one is existence of rare cases described by outliers which cannot be ignored but must be considered in successive analysis. In medicine there are more frequently utilized interactive methods of knowledge discovery where an expert is part of the process and individual analysis steps are controlled by his/her knowledge. Methods of structural learning and graphical models utilizing probabilities are becoming more important. Inseparable part of the big data processing is suitable visualisation of data, processing and results.

2 Big Data and Their Properties

The basic characteristics of big data can be expressed using the term 4V – volume, velocity, variety and veracity. Volume means that the data volume is growing exponentially. Velocity represents requirement of many tasks for real-time processing. Variety is related to types of data: structured, unstructured, multimedia. Last property, veracity, says that acquired data are not always hundred per cent reliable, complete, consistent, etc. An example can be data coming from communication on social networks, or measured data which can be noisy.

Data storage is an important issue. One of the possible ways is to use NoSQL databases. There exist various types that differ by used data models. They can be compared according to various criteria. Scofield [1] proposed following basic criteria: performance, scalability, flexibility, complexity and functionality. Choice depends on application area and predictability of future possible data.

Data structure and standardization represent important issue, in particular in medicine. Concerning structure, we can divide data into several categories: weak structure, good structure, bad structure (frequently used in contrast to good structure, however originally used in different context [2]), partial structure, unstructured data (mostly used for data in natural language, although text has a certain structure). Precisely, unstructured data should denote purely random data, thus noise. Duda, Hart and Stork [3] define such data property that appears due to randomness (in real world, from sensors or measurement). In informatics they can be unwanted irrelevant data without any sense and relation to other data. Further we can divide data into standardized (e.g. numerical items in laboratory report) or non-standardized (e.g. non-standardized text in patient database record, frequently denoted as free text). Requirement on standardization is very important because data standards ensure that information is presented in such form that supports system interoperability [4] and allows the end user to compare data during interpretation. Standards support reusability of data, improve efficiency of health care services and prevent from errors caused by e.g. duplicity of inputs. Standardization concerns following areas: data content; terminology used for data representation; data communication; knowledge utilization.

3 Data in Medicine

Data in medicine are mostly stored isolated in various storages and quite frequently it is difficult to analyse them easily in frame of a single country. In general, the health care institutions collect and use two sets of data:

retrospective data (basic information about events gathered from medical records) and clinical data (collected in real time and presented at the point of care. Data mining and data analysis methods allow to interconnect both data types. Then clinicians can use relevant information and identify trends having impact on future health care – area known as predictive analysis. It may help to develop preventive and long-term care individualized for corresponding patient groups.

Going further and adding data on gene sequences, we can discover connection between genes and sensitivity to certain diseases. Then effective treatment may be started earlier.

Big data on one side and small data and many features on the other side – these are two sides of medical data. That means that it is necessary to develop methods for analysis and discovery of connections in rare cases [5]. Quite frequently the rare cases show up as outliers that are neglected or denoted as noise by standard methods. But they must not be neglected. They must be considered in successive analyses. Very promisinge in this area are interactive methods of knowledge discovery, when an expert participates actively in the process and contributes with his/her knowledge.

4 New Trends in Medicine

One of the greatest future aims is modelling of the whole patients in their complexity that could support medical decision making, health care practice and treatment tailored to the needs of each individual patient. However this trend towards personalized medicine generates greater data volumes in high dimensions. This complicates manual analysis and in some cases makes it even impossible.

Therefore efforts to develop efficient, applicable and useful computational methods and visualization tools have been appearing recently. Adaptation of data representation is another important area. Its main task is mapping of multidimensional data to lower number of dimensions for easier interpretation and visualization since the outputs must be transparent and interpretable for the users.

Thus it is necessary to search for suitable combinations of areas and tools that can offer good conditions for processing of big biomedical data. Such synergic combination is represented by human-computer interaction and knowledge discovery from data. Recently the idea of big data analysis utilizing machine learning methods in combination with “doctor/expert-in-the-loop” has been intensively developed. It means that the whole process is not fully automatic but in certain phases a human with his/her expert knowledge enters into data processing tasks with the aim to focus attention of the system into the right direction for the next step.

The steps of data processing in this approach include:

- Data integration (fusion, mapping, cleaning, preprocessing).
- Interactive machine learning (and cognitive computation, knowledge discovery).
- Visual analytics (and expert-in-the-loop, decision support).
- Data privacy, security, ethics, evaluation.

5 Conclusions

In addition to continuous increase of data volume in medicine, multidimensionality of the data and grow of unstructured information are substantial features. Since human cognitive abilities to process such data are limited, a great challenge is represented by the development of visualization tools for expressing semantic aspects of information [6] and understanding of cognitive and communication processes of information perception.

Acknowledgements

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.

Conflicts of Interest

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

References

[1] Scofield, B. (2010-01-14). ”NoSQL - Death to Relational Databases(?)”. Retrieved 2015-02-26. http://www.slideshare.net/bscofield/nosql-codemash-2010

[2] Simon, H.A. (1973). The Structure of Ill Structured Problems. Artificial Intelligence, 4 (3-4), 181-201

[3] Duda, R.O., Hart, P.E., Stork, D.G. (2000). Pattern Classification, 2nd edition, Wiley, New York.

[4] Huptych, M., Lhotská, L. (2014). Multi-layer Data Model: A Contribution to Semantic Interoperability. In IFMBE Proceedings - 6th European Conference of the International Federation for Medical and Biological Engineering. Heidelberg: Springer, vol. 45, 749-752.

[5] Hrachovina, M., Huptych, M., Lhotská, L. (2013). Data Acquisition and Storage System in a Cardiac Electrophysiology Laboratory. In Information Technology in Bio- and Medical Informatics, ITBAM 2013. Heidelberg: Springer, 77-87

[6] Burša, M., Lhotská, L., Chudáček, V., Spilka, J., Janku, P., et al. (2013). Visualization in Information Retrieval from Hospital Information System. InSoft Computing Models in Industrial and Environmental Applications. Heidelberg: Springer, 459-467

Diagnostic Software for Decision Support of Detection and Interpretation of Tumor Markers

Ladislav Pecen[1], **Marcel Jiřina**[2], **Jakub Novák**[2]

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

[2] Department of Theoretical Informatics, Faculty of Information Technology, CTU, Prague, Czech Republic

Abstract

Values of immunological examinations of tumor markers, various hormones and relating parameters are significant identifiers providing important information about pathological situation and possible serious illnesses of a patient. Laboratory analysis provides primary information about values of measured parameters that can be further utilized together with other clinical data as inputs to diagnostic software applications which can reveal particular illnesses and clinical status of a patient. The paper briefly describes cloud-based software applications that can serve as supporting diagnostic tools.

Keywords

Diagnostic software, tumor markers, immunological examination

Correspondence to:

Marcel Jiřina
Dept. of Theoretical Inf., FIT, CTU in Prague
Address: Thákurova 7, 160 00 Prague 6
E–mail: marcel.jirina@fit.cvut.cz

IJBH 2015; 3(1):33–36
received: April 30, 2015
accepted: May 6, 2015
published: June 15, 2015

1 Introduction

Immunological examinations of tumor markers, various hormones and relating parameters belong to standard diagnostic methods in the present day. More than one diagnosis is linked to pathological values of most of these parameters, and several parameters need to be measured simultaneously. In the course of a patient's illness, laboratory parameters are usually measured repeatedly, with a particular importance of observing the dynamics of the development of these parameters over time. Such and other circumstances have led to the beginning of co-operation between physicians and mathematicians, resulting in several mathematical models having been implemented into the respective programs [1]. The purpose of such models is effectively to make use of information contained in laboratory measurements together with some other clinical data, and to summarize them as probabilities of particular illnesses and clinical status. Model parameters are estimated on the basis of statistical processing of a suitable reference file of patients, taking into consideration information from the literature and from experts in the field. The models also take into account epidemiological data, in particular incidence and prevalence of illnesses [3]. Correlations between laboratory parameters and their mutual influencing are also taken into consideration. The model output – quantified assessment of current and previous laboratory findings for the given patient – then becomes one of the components serving in the physician's decision making, along with other assessments. Models and their software implementations are being constructed since 1990 and in the stage of prototypes have been used for several years in various clinical settings. Since 1993 they have been commercially distributed in the form of program packages BIANTA and CRACTES (interpretation of tumor markers) [2]. New versions incorporating modern cloud solution are now available.

2 Bianta

BIANTA is a software decision support system for primary and suspect cancer diagnostics. The key question evaluated by the program is the determination of an unknown type and location of primary tumor. This complicated situation arises when tumor occurrence has been clearly shown, but several possible locations can be taken into consideration, i.e. if metastases were found but histological findings do not confirm explicitly any specific site of origin. Such cases form approx. 3-5 % of all primary tumor illnesses. The program can also be applied with success in the case of suspect tumor diagnosis (i.e. tumor illness suspected), as well as for primary tumor di-

agnostics in order to exclude the possibility of mistaking metastases for primary tumor, and in order to exclude duplex (i.e. two independent primary tumors). The system is based on the theory of Bayesian Network. The input data are the results of patient's blood serum measurements of tumor markers, the patient's age, sex and location of metastases if found.

