EXPERIMENTAL DATABASE IMPLEMENTATION FOR THE CEREBROVASCULAR DISEASES RESEARCH INTEGRATES TOGETHER DIFFERENT KINDS OF MEDICAL DATA

Petr Včelák, Jana Klečková
Department of Computer Science and Engineering, University of West Bohemia, Univerzitni 8, Pilsen, Czech Republic

Vladimír Rohan
Department of Neurology, University Hospital, Pilsen, Czech Republic

Keywords: Cerebrovascular diseases, Clinical data, DASTA, Database, DICOM, Medical data, Metadata, Resource Description Framework (RDF), Stroke, Thrombolysis.

Abstract: The cerebrovascular diseases are one the most common cause of death worldwide. The second most frequent cause of death in the Czech Republic. The proposed database should notably contribute to the solution of this complex problem. Its profit is based on medical data interconnection and aggregation of collaborating centers. There are stored miscellaneous de-indentified medical data in the database such as (1) a set of patient's clinical data in a DASTA file format, (2) a set of brain scans like computed tomography in a DICOM files, (3) data from Safe Implementation of Thrombolysis in Stroke (SITS) register. The experimental database project has an extensible support for any other fitted data format.

1 INTRODUCTION

Stroke is one of the leading causes of morbidity and mortality worldwide (Lopez and Blobel, 2009). Large differences in incidence, prevalence and mortality have been noted between Eastern and Western Europe. This has been attributed to differences in risk factors, with higher levels of hypertension and other risk factors resulting in more severe stroke in Eastern Europe (Brainin et al., 2000). Notable regional variations have also been found within Western Europe. Stroke is the most important cause of morbidity and long term disability in Europe, and demographic changes will result in an increase in both incidence and prevalence. It is also the second most common cause of dementia, the most frequent cause of epilepsy in the elderly, and a frequent cause of depression. (Awareness et al., 2008; Rothwell et al., 2005; O'Brien et al., 2003) The cerebrovascular diseases are one the most common cause of death worldwide. It is the second most frequent cause of death in the Czech Republic. Cancer incidence and mortality are two to three times greater in the Czech Republic than in other developed countries in Europe. (MZČR – Ministry of Health of the Czech Republic, 2010; ČSÚ – Czech Statistical Office, 2010)

The experimental database is as a meta data mining tool used primarily in a research. It allows studying of crucial values and dependencies of particular parameters in high amount of examinations with clinical data correlations. Captured data and obtained knowledge should notably contribute to the solution of complex cerebrovascular brain diseases problem. Later, the database can be used as a practical education and diagnostic tool. The experimental database concept was announced in (Včelák et al., 2009).

2 MEDICAL DATA TYPE SUPPORT

This paper experiments and hypothesis are based on an evaluation of heterogeneous medical data. It is a patient's clinical data set in a DASTA (Karlova univerzita v Praze – 2. lékařská fakulta v Praze (Charles University in Prague – 2nd Faculty of Medicine), 2010) format and radiological studies in a DICOM format (National Institute of Neurological Disorders and Stroke, 2010) of patients affected by a stroke at

Včelák P., Klečková J. and Rohan V. (2010).
EXPERIMENTAL DATABASE IMPLEMENTATION FOR THE CEREBROVASCULAR DISEASES RESEARCH INTEGRATES TOGETHER DIFFERENT KINDS OF MEDICAL DATA.
In *Proceedings of the Multi-Conference on Innovative Developments in ICT*, pages 111-114
DOI: 10.5220/0003037501110114

the University hospital in Pilsen. The hospital is a participant of the Safe Implementation of Thrombolysis in Stroke (SITS) program, an internet-based interactive thrombolysis register (Safe Implementation of Thrombolysis in Stroke, 2010). Medical doctor records details about a stroke into the SITS register. The SITS register contains e.g. therapeutic details, adverse drug reaction, National Institutes of Health (NIH) data, computed tomography (CT) description, death cause description. These data are not structured in a clinical data from the hospital information system. Feedback missing in the SITS register is a huge disadvantage for a medical doctor. That is why we are trying to interconnects clinical and imaging data (Rohan et al., 2007) with SITS the register data in the experimental database.

