MEDICAL IMAGING DATA IN MULTICENTRIC DATA COLLECTION

Martin Klimek

Doctoral Degree Programme (1), FEEC BUT

E-mail: xklime23@stud.feec.vutbr.cz

Supervised by: Jiří Kozumplík

E-mail: kozumpli@feec.vutbr.cz

Abstract: The presented article deals with the issue of storing and sharing data from medical imaging systems. This paper, inter alia, consists of organizational and informatics aspects of medical imaging systems data in multicentric studies containing MRI brain images. This paper also includes technical design of a web-based application for image data sharing including a web interface suitable for manipulation with the image data stored in a database.

Keywords: EEICT, multicentric data collection, clinical registry/registries, anonymization

1. INTRODUCTION

Nowadays, medical imaging modalities present large variety of methods which use different principles to create images of the examined part of patient's body. Every day new computer and communication technologies enter the medical field in order to satisfy the ever growing demands and requirements. The following article focuses on radiology, especially on the issues involving technical and organizational features of MRI images processing, where the amount of medical image data, which are created and stored in electronic form, is quickly increasing. MRI is widely used in many different medical disciplines, including psychiatry, [1]. The study of the connection between anatomical changes in the brain and various neuropsychiatric disorders such as schizophrenia, Alzheimer's disease, and other different forms of dementia has in the last decade become the center of interest for many psychiatric congresses and scientific articles in the field. Unlike in other diseases and injuries when neuropsychiatric disorders are concerned morphological changes are not visible in MRI at the first sight, not even to the eye of an experienced expert. The detection of these changes is usually a result of statistic comparison of groups of patients and healthy volunteers, [2].

2. DATA SHARING

Gradually, the Internet has become ever more used as an information medium and a new conception of service provision has been started. Fast and exact diagnostics, immediate processing and distribution of data among doctors, the possibility of consultation of diagnostic conclusions with specialized medical workplaces, these are the reasons why so called digital data storages are being created on the Internet. These storages can be described as systematically organized and managed files mainly consisting of electronic sources, that is of digital documents and documents which were transformed into digital form. Moreover, these storages provide us with all the necessary and appropriate electronic services. [3]. The field of data sharing is now aiming to build a shared archive of digital medical data which would ensure immediate availability of data to doctors and a possibility of long-distance co-operation of a large number of doctors via the Internet. Another important step in this field is the usage of the acquired data and technology for educational and research purposes. The established databases can be used, provided very strict anonymization rules will be complied with, as a support for starting doctors and/or students of medicine and other related fields. The use of this computer application in medical education, research, treatment, and many

other areas will enhance decision-making, management planning and medical research, which will eventually improve the quality of patient care.

3. THE STRUCTURE AND SECURITY OF DATA

According to the up-to-date legislation of the Czech Republic, all medical facilities are obliged to keep medical records (this applies in full scope for all medical facilities since 2001 and according to the information protection act these facilities are also considered personal information managers and processors), [4]. Medical facilities are therefore required to take the necessary preliminary precautions in order to prevent unauthorized and/or random access to personal information, any changes and/or loss of personal information, their unauthorized transfer and/or processing, and any other possible abuse of these data.

3.1. CLINICAL REGISTRIES

Clinical registries systematically collect all information concerning the health of the monitored individuals. It is important to note that clinical registry (unlike clinical studies, which are done in order to test the safety or efficiency of a new therapeutic process or drug) collects data directly from the medical practise, [5]. Electronic registry is usually based on a computer database which has to be dimensionalized enough for state of the art equipment and has to meet strict security criteria. Clinical registries are used mainly when it is necessary to carry out studies of large numbers of patients. These studies are especially useful in the research of so called complex diseases, where we can also find the majority of psychiatric disorders. Even though the incidence of these disorders is relatively high, their causes are not exactly known. The primary purpose of clinical registries is the analysis of the statistic relevance of these information when the above mentioned disorders are concerned.

3.2. CLINICAL IMAGE DATA ARCHIVES

One of the basic requirements for the building of digital medical archives is the development of high-quality archiving system. This enabled especially thanks to the great increase in the storage technology capacity and the speed of the Internet transfer. The possibility to further process and archive data (analysis, filtering, compression) was one of the fundamental reasons for the introduction of digitalization, [6]. PACS systems (Picture Archiving and Communication Systems) are trying to integrate all necessary patients' data. It is also trying to enable easy access to these data no matter the time and place of their acquisition. The activity of every PACS system can be divided into acquisition, processing, archiving and distribution of images. The increasing usage of computers in clinical practise has provoked the need to standardize the access to image information.

4. DATABASE FOR MULTICENTRIC DATA COLLECTION

The aim of the work was a design of a web interface which would work as a tool for collection and organization of work with image documentation acquired from MRI within multicenter study focused on modern psychiatrich research. Even though there is a whole range of clinical registries in the Czech republic, it is not known that they would have an interface for the collection of medical image data. For the development of the final web application for registration of medical documentation the script language PHP was used. For the development of the databases the database server MySQL was used. The structure of the developed system is shown in the implementation diagram in picture no.1. The implementation diagram describes the individual parts of the final product and it also determines the relations between these components.