3 Cractes

CRACTES is the decision support software for tumor markers interpretation during patient's post-cancer treatment follow-up. The main task of the program is early diagnostics of the recurrence of tumor and its most probable location. The examination of markers may be a warning signal, but it also may be a false alarm; this reminds of the jay bird which signals an enemy but has to be taken with a pinch of salt to a certain extent. After reaching the clinical status of full remission the patient is usually invited to be checked once in 1-4 months intervals (according to time elapsed since treatment of primary tumor). It is desirable to identify the metastatic process earlier than it is manifested clinically. The system helps to answer the following question: "Will clinically evident metastases develop until the next check-up of the patient in remission?" The risk of genesis of distant metastases is modelled by the methods of statistical survival analysis. A specific analysis was made for every type of tumor diagnosis. The risk of genesis of metastases was estimated as a function of all previous measurements of tumor markers, specific for a given diagnosis.

4 Software

Strengths of the software:

- High-quality selection of mathematical methods
- Based on published high-quality scientific results
- The mathematical methods used are thoroughly tested and compared with other clinically validated data files
- The coefficients of the models are estimated on high-quality and large datasets leading to high accuracy of output results
- The software is clinically used for over 16 years
- Easy and universal implementation to different software applications of clients (LIS, LIM)

4.1 Overview

The software is modular based to provide easy access to particular modules (Bianta, Cractes ...) and future possibilities to update or extend any of them without disabling the others. Usage of Java technology enables the software to run on every platform and therefor targets wider audience. It is created as a cloud-based application with access through desktop graphical user interface.

4.2 Architecture

General overview of the software can be divided into four parts as can be seen on Figure 1.

- Cloud-based modular application
- Desktop graphical user interface (GUI)
- Access database
- Patients database

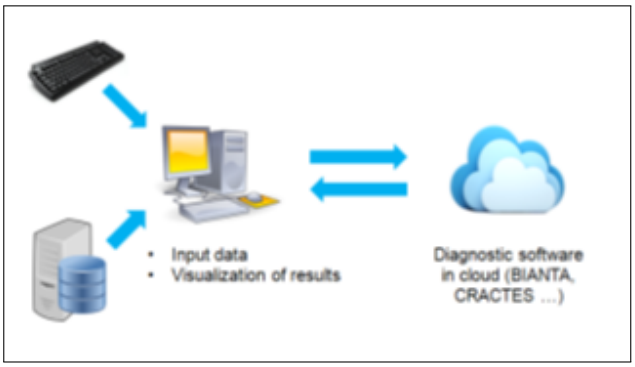


Figure 1: Overview of the cloud-based diagnostic software implementation.

4.3 Workflow

The user chooses the needed module based on the task to be solved. He then runs the desktop graphical user interface (GUI) and loads the data. This can be performed either by creating a 'new patient' or more patients manually or by importing a set of patient data from patients' database. The desktop GUI then creates a text document in valid format which reflects patient's data. It is necessary to anonymize the data before sending them to cloud-based application. These text strings are then sent to the cloud-based application which computes the results. Connection between GUI and cloud-based application can be established only if the user certifies himself with a unique access key from the access database. After receiving the output from the cloud-based application, the GUI is used to perform visual representation of the results in a user-friendly format.

4.4 Cloud-based Application

Cloud-based application contains various modules (Bianta, Cractes etc.). If any of the modules receives data in a valid format the results will be computed. Valid data format looks like the one in Figure 2.


```
pacient: JOHN SMITH [330]
18.3.2015 MIRA 2
18.3.2015 CA125 104.0
18.3.2015 CEA 105.0
18.3.2015 CYFRA 18.0
18.3.2015 CA19-9 48.0
```

Figure 2: Example of valid data format as input to any module of the cloud-based application.

This data format contains patient's name and three numbers. First two are the last two digits of the patient's year of birth and the third one reflects patient's gender. The MIRA parameter corresponds to the degree of cancer suspicion. It ranges from 1 to 4. Number one means "a patient without clinical difficulties" and number four means "sure tumor with unknown origin". The remaining lines consist of a date when the tumor marker was measured, the marker's abbreviation and its measured value.

Output of the Bianta module for example is a text document with a list of possible cancer types ordered in respect of their probabilities. Bianta module also includes the benign diagnoses. If the module doesn't have enough parameters to provide diagnoses it recommends other markers to be measured.

4.5 Desktop Graphical User Interface (GUI)

GUI handles the patients' data loading phase, the data format translation to anonymize them, the communication with the cloud-based application and its authorization in access database and displaying the results in an understandable format. It could also be used to create a patient or a set of patients. The interface to create a patient uses the fields in Figures 3, 4, 5 and 6. It also enables the user to edit imported patients' data from a patients' database.

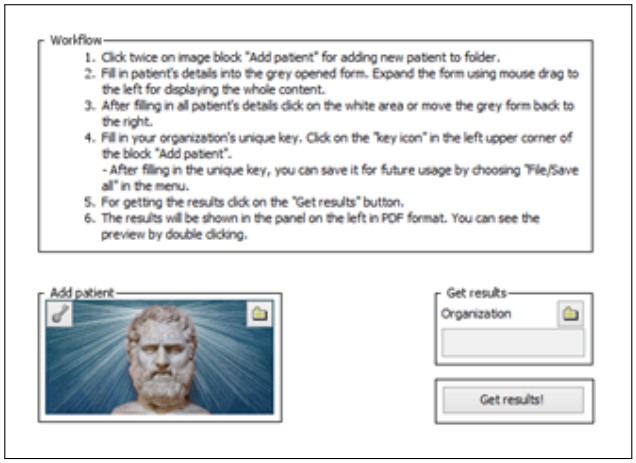


Figure 3: Main view of desktop graphical user interface.

Figure 4: A field for adding the patient's information.

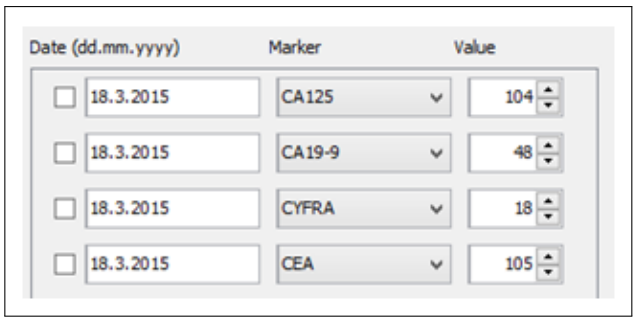


Figure 5: A field for adding the measured markers and their values.

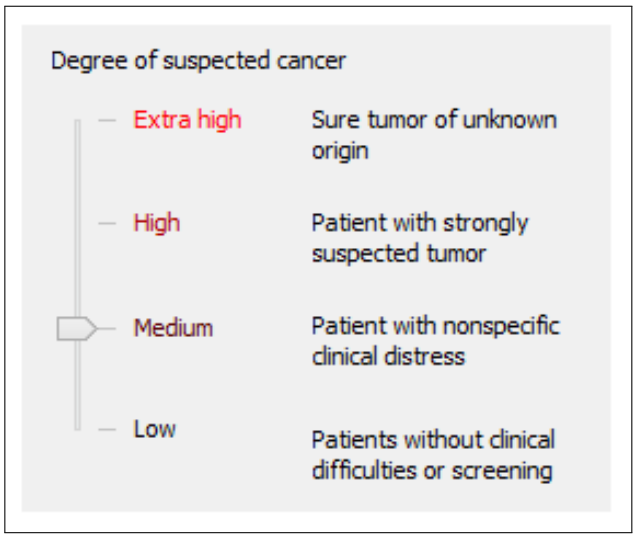


Figure 6: A field for selecting the degree of cancer suspicion.

GUI can be used to view the output of the cloud-based application in for example a pdf or another understandable format.

5 Conclusion

The software is being successfully used at a number of medical offices in the Czech Republic. We are planning to continue developing other programs/ applications dealing with various other diseases, such as bone metabolism, diabetes, or thyroid diseases. Furthermore, we are aiming to integrate population epidemiological data from other countries, such as Germany and France, to make the program applicable and usable outside of the Czech Republic.

References

[1] Pecen L., Eben Kryštof, Topolčan O., Holubec L., Pikner R., Nekulová M., Šimíčková M., Kaušitz J.: Programs for the Interpretation of Laboratory Results in Oncology. Anticancer Research, Vol. 22, 2002, No. 1B, pp. 531-532 (ISSN: 0250-7005) Held: International Hamburg Symposium on Tumor Markers /11./, Hamburg, Germany, 2002.01.27-2002.01.29

[2] Pecen Ladislav, Topolčan O., Nekulová M., Šimíčková M., Kaušitz J., Eben Kryštof, Vondráček Jiří, Pikner R., Holubec L. Evaluation of the CRACTES and BIANTA programs for the result interpretation of tumor marker assessment. Journal of Tumor Marker Oncology, Vol.15, 2000, No.1, pp.57-58 (ISSN:0886-3849)

[3] O.Topolcan, L.Pecen, J.Kausitz, M.Nekulova, M.Simickova, D.Valik, R.Pikner, L.Holubec: Computer assisted intepretation of tumor markers - possibilities for new application of a mathematical models. Journal of Tumor Marker Oncology, Vol.16, 2001, No.1, (ISSN:0886-3849)

Big Data and Personalized Medicine

Jaroslav Pokorný[1]

[1] Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Abstract

The paper describes the field of Big Data storage and processing in relationship to healthcare, particularly to personalized medicine. We discuss the basic characteristics of Big Data and its main research and application area – so called Big Analytics. We also briefly mention the challenges, and offer some conclusions.