Patient's health records are provided mostly in the native language text data. A laboratory report is a structured document used while making a decision, calculation or graphing. These clinical data can be stored in a DASTA or a HL7 format. A diagnostic imaging procedures creates multimedia data, that originates from e.g. electroencephalograph (EEG) and Computed Tomography (CT).

2.1 DASTA Type

The DASTA is Data Standard abbreviation (Karlova univerzita v Praze – 2. lékařská fakulta v Praze (Charles University in Prague – 2nd Faculty of Medicine), 2010). It is a national electronic communication standard format of a public health service in the Czech Republic. The DASTA is a Health Level 7 (Health Level Seven, Inc., 2010) standard equivalent and was developed by the Czech Public Health Informatics and Scientific Information Organization supported by the Minister of Health of the Czech Republic. It is based on the XML-based mark-up standard with XSD schema.

2.2 Health Level Seven Type

Health Level Seven (HL7) is a non-profit organization involved in the development of international healthcare standards. It is widely used for interchange between hospitals and physician record systems and between electronic medical record systems and practice management systems. (Health Level Seven, Inc., 2010)

The HL7 Clinical Document Architecture (CDA) documents are used to communicate documents such as physician notes and other material (Dolin et al., 2006). The CDA is an XML-based mark-up standard intended to specify the encoding, structure and semantics of clinical documents. The CDA document consists of a mandatory textual part and optional structured parts. The mandatory textual part ensure human interpretation of the document contents.

The experimental database bargain for the HL7 documents support. The HL7 storage is supported, but a next layer with HL7 format support is not implemented, at this time.

2.3 DICOM Type

The Digital Imaging and Communications in Medicine (DICOM) standard has been developed with an emphasis on diagnostic medical imaging. It is applicable to a wide range of image and non-image related information exchanged in clinical and other medical environments. This standard is widely used for representing and communicating radiology images scans and reporting. The DICOM standard is supported by the National Electrical Manufacturers Association. (National Institute of Neurological Disorders and Stroke, 2010)

The searching in the set of DICOM files has some difficulties. You need a special library to acquire any information from the DICOM file.

2.4 The SITS Register Data

The SITS register is an academic driven, non-profit, international collaboration. It is an initiative by the medical professionals to certify excellence in acute stroke treatment. The SITS initiated an internet-based interactive thrombolysis register, to serve as an instrument for clinical centers to follow their own treatment results and compare with other centers in their countries and in the collaborating countries. (Safe Implementation of Thrombolysis in Stroke, 2010)

Treatment results are not available to a medical doctor for reuse this data in a structured way for a research purpose. We created a set of scripts that download and parse a treatment file report page.

3 DATA PROCESSING

The research is based on an anonymous data only. All of identification information are necessarily removed before an upload from the hospital to the experimental database. The data anonymization process kept all relationships. That is why we can use so much more data to research. The patient's clinical data are stored together with imaging examinations in a single database. It can be extended with any other proper kind of data.

The hospital information system exports a particular patient with a stroke medical data sets into a single directory. There are exported DASTA and DICOM files with unique names. DASTA file links all clinical event relevant DICOM files. The links are DICOM file names. The DASTA file can contain the SITS register identificator in a comment attribute. Each stroke corresponds to one SITS register identificator. De-identified DASTA, DICOM and created SITS-XML files are raw input files in the experimental system. All raw input files are processed and extraction of meta data is done. A new information or knowledge is referred as a meta data. We decided to use a modern Resource Description Framework (RDF) model for a meta data manipulation (W3C, 2010). The raw data processing tool is extensible. It can be a simple script that extracts any attribute value or an application with complex queries, images evaluation, computing, etc. A processing tool can operate on a meta data extracted previously, not only on a raw data.

