New Generation Advanced Analytics Tools in Medical Systems

Vesselin Evgueniev Gueorguiev
Ivan Evgeniev Ivanov
Technical University Sofia, TUS
Sofia, Bulgaria
e-mail: veg@tu-sofia.bg, iei@tu-sofia.bg

Desislava Valentinova Georgieva
New Bulgarian University, NBU
Sofia, Bulgaria
e-mail: dvelcheva@nbu.bg

***Abstract*—Research on the opportunities to create tools for a new generation of advanced analysis is considered to be the most promising direction for investigation, which will enable the successful functioning of healthcare, for better treatment of patients at lower cost. So, today these tools are in the priority group. The paper presents a new project for defining a new generation of advanced analytics tools requirements, constraints, and tasks in medical systems.**

Keywords-e-Health; advanced analyses in medical systems; medical informatics.

I. INTRODUCTION

Since 2005, there has been an explosion of e-Health scientific publications and new surveys and new strategies funded by governments, as well as expansion of the themes within the framework of the European Union at the national and multinational level. At the same time the number of healthcare topics funded by the European Union has grown rapidly.

According to the European Commission e-Health Taskforce report 2007 [6], by the end of 2007, over 10% of jobs in Europe were in the healthcare sector and this sector generated over 9% of the EU gross revenue. According to COCIR, in 2010, the gross income of the companies of the e-Health sector amounted to € 2.5 billion and until the end of 2015, it is expected to rise to € 2.7 billion [8].

In the initial period (2007-2013), the e-Health researches were focused on a wide range of topics: from medical sensors and safety of the technologies to the legal aspects of the protection of medical data and information. The reason for this was the imbalance between the available IT infrastructure and its application - in 2005 Véronique Lessens [9] showed that only 2.3% of the hospitals had decision-making systems and only 18.7% had sustaining banks of clinical results and prescriptions. At the end of 2010, the cost of healthcare services increased and the main reason was the increase of prices for personnel and the increase of use of these services without substantial increase in the quality and accessibility of healthcare. This has set the objective to overcome the obstacles and after 2012 the European Commission [7] set new tasks and activities for research, development and deployment of a next-generation advanced analytics tools in medical systems. These will assist both the processes of administrative management of medical activities and the existing clinical practices.

The necessity and usefulness of the research for a new generation of advanced analytics tools in medical systems has been discussed at length in the last 2-3 years. Very indicative of this are the analyses of the IBM Institute for Business Value [1], Deloitte [2], McKinsey & Company [3][4], and Markets and Markets [5]. Other analyses are presented in the works of Nikolova at al. [10] and Tcharaktchiev at al. [11]. Some of these analyses together with the research of the archetypes applications for medical purposes [12] have led to the present research.

This paper presents the objectives of a new project, oriented to advanced analytics in medical systems. The aim of the project is to conduct an extensive research in the field of e-Health: a comprehensive study of the concepts and methods for a new generation of advanced analytics in medical systems. Research on the opportunities to create tools for a new generation of advanced analytics is considered to be the most promising direction of investigation, which will enable the successful functioning of healthcare, for better treatment of patients at lower cost. So, today they are in the priority objectives group. According to the available data bases, the first medical areas to be targeted are endocrinology (diabetes) and pulmonology (COPD).

The project objectives include: a research of the sources of medical and biological data; a research of the possibility to use information/knowledge about illnesses and their treatment in patients with similar symptoms; structuring the medical data and information following requirements for the methods of advanced analyses.

To achieve the objectives we will conduct research, develop and evaluate architectures of medical systems; new methods for processing, storage and modification over time of medical and biological data / information / knowledge will be developed; changes in the ways of gathering and integration of information in real time will be effected; applicability of mobile agents for semantic search across heterogeneous and distributed sources will be studied and developed; automatic and semi-automatic methods for imaging data structuring will be introduced.

The present paper is structured as follows: Section II presents the scientific objectives of the project; Section III presents the available current results; and Section IV is the conclusion.