Keywords

Big Data, Big Analytics, Personalized medicine

Correspondence to:

Jaroslav Pokorný
Department of Software Engineering,
Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic
Address: Malostranske nam. 25, 118 00 Praha 1
E–mail: pokorny@ksi.mff.cuni.cz

IJBH 2015; 3(1):37–40
received: April 30, 2015
accepted: May 6, 2015
published: June 15, 2015

1 Introduction

Usually we talk about the Big Data when the dataset size is beyond the ability of the current system to collect, process, retrieve and manage the data. Authors of [6] describe Big Data as large pools of unstructured and structured data that can be captured, communicated, aggregated, stored, and analyzed. Big Data is now becoming part of every sector and function of the global economy. Data created both inside corporations and outside via the Web, mobile devices, IT infrastructure and other sources increases exponentially each year [3].

Obviously, Big Data occurs in medicine area and provides a nice example of a high rate of growth. The area produces more data in more forms than most industries. For example, U.S. healthcare will soon reach the zettabyte (1021 GB) [9]. Without doubts, by combining data from many current and future devices and from different data sources, healthcare providers can have greater insight into medicine. On the other side, e.g., translation and personalized medicine based drug development requires integration with clinical and real-world patient and physician data which are spread across disparate data stores across the globe.

The purpose of this paper is to introduce some notions usually discussed in context Big Data applications in Business Intelligence (BI) into the domain of medicine, particularly so called personalized medicine. First, we accept a short definition of the notion personalized medicine. According to [11], personalized medicine is the precept that the treatment of disease is most effective when the definition of both the disease and the treatment is individualized at the patient level. In other words, personalized medicine is a medical model with the idea of customization of healthcare.

McKinsey Global Institute says that personalized medicine holds the promise of improving health care in three main ways [6]:

- offering early detection and diagnosis before a patient develops disease symptoms,
- more effective therapies because patients with the same diagnosis can be segmented according to molecular signature matching (i.e., patients with the same disease often do not respond in the same way to the same therapy, partly because of genetic variation),
- and the adjustment of drug dosages according to a patient's molecular profile to minimize side effects and maximize response.

Personalized medicine is now driven by the informatics revolution popularly referred to as Big Data. In general, the context of Big Data contains adequate tools and methods for its storage and management. On the application level, new approaches to Big Data analysis are now developed. It is usual to call them Big Analytics.

In Section 2 we introduce a list of Big Data characteristics as they occur in literature. Section 3 mentions some consideration about usability of Big Data storage and processing for personalized medicine. Conclusions summarize basic observations concerning the relationship of Big Data techniques and personalized medicine.

2 Big Data

In the past we used Data Warehouses (DW) to make data available for new users. But DW are limited to one compute center and (often) one server. They also need structured tabular representation. A big motivation for Big Data management was the textual Web content which people wanted to easily consult and search. Challenges in this area included primarily document summarization, personalized search, sentiment analysis, and recommender systems. Moreover, the social structures formed over the Web, mainly represented by the online social networking applications such as Facebook, LinkedIn and Twitter, contributed intensively to development of Big Data tools. Large datasets are typical also for medicine, e.g., existing healthcare data includes personal medical records, radiology images, human genetics and population data genomic sequences, etc. Newer forms of Big Data, such as 3D imaging, genomics and biometric sensor readings, are also fueling the exponential growth of medicine data.

2.1 Big Data Characteristics

Big Data is mostly characterized by several V's:

Volume: Data scale in the range of TB to PB and even more. The big volume is not only storage issue but also influences an analytics. Not only data samples, but often all data are captured for analysis.

Velocity: Both how quickly data is being produced and how quickly the data must be processed to meet demand (e.g., in context of streaming data).

Variety: Data is in many format types – structured, unstructured, semistructured, text, media, etc. Data does not come only from business transactions, but also from machines, sensors and other sources, making it much more complex to manage.

Veracity: Managing the reliability and predictability of inherently imprecise data.

Value: Indicates if the data is worthwhile and has value for business.

Visualization: Visual representations and insights for decision making.

Variability: The different meanings/contexts associated with a given piece of data.

Volatility: How long the data is valid and how long should be stored (at what point specific data is no longer relevant to the current analysis).

Usually the first three V's are taken into account. Clearly, these V's are mutually not too consistent, their semantics overlap, etc. Some of them are different in specialized contexts. For example, data value vision in business includes creating social and economic added value based on the intelligent use, management and re-uses of data sources with a view to increase BI. Data value is different in medicine, namely in personalized healthcare, drug repurposing, new biomarkers, etc.

Velocity means numerous real-time data streams coming in from medical devices, RFID devices, logs, etc.

Variety in medicine data faces many of the same issues as in today's enterprise data. Structured and semistructured data as well as unstructured data is processed there. Variety can be observed also in rather database processing of medicine data. For example, in the Drug Encyclopedia project [4, 5], an integrated dataset drug related data has been created. Integration of 9 heterogeneous datasets is done via Resource Description Format (RDF) [12] and Linked Data principles [1].

A soft problem of Big Data in medicine is its veracity. Data quality issues are of acute concern in healthcare, because decisions dependent on associated data processing can significantly influence patient's health and life.

Of a special importance in medicine is visualization and so called visual analytics. Visual analytics has been defined as "the science of analytical reasoning facilitated by interactive visual interfaces" in [10]. Visual Analytics is more than just visualization of the data; it is an approach that combines visualization, human factors and data analysis. In combination with Big Data its utility for decision making processes is increasing.

Big Data variability in medicine is reflected, e.g., in responses to treatment. They depend on the underlying genetic makeups of individual patients.

Finally, volatility means, e.g., that the data from patient's health applications does not need to be retained, but can rather be analyzed and then archived or deleted, in accordance with legislative requirements for health information.

From the database point of view, a typical feature of Big Data is the absence of a schema characterization, which makes difficulties when we want to integrate heterogeneous datasets.

Despite of the fact that a common definition of Big Data is still unclear, this paradigm is recognizable and usable in practice including medicine.

2.2 Big Data Processing and Big Analytics

As data is becoming more and more complex its analysis is also becoming increasingly complex. To exploit this new resource, we need to scale both infrastructures and standard data management techniques. Now the problem with data volume is it's speed (velocity) not only size. Big Data processing involves interactive processing and decision support processing of data at rest, and real-time processing of data in motion. The former can be warehoused in a relatively traditional way or stored and processed by inexpensive systems, e.g., NoSQL databases. The latter is usually performed by Data Stream Management Systems. Time is an integral dimension of data in a stream which influences its processing. A velocity also can be a prob-

lem, since the value of the analysis (and often of the data) decreases with time. If several passes are required, the data has to be put into a DW where additional analysis can be performed.

Big Analytics is a process of analyzing large amounts and different types of data to uncover hidden patterns, correlations and other useful information. Big analytics could accomplish much more than what can be done with smaller datasets. For example, Big Analytics allows move beyond linear approximation models towards complex models of greater sophistication because small datasets often limit our ability to make accurate predictions and assessments. Additionally, Big Data significantly improves the ability to locate and analyze the impact of rare events that might escape detection in smaller datasets. Access to larger datasets, affordable high-performance hardware, and new powerful analytical tools provide means for better accuracy in predictions.

Big Analytics is about turning information into knowledge using a combination of existing and new approaches. Related technologies include:

- data management (uncertainty, query processing under near real-time constraints, information extraction, explicitly managed time dimension),
- new programming models,
- machine learning and statistical methods,
- complex systems architectures,
- information visualization.

It is important to emphasize that Big Analytics does not involve only the analysis and modelling phase because noisy context, heterogeneity, and interpretation of results are also necessary to be taken into account. All these aspects influence scalable strategies and algorithms, therefore, more effective preprocessing steps (filtering and integration) and advanced parallel computing environments are needed.

In any case the main problems of current data mining techniques applied on Big Data are related to their inadequate scalability and parallelization. Consequently, new technological tools, such as Big Data Management systems, NoSQL databases, NewSQL databases [7] support Big Data storage and processing and create now a challenge for both database specialists and people doing BI and tasks of Big Analytics. Particularly virtualization and cloud computing, are facilitating the development of platforms for more effective capture, storage and manipulation of large volumes of data.

3 Usability of Big Data Storage and Processing for Personalized Medicine

Big Data is often mentioned only in context with BI; however, not only BI developers but also e-scientists analyze large collections of data. A challenge for computer specialists or data scientists is to provide these people with tools that can efficiently perform complex analytics considering the special nature of Big Data [8]. A special challenge is provided in the area of personalized medicine. Big Data storage and processing will help

- to work with large datasets, e.g., genome data, and use them for individual healthcare processes,
- to detect correlations between people and illnesses,
- in drug repurposing,
- in adverse reaction monitoring and detection,
- progress in personalized medicine,
- in patient pre-profiling.