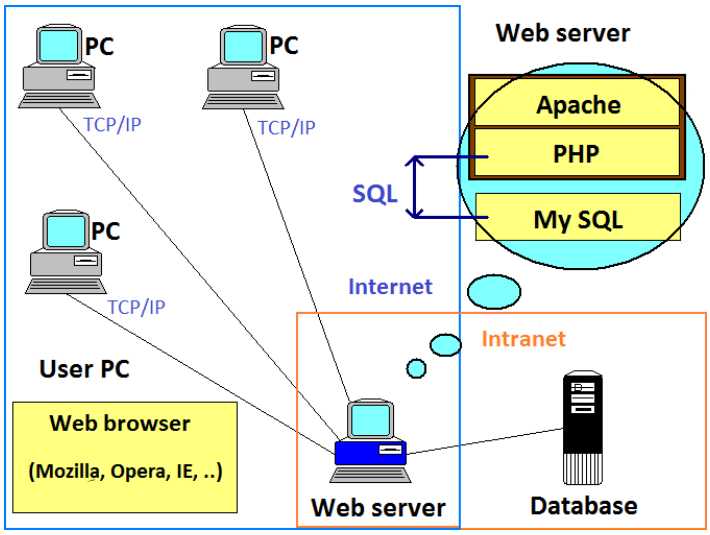

Figure 1: Architectury of web application.

4.1. WEB INTERFACE

Web interface functions as a primary tool for the exchange of information between a user and a computer via a connected network. (That is either the Internet or a local Intranet). Desktop applications, which communicate with the central database, can also be used for the shared data manipulation. Nevertheless, the disadvantage of these applications is that whenever the server part is updated, all the workplaces with the client application must be updated as well. This, however, increases the operating costs. The undoubtable advantage of web interface, taken into consideration in the work, is the fact that these interfaces, on the client-side, need only a web browser for the insertion and editing of data. This can be seen in the previously mentioned implementation diagram in picture no.1. All service including maintenance takes place on the software provider-side.

4.2. 3-D IMAGING DATA COLLECTION

When using DICOM there is a large file of individual images created during a patient's examination. In order to eliminate the time-taking uploading of individual images, an electronic form was designed. This forms allows the upload of compressed ZIP files. The collection of large number of MRI images is provided by a programmed component for ZIP file collection. After these ZIP files are uploaded on the server they are automatically decompressed and stored in the appropriate patient file. When the NifTI format is used this problem is fully eliminated because of the nature of the format, in this case only one NifTI file is uploaded.

4.3. SECURITY ASPECTS AND THE PROTECTION OF PERSONAL INFORMATION

Because it is strictly given the doctor-patient relationship must not in any way be disturbed and all the information acquired in the relationship must be confidential and secure, the security aspects of the web application and the documentation it stores play a key role in the design of this application. The access to the data center will be allowed only after proper authorization which should eliminate the access of unauthorized people. Nowadays, there are many technical ways of securing a login to a internet application, however, the authorization via user name and password is used here. After user verification, it is possible to proceed to editing, uploading, and transferring of data. Another possible place of a potential attack is the Internet way used by the data floating during particular activities such as the above mentioned uploading, etc. Therefore, it is important to make sure it is not possible to intercept the communication between the client computer and server and this way read or even change the transferred data. In order to eliminate this threat, the data floating in both ways must be encrypted. For these security precautions encrypting SSL (Secure Sockets Layer) was used. On the client-server level this protocol uses a combination of encrypting algo-

rithms, [16]. Another important part of the poject was the anonymization of medical data. Anonymization can be, to a certain degree, understood as a contradiction to identification. According to the level of anonymization of direct data between two subjects – a data manager (a university) and a data provider (a hospital = HCP – health care provider), anonymization can be devided into four types:

1. Patient's birth number itself is stored in the registry.

2. MD5 hash of a birth number is stored in the registry.

3. Only an identificator of a birth number is stored in the registry.

4. Local data collection only

The first possibility allows very direct pairing of HCP and registry data. This pairing can be done only by the data manager. This pairing is impossible without a contract with the given hospital. The second level of anonymization, when a hash print of a chosen algorithm is stored in the registry, this way a mentally not very complex pairing of documents is created (a large number of HCP records, their encryption, and pairing). Technically, the pairing is again possible only at the data manager side. According to the law, this alternative is the same as the previous possibility because this way of encrypting has a liability of backward searching and therefore it is again not considered to be a proper anonymization, even though it is much more difficult to find a particular birth number without the necessary expert knowledge. When the third level of anonymization is concerned, only certain identifiers of birth numbers can be found in the registry. Technology allows to use several possibilities. The first possibility is the situation when an authorized employee of a hospital enters data into a registry and the birth number conversion table is kept somewhere in the hospital. However, this solution is rather unpractical and there is a great risk of loss of data or incompleteness of data and therefore further manual data correction. This solution requires active participation of the hospital facility at the data pairing. There are no direct data stored in the registry and therefore according to the law, this solution is considered to be a proper anonymization. The second possibility of this level is a state when all the birth numbers entered into the registry are encrypted by an encrypting algorithm but before the encryption is done, a previously agreed prefix, known only to a certain hospital employee, is added to the birth numbers. This possibility is technologically and logistically the most demanding solution which deserves a separate analysis and a detailed design based on a particular solution options of the given registry. For pairing a co-operation with the prefix keeper is necessary, which can be troublesome if there are more keepers of different prefixes. In this case the pairing of documents can be rather time and technologically demanding. In the registry there is only the birth numbers' print, which is unreadable without the hash password and therefore no backward searching can be done. According to the law this is also cinsidered a proper anonymization and this very option was used in this project (see below). The last level of anonymization is the option when the collection of data is done locally. The registry is run in the hospital only and there are no requirements concerning its content. If an analytic processing is needed, anonymous exports are made. Technologically, this level is conditioned by the possibilities on the client-side (hospital). In this project, when anonymization of a data file, which are to be stored in a database accessible via web interface, is concerned, the emphasis was put on the fact that the information in the registry must not in any way lead to direct identification of a person. One of the possible options was to substitute these data by a certain identification number (ID). This number would be connected with other parts of the database by a key, so that potential new files could be added to the correct patient. However, if the security is somehow breached and an attacker gets past the technical barrier, they have a chance to identify the patient by using this key (indirect identification), therefore this can be considered only a pseudo-anonymous file and not an anonymous one. The aim was to limit the amount of personal information in the file while the information present must be encrypted. It is necessary that the stored data be stored in an encrypted form. This way the data cannot be read directly. For the purpose of one way encryption of birth numbers MD5 (Message-Digest algorithm) or SHA (Secure Hash Algorithm) algorithms can be used. Both of these encrypting algorithms change the given data into a print of a fixed length. Their main advantage is

that only a small change in the input data means a great change in the output (this means a creation of a fundamentally different print). In the project the MD5 algorithm encryption was used. For example, if there is an individual born on 1.1.1990 with birth number 900101/xxxx (x can stand for a zero – this is not relevant here), after MD5 encryption the print of this birth number will be "54007264e63810abc626f197eb18be17". As it was already mentioned when speaking about the second level of anonymization, this way is vulnerable because of the possibility of backward searching and it is not considered a proper anonymization. This is the reason why before the MD5 encryption the birth number chain is supplemented by a special code which is stored in a different place than the birth number chains. If a special code, for example number one, is added at the beginning of the previously chosen birth number "900101/0000", The print created by the MD5 algorithm stands as follows „c68aa8de85d88148e32fdd28b8a38523", this is obviously an absolutely different code when compared with the previous one. Moreover, if the data are stolen from the database the patient cannot be identified without the special key, unlike in the first case. According to the law this is considered a fully legitimate anonymization.

5. CONCLUSION

In the Czech Republic so far, clinical registries work solely with data stored in form of text. Image data are used very rarely and only in foreign registries. Legal aspects of multicentric database using MRI data are very similar to those of clinical registries, however, since clinical registries do not use image data, the issue of data sharing in multicentric database has not been processed. The issue of anonymization of patient data was solved by the MD5 algorithm (with prefix) and it meets the requirements of the Czech legislation. Test implementation of the developed application already runs and uses about 20 patient's medical record. In clinical practice it is important to prevent any possible interception of communication and therefore use encrypting channel. It is also important that the server be dimensionalized enough for peak load.

REFERENCES

[1] KAWASAKI, Y. et al.: Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. NeuroImage 34:235-242, 2007

[2] THOMAZ, C. E. et al.: Multivariate Statistical Differences of MRI Samples of the Human Brain. Journal of Mathematical Imaging and Vision 29: 95-106, 2007

[3] KUKUROVÁ, Elena; VLČÁK, L'ubomír. *Pricípy e-health*. Olomouc : SOLEN PRINT, 2009. 154 s. ISBN 978-80-903776-7-7

[4] *Registry.cz* [online]. 2008 [cit. 2011-02-23]. Legislativní aspekty: Česká republika. Dostupné z WWW: <http://registry.cz/index.php?pg=legislativni-aspekty--ceska-republika>

[5] *Systémy pro sběr klinických dat* [online]. Brno: Institut biostatistiky a analýz, 19. 10. 2007, 4. 5. 2009 [cit. 2009-05-04].

[6] MCNEIL, J. J., et al. *Guidelines for the establishment and management of clinical registri: Clinical registries*. 2nd draft version, 2008.