 ISBN: 978-1-61208-384-1

II. THE SCIENTIFIC OBJECTIVESOF THE PROJECT

The analyses of the development of the Healthcare Industry show the expectations of the society for higher service quality, better results and lower cost. This poses a number of serious challenges to this sector, since increasing expectations contrast with the evident critical shortage of resources. Considering the ongoing ageing of the population, each year the use of these resources increases on a daily basis. At the same time there is an increase of the number of chronically sick people, thus additionally limiting the available healthcare resources. Those challenges have led to the development of various approaches and solutions during recent years, the most widely spread being e-Health and Telemedicine. In the early phases of integration of those systems into the existing medical information systems the results have shown improvements in the performance indicators of the health system. These advantages have reduced significantly with the development of medical hardware/apparatus and the appearance of new approaches for patient treatment and tracking. This brings the issue of new paradigms for work with medical information. One of the most promising sphere of research is considered to be the exploration of the possibilities to create a new generation of advanced analytics tools which will allow healthcare to function successfully – the expectations are to improve the balance between the demands and expectations of patients and society, to optimize the use of existing resources, and to increase the ability to respond adequately to changes in medical systems and practices. In order to achieve this, the tools for advanced analytics have to be able to use the increasingly broadening range of information about the patient, thus allowing a much earlier medical and administrative intervention.

The existing academic studies focus only on individual aspects of the problem and most often they only cover a narrowly defined field of application. Corporate developments are oriented towards the possibility of renewal and development of old company systems, through integrating new approaches to collection, unification, search and processing of data, information retrieval and generation of knowledge. The most frequently reported result [2] is the identified impossibility to introduce substantially new tools for advanced analytics due to outdated design of systems of older generations.

The attempt for integration of databases in widely different medical fields, having different requirements for the examination and accompanying the patients, is also an uncommented element of such classes of systems and will represent an innovation.

A. *Heterogeneity, distribution, variability and interoperability of the data issue*

The existence of multiple databases, storing a variety of information, raises the issues of heterogeneity, interoperability, complex data structures, and integration.

The data are derived from various sources: internal (electronic health records, clinical systems for decision making, etc.) and external (laboratories, pharmacies, insurance companies, etc.). The data are in multiple formats (flat files, relational tables, text files, etc.) and they come from various geographical locations. Nowadays data sources include: data from websites and dedicated servers, social networking and blogs; remote sensors and measuring devices; invoices related to health care (both in unstructured or semi-structured formats); biometric data (medical images, blood pressure, etc.), unstructured and semi-structured data such as electronic health records, annotations, medical prescriptions, e-mails and paper documents. These are data with an extremely high degree of heterogeneity in respect to the type of the used data model, as well as the incompatible formats and nomenclatures of the values.

At the same time the data are highly decentralized, with a high degree of terminological variations, records specifics, data presentation formats and applications. This in turn is associated with problems when conducting manual search for specific data or information.

B. *The retrieval of semantic information from textual medical data*

Semantics is a science which studies the meaning or relationship of words, phrases or symbols. This determines the priority research on its use in the modern information systems, designed to work with large amounts of medical or biological data. This includes a research on the possibilities for access, capture, storage, search, sharing, transfer, analysis, and visualization of data, according to their size, velocity, variety and value.

The current systems work with very different medical or biological data. From the semantic point of view the most common operations are search and interpretation of big data. This is the lowest level of use of the semantic systems, leading to retrieval of information from data. The main problem is that the information is not generated to be appropriate for its end-user (the problems identified by the user's cognitive psychology are not taken into consideration). We must explore the possibilities to move to a new level, at which the information is used for retrieval of knowledge. When handling patient data, or studying the characteristics of diseases this will allow systems operations to be influenced by the specifics of the end-user. The expected result is that future medical systems using advanced analytics will be able to derive knowledge about successful approaches for treating a particular patient. They will be based on analysis of the decisions made by other physicians in similar cases.

The purpose of our work is to explore the possibilities of semantic retrieval of information and knowledge from textual sources (medical records, annotations, emails, blogs, social networks, etc.).

C. *Information extraction from unstructured data*

Modern non-invasive medical imaging techniques allow a generation of highly detailed anatomical and

 ISBN: 978-1-61208-384-1

physiological information about the human body to be easily accessible. This information is usually represented by a sequence of high-quality medical images (slices) stored in specialized and non-standardized formats. In general, this information is two-dimensional (static images, such as chest X-rays), three-dimensional (3D reconstruction of bodies from a set of slices) and four-dimensional (information about changes of 3D structures in time, e.g., the fetus movies). In addition, a pseudo-colour may also be used as an additional procedure in order to extract specific information about the patient. All these data are unstructured.

An additional problem of the use of medical images as an information source is that often the images are subject to linear or non-linear distortions, shifts, rotations, scaling, etc. This determines the ineffectiveness of many of the existing algorithms.

All these problems reduce the possibilities to structure the imaging data. This is the base for improving the efficiency of the information retrieval process.