Authors of [9] mention other scenarios, e.g.:

- analyzing patient characteristics and the cost and outcomes of care to identify the most clinically and cost effective treatments and to offer analysis and tools, thereby influencing provider behavior,
- applying advanced analytics to patient profiles (e.g., segmentation and predictive modeling) to proactively identify individuals who would benefit from preventative care or lifestyle changes,
- broad scale disease profiling to identify predictive events and support prevention initiatives,
- collecting and publishing data on medical procedures, thus assisting patients in determining the care protocols or regimens that offer the best value,

Considering Big Analytics, the following contributions for medicine can formulated [2]:

- Machine learning software can point to abnormalities and predict health issues.
- Big Data analysis can help to move from corrective to preventive medicine.
- Doctors will increasingly rely on Big Data technology for triage, diagnosis and decision making.
- Consequently, doctors will perform better, health care costs can decrease, and patient care can improve.
- Pharmaceutical companies when screening large clouds of Big Data can find and corroborate seemingly weak correlations and target medicines to subgroups of patients with similar genetic backgrounds.

Data produced by real-time streaming data monitors is also common. The ability to perform real-time analytics against such data in motion and in context other data is a big challenge in personalized medicine and in healthcare in general.

4 Conclusions

We have described Big Data movement towards healthcare area and some opportunities how to develop new types of applications appropriate especially for personalized medicine.

The success of Big Data technologies will depend on natural language processing capabilities, pattern recognition algorithms for image and video sources, predictive modelling, and on new statistical analysis methodologies, large storage capacities in the cloud and advanced search technologies. This holds not only for BI tools used in business area, but in medicine too.

Big Data technology enables to integrate genomic and clinical data in healthcare. An important point is that the industry will be able to advance personalized medicine. As remarked in [9], ideally, individual and population data would inform each physician and his patient during the decision making process and help determine the most appropriate treatment option for that particular patient.

References

[1] T. Berners-Lee. Linked Data. `http://www.w3.org/DesignIssues/LinkedData`, 2006. Accessed: 2015-04-29.

[2] H. Broda, Big Data Trends—A Basis for Personalized Medicine, Infosys Limited, Bangalore, India,

[3] J. Kelly, Big Data: Hadoop, Business Analytics and Beyond, Wikibon, `http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond`. Accessed: 2015-04-29.

[4] J. Kozak, M. Necasky, J. Dedek, J. Klimek, J. Pokorny. Using Linked Data for Better Navigation in Summaries of Product Characteristics. In: Proc. of the 6th Int. Workshop on Semantic Web Applications and Tools for Life Sciences, Edinburgh, UK, December 10, 2013.

[5] 7. J. Kozak, M. Necasky, J. Dedek, J. Klimek, and J. Pokorny. Linked Open Data for Healthcare Professionals. In: Proc. of Int. Conference on Information Integration and Web-based Applications & Services, IIWAS '13, New York, NY, USA, 2013, ACM, pp. 400-409.

[6] J. Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, Ch. Roxburgh, A.H. Byers, Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Inst., 2011.

[7] J. Pokorny, NoSQL Databases: a step to database scalability in Web environment, Int J Web Info Syst 9(1), 2011, 69–82.

[8] J. Pokorný, P. Škoda, I. Zelinka, D. Bednárek, F. Zavoral, M. Kruliš, P. Šaloun, Big Data Movement: A Challenge in Data Processing. Chapter in Big Data in Complex Systems, A.E. Hassanien et al. (Eds.), Springer International Publishing in Switzerland, Studies in Big Data 9, pp. 29-69, 2015.

[9] W. Raghupathi, V.Raghupathi, Big data analytics in healthcare: promise and potential, Health Information Science and Systems 2:3, 2014.

[10] J.J. Thomas, K.A., Cook (Eds), A Visual Analytics Agenda, IEEE Computer Graphics and Application, 2005.

[11] J. Ward, Oncology Reimbursement in the Era of Personalized Medicine and Big Data, Journal of Oncology Practice 10(2), 2014, 83-86.

[12] W3C, Resource Description Framework (RDF), `http://www.w3.org/RDF/`. Accessed: 2015-04-29.

Keystroke Dynamics for Security Enhancement in Hospital Information Systems

Anna Schlenker[1,2], **Adam Bohunčák**[3]

[1] Department of Biomedical Informatics, Faculty of Biomedical Engineering,
Czech Technical University in Prague, Kladno, Czech Republic

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

[3] Joint Department of Biomedical Engineering Czech Technical University in Prague and Charles University in Prague,
Faculty of Biomedical Engineering, Czech Technical University in Prague, Prague, Czech Republic

Correspondence to:

Anna Schlenker
Department of Biomedical Informatics,
Faculty of Biomedical Engineering,
Czech Technical University in Prague, Kladno, Czech Republic
Address: nám. Sítná 3105, 272 01 Kladno
E–mail: schlenker.anna@gmail.com

IJBH 2015; 3(1):41–44
received: May 1, 2015
accepted: May 7, 2015
published: June 15, 2015

1 Big Data in Hospital Information Systems

In the healthcare sector amount of data has been accumulated over the years. Today it is still most of them on paper, but gradually we are moving towards computerization of health care. As a part of this eHealth implementation it is necessary to convert all paper records into electronic form too. There is, of course, involved the processing of large volumes of data from electronic health records, which will gradually keep increasing [1].

Using analysis of big data can improve the quality of health care. In addition to support decision-making can analysis of big data also help with risk-analysis or cost-analysis in healthcare. Another possibility is use in developing medical guidelines [1].

Big Data can be defined using so-called "4V definition":

1. *V=volume;* This means that the volume of data increases exponentially.
2. *V=velocity (speed);* There are jobs requiring immediate processing of large volumes of data which are continuously generated. A suitable example may be processing data produced by the camera.
3. *V=variety (variability);* Besides processing the structured data, there are jobs for processing unstructured text, but also various types of multimedia data.
4. *V=veracity (credibility);* Uncertain credibility of data due to their inconsistency, incompleteness, ambiguity etc. A suitable example can be illustrated on the communication on social networks.

2 Security Enhancement with Multifactor Authentication

The topic of data security is increasingly transferred to the field of biomedicine and health care. And in this area it is not only a medical secret as such, but with a number of other issues related to eHealth. Today workplaces working without a computer and without the information system are, thankfully, rare. Many people are becoming interested in who has access to these systems and the information in it [2].

Today, we put great emphasis on the fact that each user need their own login information and also to have set different editing rights for individual users on the system (physician, nurse, laboratory technician, radiologist, technical staff, etc.). There are also organized training sessions where health care professionals are getting familiar with the need for **multifactor security** of sensitive patient data. This is mainly to make staff understand that entering the password is not only thing that delays them from work, but also the thing that can protect them. And

that the information system is not only annoying software that do not let them continue without filling some fields, but it can be significant and this box can be important [2].

Multi-factor authentication is a security system in which more than one form of verification is used in order to prove the identity and allow access to the system. In contrast, single factor authentication involves only one form of verification, most frequently a combination of user ID and password [2].

2.1 Current State of Security in Hospital Information Systems

Currently the most of the hospital information systems remember on the safety and can set different rights for different users. A user who logs into the system, is typically verified just by one method, typically password. This password might not fulfil any of the conditions needed for password resistant from dictionary attack. For the attackers is obviously not a problem to use dictionaries of all languages. Users should, therefore, remember that their password should not consist of the name of the husband/wife, pet or other full-semantic word. Generally, the password should be long enough (at least 8 characters), composed of large and small letters, numbers and special characters. Secure password the user must not tell anyone (not even his/her husband/wife) and not write it anywhere. There is nothing easier than to copy the password that the user entered in the notebook [2].

Another mistake in most hospital information systems is that there is no automatic logoff. The reason is probably a "waste of time" health care professionals at constant logging. On the other hand, users of these systems should realize that in this case there is nothing easier than to have the system logged in and unattended, need to necessarily leave (which in the healthcare is probably not a rarity) and expose the computer and all sensitive data to the world. Likewise, in this way hospital information systems cannot prevent a user has been logged on multiple computers at the same time, which provides potential attackers same options [2].

2.2 Biometric Authentication

After realizing all the risks that entails the use of these systems, offering a variety of methods for high data security. These are mainly the **biometric characteristics** that cannot be written anywhere, they cannot be forgotten and cannot be lend to anyone. Among the most commonly used biometric characteristics include anatomical-physiological characteristics such as fingerprints and palm prints, scanning bloodstream of palm or back of the hand, hand geometry, facial recognition, retinal scan, etc. Another group is called behavioural characteristics which are yet used more in criminology. These include recognizing people by walking or voice [2, 3].

2.3 Keystroke Dynamics

Very nice compromise that offers a high level of security without unreasonable burdening of staff, is the **keystroke dynamics**. This method is a behavioural biometric characteristic that captures the dynamics of typing. This means detecting the times at which the keys were pressed with milliseconds precision. For these times, then calculating the time vector, which is composed of the keystroke duration and keystroke latency (see Figure 1).

The *keystroke duration* means the period of time a key is held for. The *keystroke latency* means the time between individual keystrokes.