The advantage of our experimental database is the incremental growing of a RDF model and increasing knowledge base. Suppose you are interested in a treatment time delay of a stroke. First, you know nothing, except the algorithm. So you implement an evaluation of a time interval between a stroke and a treatment's begin. The RDF model will be extended after the implemented algorithm execution. After that, you are able to confirm a treatment time delay impression on patients treatment result. Other advantage example is when you want to search for a set of specific files based on any conditions then it will be too time consuming to process the raw data. Apparently it is a better way to acquire useful data earlier and store them into the RDF graph. Then the search can be more quickly and efficiently.

4 HYPOTHESIS AND EXPERIMENTS

4.1 Data Set Description

The whole stroke data set characteristic in the experimental database is:

- Stroke was in years 2005 - 2009
- Patient count 244; Middle age 68
- Male/Female count 148 / 96
- Male/Female middle age 66.5 / 71
- Male/Female average age 66.1 / 67.2

4.2 Patient's Age and Sex Distribution Statistic

Advanced age is one of the most significant stroke risk factors. The incidence of stroke increases exponentially from 30 years of age, and etiology varies by age (Ellekjær et al., 1997). Our real patient's data set are shown on the Figure 1 grouped by age and sex. It shows us that 95 % of strokes mostly occur in people age 45 and older. Two-thirds of strokes occur in those over the age of 65. Similar results are also referred in (Senelick and Dougherty, 2001; National Electrical Manufacturers Association (NEMA), 2010). A person's risk of dying if he or she does have a stroke increases with age. However, it is evidence that stroke can occur at any age.

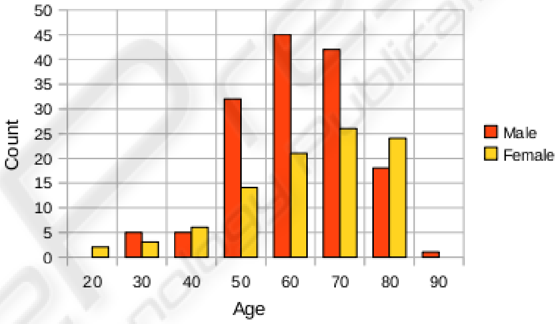

Figure 1: Patient's Age and Sex distribution over the stroke data set.

4.3 rt-PA Treatment Time Delay

The last basic example of data usage is a treatment time delay evaluation. It is a time interval between a stroke and an rt-PA treatment begin. The distribution of the treatment time delay by hours is shown on a Figure 2. Any more complex query evaluation research can continue with this values, e.g. we can identify dependency between the treatment time delays and a treatment success.

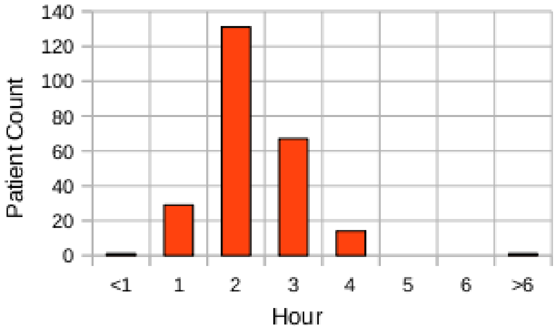

Figure 2: Treatment time delay. It is a time interval between a patient's stroke and begin of an rt-PA treatment.

5 CONCLUSIONS

The collaboration vision of medical doctor and computer programmer team were established. The medical doctor informs what to do with the data and how to interpret data for the research purposes. The computer programmer creates and implements a right algorithm. The experimental database is ready to new participants, now.

The project benefits from the medical data relationships. Related data and a RDF data model is the project base. Meta data stored in a RDF format are acquired by raw and previously processed medical data processing. By this way we create information and build a knowledge base. The knowledge base speed up and ease searching for resources. A user can take advantage of this meta data referred to the raw medical data. It is better to querying in a huge amount of data. Generally, the meta data can contain any data type. With RDF model we can refer to any data type.

Basic part of the project is implemented and all ordinary medical data types are supported. Currently, we are working on a new data mining algorithms that extends the project knowledge base.

ACKNOWLEDGEMENTS

The work presented in this paper was supported by the project Czech Science Foundation number 106/09/0740.