D. The use and analysis of biological data.

Bioinformatics is one of the fastest growing sciences in the 21st century and belongs to the so called "life sciences", covering the studies of the living organisms such as plants, animals and human beings. Prior to the era of bioinformatics there were only two types of biological experiments: in a living organism (*in vivo*) or in an artificial environment (*in vitro*). It is commonly assumed that bioinformatics is "*in silico*" biology, i.e., the biological experiments are realized through computer models simulated on silicon chips.

The potential of bioinformatics to identify useful genes resulted in studies of the changes of the normal cell activities in various disease conditions. It also let to the creation of new gene products like drugs and vaccines. All this has led to a paradigm shift in biology and in biotechnology. The existing science paradigms have changed and now the genome science research provides an opportunity to carry out scientific experiments using computer modeling and simulations in such areas as drugs and vaccines synthesis, genomics, gene therapy, the study of the evolution, etc.

A major challenge in the analysis of biological data is to offer an integrated and contemporary access to exponentially growing amounts of data in multiple formats, as well as efficient algorithms for their processing.

The purpose of the study is to investigate how to retrieve and how to integrate biological data and biological information in on-line medical systems and services.

E. Possibilities for integration of medical information

Organization, storage and maintenance of a huge diversity of medical and biological data remains a challenge due to the following factors:

- The volume of the data has been increasing almost exponentially in the last decade.
- New data types are emerging and new medical and biological concepts are being developed.
- There is no standardization in the nomenclature of the data.
- The data is most often stored in flat files and relational databases: about 70% of the data is stored in text format or as static images; the remaining 30% of data is stored in different types of databases, organized in indexed files or in specialized relational databases.

The strong decentralization of the medical and biological data, the considerable differences in terminology and the peculiarities of the generic data sources description, as well as the difference in format of data search queries requires automated procedures for databases integration to be developed. The aim is to achieve more than just retrieving and modifying of data because nowadays the professional performance in any field is increasingly dependent on accessing the proper data and information. This requires surveys which are comprehensive, easy to use and linked to the other databases so that they will provide the necessary data resources. The heterogeneity and decentralization require suitable methods to provide access to the actual data associated with a specific disease or a specific medical problem. This involves the integration of large and diverse databases / information / knowledge associated with different levels of performance.

The goal here is to investigate the platforms of advanced analytics, the input data formats and the systems analysis of big datasets and to propose a conceptual architecture for an advanced analytics system of a new generation in medical practice. The proposed architecture must provide an opportunity for design of an integrated and modern access to the exponentially growing amounts of data in multiple formats. The further exploration includes providing an access to a constantly updated representation of the accumulated knowledge in the medical field.

III. CURRENT RESULTS

At present, the project is in its first stage. This decreases the amount of current result. Nevertheless, some of the obtained data and achieved results can be summarized.

One of the main project goals is to propose requirements for the logical design of the new generation of tools for advanced analytics in medical systems, as well as recommendations for the limitations and capacities for generating new approaches for this category of tasks. To achieve this goal, our investigation starts from determining input/output data streams to/from hospital and hospital networking. As a generic source of raw data we use data streams in Sofia Medical University hospitals complex - the biggest Bulgarian hospitals complex.

 ISBN: 978-1-61208-384-1

The study of the hospital information systems and their networking determines the following data/information/knowledge sources: medical staff computers, clinical workstations, microbiology, radiology, clinical laboratories, pharmacy, clinical databases (electronic medical records), patient computers, financial systems (billing, cost accounting), material management, administrative systems, research databases, library system, and educational resources.

The study of the input/output data streams has determined the following sources and consumers of data/information/knowledge: patients (home workstations), other hospital systems, other physicians, government healthcare systems (e.g., electronic medical records), pharmaceuticals regulators, insurance agencies, medical research groups/institutions, the Internet, other information resources/libraries/databases, vendors and providers of various types, and medical education centers (e.g., medical schools).

The heterogeneity and the distribution of data at this stage show that the main problems of multiple data sources are the following:

- Heterogeneity of names – different databases store the same values, but the names of the attributes given are different.
- Heterogeneity of relational structure – the composition of attributes in a complex structure is varies, but the stored values are identical.
- Heterogeneity of values – the method of presentation of values differs in different data sources.
- Semantic heterogeneity – according to the type of storage, different assumptions can be made about the data relevance, reliability and usefulness.
- Heterogeneity of the models of data storage – this raises the issue of transformation between models.
- Heterogeneity by time – different data are obtained at different times.

The reduction in number of all of these kinds of heterogeneity is important for investigation and evaluation of data variability and interoperability.