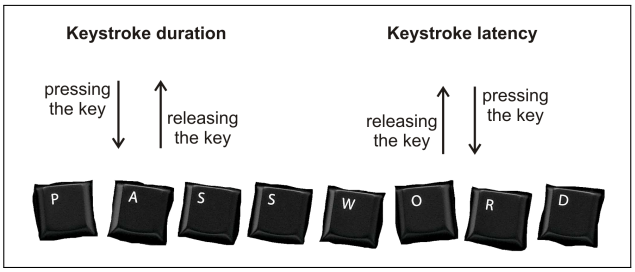


Figure 1: Keystroke duration and keystroke latency.

Keystroke verification techniques can be classified as either *static* or *continuous* [3].

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

Advantages of Keystroke Dynamics

1. The ultimate goal is ability to continually check the identity of a person as they type at a keyboard [4, 5].
2. Neither login nor verification affect the regular work flow because the user would be typing the needed text anyway. Easy to use for example with login and password during a logon process [6].
3. Unlike other biometric systems, keystroke dynamics is almost free. The only hardware required is the keyboard [4, 7].
4. Time to train the users is minimal and ease of use is very high [6].
5. Public acceptability is very high. There are no prejudices such in a case of fingerprint verification or discomfort such as retina pattern scanning [3].
6. Keystroke dynamics is ideal also for remote users.

Disadvantages of Keystroke Dynamics

1. Keystroke dynamics is a non-static biometrics like for example voice. This can change quite fast during time, also one-hand typing (due to injury), etc. can influence typing rhythm [4].
2. Low accuracy – keystroke dynamics one of the less unique biometric characteristics [6].
3. Small commercial widespread of technology [6].
4. Dependency on keyboard characteristics, for example layout of keys. Some users may be used to a full-sized keyboard, while the others may prefer to use a laptop, where the typing behaviour will probably be very different [8].
5. Typing style usually differs depending on the language (native vs. foreign) [9].

The main advantage of this method is that there is no need to purchase any additional hardware. In other biometric methods you need to buy for example fingerprint scanners or webcams to scan your face. Here it is sufficient keyboard which already is connected to the computer. This avoids the need to train staff too. A great advantage is also the possibility of a continuous record throughout the user's interaction with the device. For healthcare it is advantageous especially in time when the staff has to leave the computer suddenly and has no time for logout the software. In this case, the software will recognize that the stranger working with system and software will be locked.

2.4 Mouse Dynamics

Another option to further increase the security of hospital information system, is collateral use of keystroke dynamics and mouse dynamics. **Mouse dynamics** is behavioural biometric characteristic, tracking the movements of the mouse during user's interaction with the device. Here are measured four types of actions: the general movement of the mouse, drag and drop, point and click, and silence.

3 Application in Biomedicine

We have designed an application that uses multi-factor security (user name + password + keystroke dynamics when entering a password). Its advantage is the data collection directly from the operation system of the computer, without any delay. The application records the key code, key name, press time and release time automatically. Capturing starts by pressing the 'Start' button and ends with the 'Stop' button (see Fig. 2). After the recording and pressing the 'Save' button, all the data will be exported to a CSV file. All parameters needed for the keystroke dynamics can be calculated from these data. The application is programmed in C♯ and runs on any computer (running Windows with .NET Framework Version 4.0).

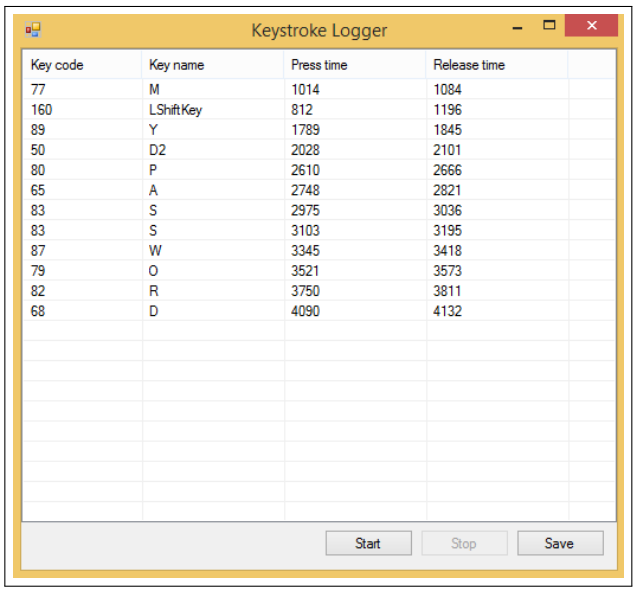


Figure 2: Application records the parameters for keystroke dynamics.

The application was implemented in C♯ within the Microsoft Visual Studio Express 2012. When creating the application class globalKeyboardHook.cs [10] was modified and used. This modified class addresses global interception keys. It is possible to get a code of the original class under The Code Project Open License (CPOL) [11]. The use of this class can store the exact time of pressing and releasing of the keys. Of course, it scans all the keys on the keyboard and also the possibility of shooting more keys simultaneously (see Fig. 3). Fig. 3 shows data exported to the CSV file.

	A	B	C	D	E	F	G
1	75	K	7229	7832			
2	76	L	7471	7833			
3	160	LShiftKey	6431	7948			
4	65	A	7677	7966			
5	68	D	9482	9572			
6	82	R	9828	9857			
7	85	U	9922	9978			
8	66	B	10235	10300			
9	89	Y	10477	10528			
10	48	D0	13415	13466			
11	57	D9	13588	13635			
12	187	Oemplus	15199	15237			
13	221	Oem6	15376	15414			
14	113	F2	18093	18180			
15	114	F3	18309	18374			
16	91	LWin	19351	19407			
17							

Figure 3: Data exported to a CSV file.

At the beginning of the file the 'PressedKeys' class is introduced. It contains variables 'Code' (type Keys (specify key codes and modifiers)), 'Pressed' (type long) and

'Released' (type long). Subsequently, all variables (including stops) are nulled. Then when you press the 'Start' key it allows scanning values which are stored in the appropriate variables. After pressing the 'Stop' key the scanning of values is stopped and the option to save the scanned data to a file is allowed using the 'Save' key.

The user interface consists of four columns (see Fig. 2). The first column shows the code of the key, the second the name of the key, the third the time of pressing the key and the fourth the time of releasing the key. The interface has at its top right standard buttons for minimizing and maximizing windows and also a button to close the application. This button also stops scanning keys (i.e. the button 'Stop' is pressed).

Right at the bottom of the user interface there are three buttons:

1. The 'Start' button starts scanning (key codes and times of pressing and releasing the key). After repressing the 'Start' (it is active again until after pressing the 'Stop' key), the application memory is empty (i.e. recording starts from the beginning).

2. The 'Stop' button stops scanning. The button is only active when the application starts (it does not make sense to stop something that is not running).

3. The 'Save' button saves the recorded data to a file with the suffix CSV. In the case when during runtime were not pressed any keys on the keyboard, this button is not active. This prevents the possibility of imposing an empty file.

3.1 Experiment

This application we tested in a pilot study on a group of 32 people. There are 10 men and 22 women. Each of them was asked to type the word *kladruby*, which is a name of a town in the Czech Republic.

The data include 15 variables, namely the time of 8 keystroke durations for the letters (K, L, A, D, R, U, B, Y) and 7 corresponding keystroke latencies (K-L, L-A, ..., B-Y). We performed several analyses with the aim to learn the classification rule allowing to assign the data vector with 15 variables of an unknown person to a particular individual from the database of 32 probands.

3.2 Results

The results show that it is possible to use this characteristic to distinguish the individual user with an accuracy of 93 %.

In our research, now we continue with testing the continuous keystroke dynamics. Proposed study involves the detection of users who are established in the database followed by deadlock software when detecting fake users. Another possibility is the use of the software to detect and automatically assign users to a written text without need to login.

The next stage will be, of course, modification of our application for use on touch devices (tablets and smart phones) that are part of mobile medical units (ambulances).

Acknowledgements

This work has been supported by the specific research project no. 260158 "Semantic Interoperability in Biomedicine and Health Care", Charles University in Prague and by the research project no. 494/2013 "Identification of users by keystroke dynamics", CESNET Development Fund.

References

[1] Raghupathi W., Raghupathi V.: Big data analytics in healthcare: promise and potential. Health Information Science and Systems 2014, 2(3). doi:10.1186/2047-2501-2-3

[2] Schlenker A.: Multifactor Data Security in Information Systems in Health Care. International Journal on Biomedicine and Healthcare 2014; 2(1):25-27

[3] Schlenker A., Šárek M.: Behavioural Biometrics for Multi-Factor Authentication in Biomedicine. European Journal for Biomedical Informatics. 2012; 8(5):19-24. ISSN 1801-5603

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

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

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

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

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

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

[10] A Simple C♯ Global Low Level Keyboard Hook [Internet] 2007 [cited 2013 Aug 28]. Available from: http://www.codeproject.com/Articles/19004/A-Simple-C-Global-Low-Level-Keyboard-Hook

[11] The Code Project Open License (CPOL) 1.02 [Internet] 2008 [cited 2013 Aug 28]. Available from: http://www.codeproject.com/info/cpol10.aspx

Methods for Better Health: Big Data, Personal Health and More

Rolf Engelbrecht[1], **Lorraine Nicholson**[2]

[1] ProRec-DE, Ismaning, Germany

[2] International Federation of Health Information Management Associations, Manchester, United Kingdom

Abstract

Information is an important factor in all business areas. Health care is producing data from many different sources. Individual patient data are used in many ways and for different purposes including statistical analysis and presentation. Epidemiological studies lead to new knowledge end medical evidence for diagnosis and therapy. The amount of healthcare data is increasing and new methods are needed for interpretation of these large quantities of data. Strategies and methods used in other businesses have to be analysed, possibly adapted, tested and applied. Developments in patient centred medicine and personal medicine offer new treatment possibilities. The relationships to develoments in medical records, big data and health information management are important. Analysis of these data may result in a new way healthcare is provided.