REFERENCES

Awareness, P., Referral, E., Transfer, P., Care, S., Imaging, D., Principles, G., Tests, B., Prevention, P., Pressure, H., Smoking, C., et al. (2008). Guidelines for Management of Ischaemic Stroke and Transient Ischaemic Attack 2008. *Cerebrovasc Dis*, 25:457–507.

Brainin, M., Bornstein, N., Boysen, G., Demarin, V., et al. (2000). Acute neurological stroke care in Europe: results of the European Stroke Care Inventory. *European Journal of Neurology*, 7(1):5–10.

Dolin, R., Alschuler, L., Boyer, S., Beebe, C., Behlen, F., Biron, P., and Shabo Shvo, A. (2006). HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, 13(1):30.

Ellekjær, H., Holmen, J., Indredavik, B., and Terent, A. (1997). Epidemiology of stroke in Innherred, Norway, 1994 to 1996: incidence and 30-day case-fatality rate. *Stroke*, 28(11):2180.

Health Level Seven, Inc. (2010). What is hl7? Online, 2010-03-02. http://www.hl7.org/about/index.cfm.

Karlova univerzita v Praze – 2. lékařská fakulta v Praze (Charles University in Prague – 2nd Faculty of Medicine) (2010). Data Standard (DASTA). Online, 2010-03-02. http://dasta.lf2.cuni.cz/.

Lopez, D. and Blobel, B. (2009). A development framework for semantically interoperable health information systems. *International Journal of Medical Informatics*, 78(2):83–103.

MZČR – Ministry of Health of the Czech Republic (2010). Ministerstvo zdravotnictví České Republiky: Věstník č. 2/2010: Péče o pacienty s cerebrovaskulárním onemocněním České republice. Online, 2010-03-01. http://legislativa.mzcr.cz/File.ashx?id=233&name= V%C4%9Bstn%C3%ADk_%20%%C4%8D_02_2010.pdf.

National Electrical Manufacturers Association (NEMA) (2010). Stroke: Hope Through Research. Online, 2010-03-02. http://www.ninds.nih.gov/disorders/stroke/detail_stroke.htm.

National Institute of Neurological Disorders and Stroke (2010). Digital Imaging and Communications in Medicine (DICOM). Online, 2010-03-02. http://medical.nema.org.

O'Brien, J., Erkinjuntti, T., Reisberg, B., Roman, G., Sawada, T., Pantoni, L., Bowler, J., Ballard, C., DeCarli, C., Gorelick, P., et al. (2003). Vascular cognitive impairment. *The Lancet Neurology*, 2(2):89–98.

Rohan, V., Sevcik, P., Polivka, J., Ambler, Z., Kreuzberg, B., and Ferda, J. (2007). Klinický pohled na výpoČetní tomografii u akutní ischemie mozku (A clinical Approach to Computed Tomography in Acute Cerebral Ischemia). *Česká a slovenská neurologie a neurochirurgie*.

Rothwell, P., Coull, A., Silver, L., Fairhead, J., Giles, M., Lovelock, C., Redgrave, J., Bull, L., Welch, S., Cuthbertson, F., et al. (2005). Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (oxford vascular study). *The Lancet*, 366(9499):1773–1783.

Safe Implementation of Thrombolysis in Stroke (2010). About safe implementation of thrombolysis in stroke. Online, 2010-03-02. http://www.acutestroke.org.

Senelick, R. and Dougherty, K. (2001). *Living with stroke: a guide for families*. Health South Press.

Včelák, P., Polívka, J., Maule, P., Kratochvíl, P., and Klečková, J. (2009). Experimental database system for the vascular brain diseases research. In *Neuroinformatics 2009: 2nd INCF Congress of Neuroinformatics*, Pilsen, Czech Republic. Frontiers in Neuroinformatics. Frontiers Research Foundation.

W3C (2010). Resource Description Framework (RDF): Concepts and Abstract Syntax. Online, 2010-03-02. http://www.w3.org/TR/rdf-concepts/.

Český – Czech Statistical Office (2010). Český statistický úřad: Úmrtností tabulky (Death-rate Statistics). Online, 2010-03-02. http://www.czso.cz/csu/redakce.nsf/i/umrtnostni_tabulky.