IV. CONCLUSIONS

The accumulation of large amounts of data in the process of examination of certain types of patients (especially those with chronic diseases) raises the problem of how to catalogue the obtained (often heterogeneous and dispersed) information, its machine processing in order to facilitate its understanding, as well as the ability to perform comparable and traceable measurements, especially in image (photographic) data. This calls for new research in the field of:

- analyses of medical images;
- text analyses based on given semantic criteria (most often for the purpose of real time processing);
- analysis of diverse types of clinical information to support the clinical decision making;
- comparative studies on the effectiveness of approaches and medical practices;
- predictive analysis, quality analysis of medical data and many others.

It follows from the above that the most important task for the creation of a new generation of tools for extended analyses is to analyse the modality of the various data, which in turn would allow the retrieval of information for specific diseases and the development of spatio-temporal descriptions for comparing the data and their modality. This would allow the development of new approaches which will facilitate the decision making process based on similarity between patients' data. Merging this information based on certain semantic features and generating new ideas for good medical practices will open up new vistas for patient treatment.

REFERENCES

[1] J. W. Cortada, D. Gordon, and B. Lenihan, , "The value of analytics in healthcare: From insights to outcomes", Executive report, IBM Institute for Business Value, April 2012.

[2] Deloitte, "2014 Global health care outlook: Shared challenges, shared opportunities", 2014, url: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/dttl-lshc-2014-global-health-care-sector-report.pdf [last accesed: 05.Jan.2015]

[3] J. Cattell, S. Chilukuri, and M. Levy, "How big data can revolutionize pharmaceutical R&D", Article, McKinsey& Company, April 2013, url: http://www.mckinsey.com/insights/health_systems_and_services/how_big_data_can_revolutionize_pharmaceutical_r_and_d [last accesed: 05.Jan.2015]

[4] B. Kayyali, D. Knott, and S. Van Kuiken, "The big-data revolution in US health care: Accelerating value and innovation", Article, McKinsey&Company, April 2013 , url: http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care [last accesed: 05.Jan.2015]

[5] MarketsandMarkets, "Healthcare Analytics/Medical Analytics Market by Application (Clinical, Financial, & Operational), Type (Predictive, & Prescriptive), End-user (Payer, Provider, HIE, ACO), Delivery Mode (On-premise, Web, & Cloud) - Trends & Global Forecasts to 2020", report, Markets andMarket, Dec. 2013, url: http://www.marketsandmarkets.com/Market-Reports/healthcare-data-analytics-market-905.html [last accesed: 05.Jan.2015]

[6] European Commission, "Accelerating the Development of the e-Health Market in Europe, e-Health Taskforce report 2007", ISBN-13-978-92-79-07288-8,Brussels, url: http://ec.europa.eu/information_sociaty/ehealth [last accesed: 05.Jan.2015]

[7] European Commission, "e-Health Action Plan 2012-2020: Innovative healthcare for the 21st century", Brussels, Dec. 1012, url: http://ec.europa.eu/information_society/newsroom/

cf/document.cfm?action=display&doc_id=4188 [last accesed: 05.Jan.2015]

[8] COCIR eHEALTH Market Intelligence Center, Market Information, April 2013, url:http://www.cocir.org/site/fileadmin/Publications_2013/April_2013_COCIR_eHealth_Market_Intelligence_Center.pdf [last accesed: 05.Jan.2015]

[9] V. Lessens, "Are Europe's Hospitals Ready for eHealth?", Outsourcing, 2005, issue 4, vol. 7, period 9-10, pp. 14-15

[10] I. Nikolova, G. Angelova, D. Tcharaktchiev, and S. Boytcheva. Medical Archetypes and Information Extraction Templates in Automatic Processing of Clinical Narratives. In Proceedings of ICCS 2013, 10-12 January, Mumbai, India, Conceptual Structures for STEM Research and Education, Springer, Lecture Notes in Computer Science Volume 7735, 2013, pp 106-120

[11] D. Tcharaktchiev, G. Angelova, S. Boytcheva, Z. Angelov, and S. Zacharieva. Completion of Structured Patient Descriptions by Semantic Mining. In: Koutkias, V., J. Niès, S. Jensen, N. Maglaveras, and R. Beuscart (Eds.), Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety, IOS Press, Studies in Health Technology and Informatics series, Volume 166, 2011, pp. 260 – 269

[12] D. Tcharaktchiev, V. Gueorguiev, and I. E. Ivanov, Standards for Medical Information Interchange and Design of Modern Mobile Devices and Solutions, GlobalHealth 2014 “The Third International Conference on Global Health Challenges”, Rome, Italy, 2014, pp. 134-139, ISBN: 978-1-61208-359-9

 ISBN: 978-1-61208-384-1