Keywords

health informatics, medical informatics, health information management, big data, big data analysis, EFMI working group, EFMI-HIME, EuroRec quality criteria, IFHIMA, evidence-based medicine, electronic health records, EHR

Correspondence to:

Rolf Engelbrecht
ProRec Germany
Address: Glaslweg 33, D 85737 Ismaning
E–mail: Engelbrecht@ProRec-DE.org

IJBH 2015; 3(1):45–51
received: May 25, 2015
accepted: June 2, 2015
published: June 15, 2015

1 Introduction

In health care many disciplines work together and generate data and information about their interventions on individual patients. Good practices in managing this information can promote and support the delivery of high-quality health care. The European Federation for Medical Informatics (EFMI) has set up the working group HIME (Health Information Management in Europe) for the management of health information for this purpose.

Health Information Management (HIM) is the practice of maintenance and care of health records by traditional (paper-based) and electronic means in hospitals, physician's office clinics, health departments, health insurance companies, and other facilities that provide health care or maintenance of health records. Effective sharing of patient information using paper-based systems is difficult.

The appropriate exchange of health information will be the foundation of eHIM practices for the future. There will be new and effective way of exchanging information through interoperable Health Information Exchange (HIE) networks and it will be important to mobilise healthcare information electronically across organizations within a region or community - a standards-based EHR is the foundation on which HIE will be built. The role of the Health Information Manager is the management of Information in Health and the EHR will unite them professionally. With data sharing amongst multiple providers of care there will be privacy concerns regarding patient records, mainly related to confidentiality of data. Standards and definitions must be standardised to ensure that information is shared effectively and securely between providers and standardisation of terminology will be crucial to facilitate effective communication, information exchange and data sharing. There must be a sound infrastructure for data sharing utilising data sharing protocols, codes of practice etc.

For this to become a reality a good education for health information managers will be crucial together with effective training for their staff. Institutions that offer good quality professional education including schools, colleges and universities therefore need to develop programs for small and medium healthcare enterprises. This requirement could offer an important development area for the use of telemedicine complemented by local activities to deliver high quality healthcare in an effective and cost-efficient way.

Big data is growing in importance. Volume and variety of data in organisations is growing constantly. This tremendous growth means that for operating a health care provider must understand big data in order to select the information that raelly counts, but the provider also must understand the possibilities of big data analytics and available tools and services.

Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch. [5]

2 Big Data and its Analysis

Data processing in health care started like in other business areas with administrative solutions. Structured data were stored in flat sequential data sets on tapes or small disk volumes. A next step in the evolution was the use of data base management systems (DBMS) using tables in hierarchical manner or relational way. The amount of data in health care was extended to medical data also.

Edgar F. Codd developed 1970 the relational model [3]. His motivation was to ma a clear boundary between logical and physical aspects of database management for better data manipulation, storage and retrieval. This simple model enabled programmers with a common understanding of the data and its relation to develop complex systems for big amounts of data. This was done by different teams. Data became communicable. Data were collected frm various legacy systems. A high level query syntax allowed users to work on big chunks of data mostly for manipulation/update and analysis. Data mining and statistics on big amount of data from different sources and multiple locations, e.g. several hospitals; became possible. Major problem in medicine were unstructured data. There were some solutions in coding this information for processing. But still big data volumes in health care were not accessible.

A new dimension started with the internet and its use by many persons with different backgrounds. Unstructured content found its place in the web. Information retrieval and extraction has to be made on this basis which was enabled by the dramatic increase of computer power and online storage. Specific search engines and analysis tools were developed for medical content, for social media and also for temporal analysis. The use of natural language was possible for a broad scope of users. Content and text analysis entered also new horizons. Data are stored locally but also in clouds in the web. Big data has a platform which is growing.

A new wave of data is coming from patients with access to social media and mobile devices especially sensor based monitoring data in chronical care. Again it will be sets of structured and text data. Data contain rich customer opinion and behavioral information. Classification and coding will play not a big role. Discussion will be on security of data and privacy.

There will be different problems in the future in amount and heterogenity of data. We need solutions to generate good to high quality of results from analysis. This could and should be used also for personalised medicine.

There applications in general and in health-related areas. Sometimes it is more understandable to look into applications of other areas, e.g. commercial business.

Macy's is a chain of big department stores. In addition to that, "Macy's gathers, and of course analyses, a vast amount of customer data ranging from visit frequencies and sales to style preferences and online offline personal motivations. They use this data to create a personalized customer experience including customized incentives at checkouts. Even more, they are now capable of sending hyper-targeted direct mailings to their customers, including 500.000 unique versions of a single mailing" [4]. Maybe we can derive principle for handling big data for analysis from it.

3 Personalised Health vs Patient Centred Care

Healthcare undergoes like other areas a continuous evolution with some revolution. New evidence and new technologies are the driving forces. The patient and his/her disease was always in the focus; healthcare was and is patient-oriented. Regimen for treating different diseases were developed by the specialist physicians. New technologies like Roentgen bulb and other image producing techniques have changed the diagnostics and treatment. In the last years digital imaging uses the power of computers for high quality images.

Quality enhancement and cost reduction are often goals which are not reachable in one solution. So quality was and is the topic for national and international initiatives. In Germany the bIT4health concept has developed a health telematics infrastructure. The report "Crossing the Quality Chasm: A New Health System for the 21st Century" [14] is the second and final report of the Committee on the Quality of Health Care in America, which was appointed in 1998 to identify strategies for achieving a substantial improvement in the quality of health care delivered to Americans. A workshop on "Using Information Technology to Improve the Quality of Care" gave many suggestions for a redesign of the healthcare system for the 21st century. Among this was the patient-centered approach which focusses on providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions. It offers also a set of 10 new rules to guide patient-clinician relationships.

It is a big step to personalised medicine which complements the patient-centered care. In Europe the

EAPM (European Alliance for Personalised Medicine)-REPORT on Innovation and Patient Access to Personalised Medicine summarised the actual status and brings together healthcare experts and patient advocates on major chronic diseases to improve patient care by accelerating the development, delivery and uptake of personalised medicine and diagnostics. [15]

EU-Commissioner for Health and Consumer Affairs Tonio Borg explained the EU-position: "Personalised medicine is a promising concept. As patients are divided into groups based on their individual, biological, genetic and genomic characteristics, medical interventions are tailored to those patients' needs. Hence, this new approach can help reduce the risk of undesirable adverse reactions, and at the same time make medicine more effective. Personalised medicine is an innovative, efficient and patient-centred alternative to the one-size-fits-all medicine. And it also yields a maximum return on healthcare investment - a valuable argument for decisionmakers in times of austerity. "

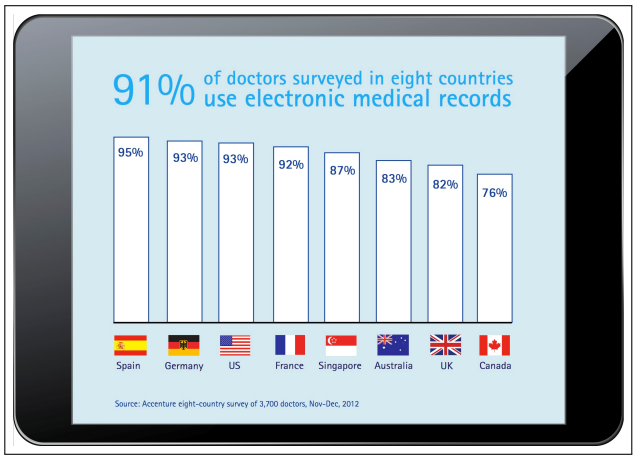


Figure 1: Survey of physicians' use of EMRs.

Early studies have shown that the effectivity of drugs is in some cases rather low, e.g. 2001 in cancer drugs sometimes for 75% of the patients a drug is ineffective [6]. To use information about patient's genome and the recommended drug can avoid spending money, raise the quality and lower the harm for the patient. The number of "Prominent examples of personalized medicine drugs, treatments and diagnostics products available" has increased from 13 (2001) to 113 (2014) [7]. The number of members of the Personalized Medicine Coalition (PMC) has increased from 20 (2004) to 230 (2014). The costs for sequencing a genome has dropped to $1000 and will come down further.

The use of digital genome data and evidence data needs appropriate IT equipment. Nearly all hospitals in Europe and up to 95% private practitioners in Europe are using computer systems for their daily work.

There was an online survey conducted of 3,700 physicians across eight countries: Australia, Canada, England, France, Germany, Singapore, Spain and the United States. [16] The survey included 500 doctors per country (200 from Singapore) and assessed the physicians' adoption, utilization and attitudes toward healthcare IT.The results are displayed in figure 1 and figure 2 and summarised : The majority of doctors in all countries reported that EMR and HIE have had a positive impact on their practice, such as reducing medical errors (76 percent) and improving the quality of data for clinical research (74 percent). However, U.S. doctors were the least likely (38 percent) to report that using EMR and HIE reduced their organization's costs. They also said that cost was the single greatest barrier to technology adoption.

In combination with personalised medicine by tailoring healthcare solutions to the individual patient and delivering the right treatment at the right time there should be another positive effect of using health IT. This should be monitored by independent groups and a strategy should be developed from the medical standpoint.

National projects and actions will play a major role such as the NHS plan described by the NHS- National Information Board (November 2014) (Report: Personalised Health and Care 2020 Using Data and Technology to Transform Outcomes for Patients and Citizens A Framework for Action) The goal is better use of data and technology has the power to improve health, transforming the quality and reducing the cost of health and care services.

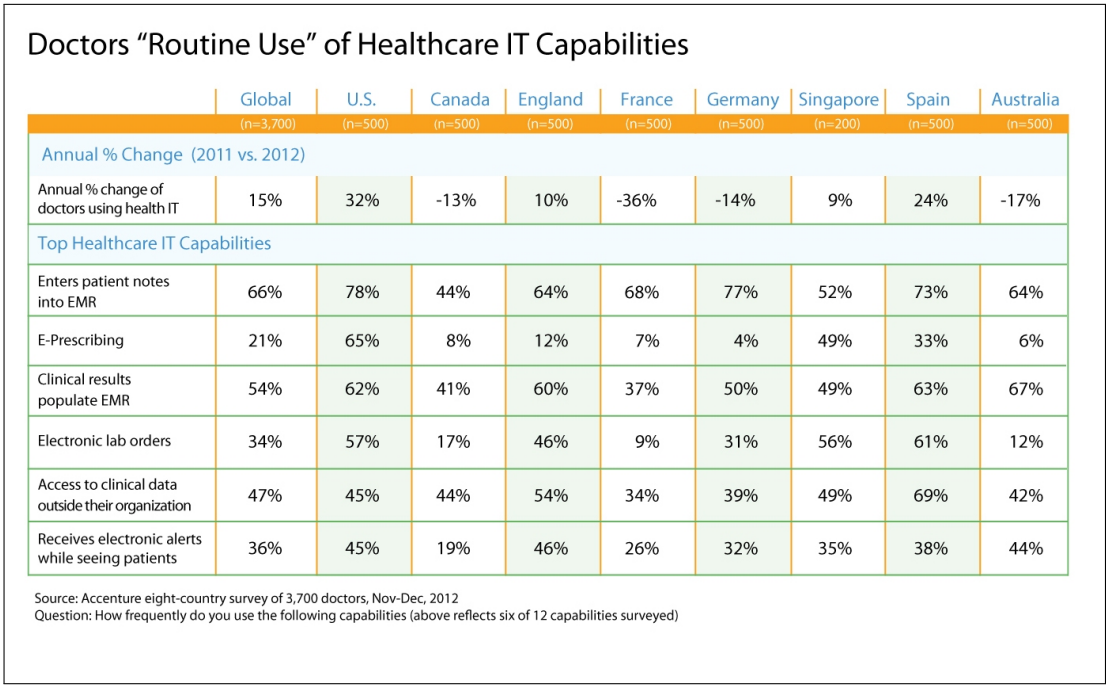
4 European Activities in Quality of Health Data

The European Institute for Electronic Health Records EuroRec [8], founded as an institute in 2003, has set itself the goal of supporting high quality design, implementation and adaptation of existing medical information systems (EHR systems). In several EU projects, both the structures of the EuroRec Institute as well as quality criteria for EHR systems were developed.

EuroRec was registered in France and now has its "office" in Ghent, near Brussels, which is ideal with close proximity to the European institutions. In EuroRec projects and other activities almost all 16 national ProRec centres are involved. The national centres usually undertake their own activities and/or participate in national or regional activities.

A focus in recent years has been the QREC project with 26 partners, which developed quality criteria for EHR systems for use in the development, evaluation and certification of systems. There are more than 1,000 criteria, which are a narrative description of properties of the system and come from many sources, such as from the American and Canadian certification, manuals of European systems and other publications. The source is indicated for all criteria, which are grouped and indexed according to usage and they are also suitable for training [9].

EuroRec was and is member of several RD projects, e.g. EHR4CR (Electronic Health Records for Clinical Re-



Doctors "Routine Use" of Healthcare IT Capabilities

	Global (n=3,700)	U.S. (n=500)	Canada (n=500)	England (n=500)	France (n=500)	Germany (n=500)	Singapore (n=200)	Spain (n=500)	Australia (n=500)
Annual % Change (2011 vs. 2012)									
Annual % change of doctors using health IT	15%	32%	-13%	10%	-36%	-14%	9%	24%	-17%
Top Healthcare IT Capabilities									
Enters patient notes into EMR	66%	78%	44%	64%	68%	77%	52%	73%	64%
E-Prescribing	21%	65%	8%	12%	7%	4%	49%	33%	6%
Clinical results populate EMR	54%	62%	41%	60%	37%	50%	49%	63%	67%
Electronic lab orders	34%	57%	17%	46%	9%	31%	56%	61%	12%
Access to clinical data outside their organization	47%	45%	44%	54%	34%	39%	49%	69%	42%
Receives electronic alerts while seeing patients	36%	45%	19%	46%	26%	32%	35%	38%	44%

Source: Accenture eight-country survey of 3,700 doctors, Nov-Dec, 2012
Question: How frequently do you use the following capabilities (above reflects six of 12 capabilities surveyed)

Figure 2: Survey of physicians' use of health care IT.

search. The future will show the relevance of this project for healthcare and healthcare-IT solutions.

5 Health Information Management

Medical Records development started from medical documentation and medical record management. There were many efforts in organising paper-based medical records structures and storing in archives in hospitals. This started in the 1970s with a focus on availability of data from thousands of patients for routine use to derive knowledge and undertake other research. The amount of data from many sources, both inside and outside the hospital, was tremendous. At the same time electronic hospital information systems were developed initially from administrative functions. The move from conventional health information management (HIM) towards electronic health information management (eHIM) also started and is still ongoing.

Health Information Management (HIM) is the practice of maintenance and care of health records by traditional (paper-based) and electronic means in hospitals, physician's office clinics, GP surgeries, health departments, health insurance companies, and other facilities that provide health care or maintenance of health records. The effective sharing of patient information to facilitate care delivery from multiple providers (an integrated care model) the use of traditional paper-based systems is difficult. eHealth integrated care models require the sharing of data from multiple sources, each holding an electronic record for the patient. These records must be brought together to eliminate "silos" of information to facilitate the delivery of high quality, safe and effective care in the patient's home environment. Therefore HIM will need to become eHIM to support the Patient-Centred Approach and in order to offer HIM expertise to Health Information Technology (HIT), the profession must train more HIM professionals in both traditional and emerging practice.

In 2003 the e-HIM Task Force established by the American Health Information Management Association (AHIMA) identified seven major factors affecting healthcare: rising costs, an ageing and mobile population, a lack of data standards, growth of technology, shrinking HIM work force, the need for consumer education, and changing public imperatives. All seven remained central when the e-HIM agenda was revisited in 2005, but former task force members singled out three factors in the forefront at that time: work force, technology, and data standards. These three factors are still in the forefront today. [11]

eHealth is concerned with promoting, empowering and facilitating health and wellbeing with individuals, families and communities, and the enhancement of professional practice through the use of information management and information and communication technology (ICT). eHealth is not just technology - it is about finding, using, recording, managing, and transmitting information to support health care, in particular to make decisions about patient care. Computers (and other ICT devices) are merely the technology that enables this to happen.

e-Health is so much wider than just hospitals – heath care, social care, health promotion, disease prevention, community health, primary care, social support, risk management, disease monitoring/surveillance, citizen support for assisted living, bio-medical monitoring devices, m-Health and tele-Health. Each health and social care professional involved in multi-disciplinary support and care

for citizens creates a record of their interaction and/or intervention with the citizen all of which need to be brought together to form a holistic record of care and support to the individual (patient centred care). Any gaps or omissions in recording could have serious treatment and safety implications for the individual. All records need to be coordinated along with the informal support and care provided by family, friends and neighbours and also actions taken by the "empowered patient" as they take responsibility for their own health and well-being e.g. by use of remote disease monitoring.

There is an urgent need to identify the emerging/developing role of Health Information Managers in the integrated, patient-centred care model (e-HIM's). Records still need to be managed, probably more than ever with records for individual patients being generated by multiple systems! The role could include the health information governance issues including data integrity, data sharing, liability (particularly in respect of tele-health delivery), the legal framework for tele-health, the 'legal health record', privacy/confidentiality and Electronic Health Record retention and disposal. Transparency and Governance are essential pre-requisites for patient centred e-Health delivery. There is also a need to collect, analyse and present health information for the "evidence base" for the new ways of working and care delivery [12].

Health information is generated by applying knowledge to data. Electronic processing of data and electronic processing of knowledge is not necessary but it is helpful and widely used. Electronic health information management (eHIM) is the method of the future for good quality health care. Principles for best practice are described by Sallyanne Wissmann [1] and reflect all parts of health information management (HIM):

- Data have to be defined clearly and understandably for all users. The best way is to use standards such as ISO EN13606. This also makes interoperability between systems possible and easier . In the European context multilinguality is necessary for the mobility of patients.
- Adequate visualization of data and information is a key factor for their use.
- Quality criteria are important for the planning, development and use of health information systems with regard to gathering, storage, processing and retrieval of data, information and knowledge.
- Without education and training of users, health professionals and patients, best systems will not work in a high quality mode.
- Data security and privacy are prerequisites for eHIM
- Availability of health information at the right place, the right time and the right format is crucial to ensure safe and effective care and treatment for the patient.

Top 10 Trends Impacting HIM in 2016 and an Emerging Role for eHealth Information Managers [2] are:

1. Clinical and business process leaders increasingly need to own the EHR and other technologies in order for these technologies to be successful.
2. Some level of Information Management is a basic competence for most who work in healthcare.
3. HIM functions are distributed and embedded throughout organizations, with greater focus on support of patient care and population health mission.
4. There is greater recognition of the importance of managing the Records Management/Information Management aspects of digital information through its life cycle. Critical functions will include data integrity, legal health record, e-discovery, privacy and access and authentication management.
5. Plans for broad payment reform are coalescing as a result of risk- and outcomes-based payment pilots and demonstrations under the 5-year-old Affordable Care Act.
6. Health systems continue to work at reducing overall costs by 20% to remain financially viable.
7. The linkage between improvements in quality and improvements in financial performance is well documented. Health systems continue to work at reducing overall costs by 20% to remain financially viable.
8. An increasing number of people rely on technology and information to assist in self-management and select providers who deliver cost-effective care.
9. Clinicians require and use tools and information to anticipate the outcomes and cost consequences of their clinical decisions at the point of care.
10. The design of ICD-11 is being evaluated, and planning for implementation is projected for 2020.

These "Top 10 Trends Impacting HIM in 2016" address the issues relating to the wider healthcare agenda in particular Trend No: 4 regarding the importance of the Records Management/Information Management aspects of managing digital information throughout its life cycle.

The International Federation of Health Information Management Associations (IFHIMA) supports national associations and individual health information management (HIM) professionals to implement and improve health information management operations in their own countries and the systems which support the discipline. IFHIMA was established in 1968 to bring together national organizations committed to improvement in the use of health records in their countries. The founding organizations recognized the need for an international organization to serve as a forum for the exchange of information

relating to health information management and information technology.

The purposes of IFHIMA are:

- promote the development and use of HIM in all countries
- advance the development and use of international HIM standards
- provide for the exchange of information on HIM education requirements and learning programs
- provide opportunities for education and communication between persons working in the field of health records/information in all countries
- promote the use of technology and the electronic health record

IFHIMA is a non-profit organization in official relations with the World Health Organization (WHO) as a recognized non-governmental organization (NGO). The Federation sends representatives to WHO meetings and works closely with WHO on specific projects of particular concern or interest to WHO in the field of health records and information systems. Currently IFHIMA is working on a project with the WHO Family of International Classifications to develop and promote an international training strategy and explore development of an international certification strategy for coders.

IFHIMA [10] represents the interests of national member associations globally. This includes joint working with other international organisations such as IMIA (International Medical Informatics Association). IFHIMA provides education modules in basic health records practice which are available to download free of charge from the IFHIMA website by practitioners working in the field of medical/health records and HIM (Health Information Management). Members also have the opportunity to serve on international committees, task groups or projects, which focus on health information management, health records, and the HIM profession.

6 Coding and Classification

Codes are the classical way for harmonisation of medical data. Many of them are available internationally. Mostly used is ICD (International Classification of Diseases) which version 10 is worldwide in use in all levels of care. Originator is WHO. National organisations are translating and distributing it, e.g. DIMDI(German institute for medical documentation and information) in Cologne providing the German version. ICD is also available in modifications for specific purposes. Version 11 is under development. The ICD code is abstracted from medical histories, stored in the EHRs and it is used mostly for statistics and legal purposes. It can be used also for communication between actors in health care.

Other specific are available for specific purposes. e.g. DRG(Diagnosis Related Group), LOINC (Logical Observation Identifiers Names and Codes, used for tests, measurements and observations), OPS (Operations and Procedures), SNOMED (Systemized Nomenclature of Medicine), ABDA (German drug classification). The granularity of the different systems is very different and should not discussed here. It is important to know that in most cases there is an information reduction from free text. Expanding the code to full text does not reflect the input text totally. In some cases the code is entry for more information, e.g. ABDA (German drug classification) code allows access to additional information like drug interactions, drug ingredients.

7 Conclusion

This paper presents some thoughts on big data and personalised medicine. There is a strong demand for new methods to handle both the use of big data analytics for new ways in person centered health care and the exchange of data generated from personalised diagnosis and therapy. Mobile health and new devices will play a major role. It will be necessary to develop ways to integrate data from different sources. This will lead to a huge information source for secondary use of data including social data such as life style.

The new possibilities of are a challenge for big surveys with more than 1 million participants in US and health information exchange (HIE) between hospitals, regional health information networks and private practitioners. A first step is done by the ICA (Informatics corporaton of America) using methods developed by the Vanderbilt university medical centers from 2000 on and available under the name of CareAlign. This software was developed "to improve efficiency and communication processes in order to deliver cohesive care across the medical center and its affiliated clinics and physicians' practice" [13]. It creates a unified electronic medical record by uniting all the clinical systems and processes, creating an easily accessible database of patient information for clinicians. Furthermore ICA offers workflow tools to improve communications, capture data, track clinical metrics and follow results.

It will be necessary to analyse tools from different organisations and see how far they are able to develop a longitudinal patient record. Special emphasis will be necessary for applying European privacy and data security law following the EU-directives and its implementations in member countries. ID management and accurate patient identification are to be considered under these circumstances when extended use of data and its misuse is under global discussion.

To use and understand the tools, its possibilities and broad scope of results a wide range of opportunities for education and training in telemedicine and partly with the help of telemedicine is necessary. Therefore, this con-

tribution is perhaps a start for a discussion of another commitment in the ongoing research and development, education and training for a better healthcare through the application of medical informatics and medical/health information management systems and processes.

References

[1] Wissmann, Sallyanne, Health information needs best practices, Hospital Aged care, November 2011, p. 21 http://www.himaa.org.au/Public/HA_HIMSS_AP_2011_Sallyanne_Wissman.pdf

[2] Kloss, Linda L., Health Information Management In 2016, Precyse Solutions, www.precyse.com

[3] Codd, E. F. A relational model of data for large shared data banks. Comm. ACM 13, 6 (June 1970), 377-387.

[4] https://datafloq.com/read/macys-changing-shopping-experience-big-data-analyt/286 Last access: 15 05 2015

[5] http://www.sas.com/en_us/insights/analytics/big-data-analytics.html Last access: 14 05 2015

[6] The Case for Personalized Medicine, 3rd edition, 2011 http://www.personalizedmedicinebulletin.com/wp-content/uploads/sites/205/2011/11/Case_for_PM_3rd_edition1.pdf Last access: 23 05 2015

[7] Personalized medicine by the numbers, 2014, http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/pmc_personalized_medicine_by_the_numbers.pdf Last access: 23 05 2015

[8] EuroRec Institute: http://www.eurorec.org/, Last access: 23 05 2015

[9] EuroRec-QREC Project: http://www.eurorec.org/RD/pastProject_Q-REC.cfm, Last access: 02 06 2015

[10] IFHIMA: www.IFHIMA.org, Last access: 26 05 2015

[11] Bloomrosen, Meryl. "e-HIM: From Vision to Reality." Journal of AHIMA 76, no.9 (October 2005): 36-41

[12] Nicholson, Lorraine, The Transition from Health Information Management to eHealth Information Management to Support eHealth and the Patient-Centred Approach, IFHIMA Global News, Edition 12,2013 Pages 18-22 https://ifhima.files.wordpress.com/2014/03/global_news_2013-issue_12-january_2013.pdf

[13] Wicklund, Eric, ICA seeks its niche in health care IT, 2007,Healthcare IT news, 36-37 http://www.nxtbook.com/nxtbooks/medtech/hitn0107/ http://www.healthcareitnews.com/news/ica-seeks-its-niche-healthcare-it Last access: 01 06 2015

[14] Institute of Medicine, Crossing the Quality Chasm, http://www.nap.edu/openbook.php?isbn=0309072808

[15] European Alliance for Personalised Medicine-Report, 2013 http://euapm.eu/wp-content/uploads/2012/07/EAPM-REPORT-on-Innovation-and-Patient-Access-to-Personalised-Medicine.pdf

[16] The digital doctor is "In", http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Digital-Doctor-Is-In-UK.pdf, Last access: 31 05 2015

