© *The Journal on Information Technology in Healthcare* 2009; **7**(4): xxx–xxx

Security and Privacy Issues in Bipolar Disorder Research

Panagiotis Rizomiliotis, Aggeliki Tsohou, Costas Lambrinoudakis, Stefanos Gritzalis

Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece.

ABSTRACT

Mental health diseases are common but research to further knowledge and understanding of them is hampered by data privacy and confidentiality regulations that apply to medical records. Centralised databases containing the relevant medical history of thousands of patients with an individual mental disease would be of great value for researchers, enabling techniques such as data mining to be applied. The major challenge in achieving this is anonymising the data to satisfy legal and ethical requirements without removing important clinical information. In this paper we propose a model that can be used to create a central repository of anonymised data for patients with bipolar disease. Knowledge obtained from the database is fed into an expert system which can guide clinicians in patient management. Security requirements are provided by access to the database being controlled by RBAC (Role Based Access Control).

INTRODUCTION

Mental disorders or mental illnesses such as mood disorders, anxiety disorders, psychotic disorders, eating disorders, and personality disorders affect approximately 1 in 4 of the population[1]. One relatively common serious mental illness is bipolar disorder (BD), which is also known as manic depression, manic depressive disorder or bipolar affective disorder. BD is characterised by episodes of full-blown mania which is defined as periods of abnormally expanded or irritable mood, along with major depression. These episodes can have devastating consequences on the professional and social life of those affected. Alcohol and drug abuse and dependence, and social and professional isolation, are the most common complications of BD. Around 10–20% of bipolar patients who have been hospitalised at some stage during their

Correspondence and reprint requests: Panagiotis Rizomiliotis, Department of Information and Communication Systems Engineering, University of the Aegean, Samos GR-83200, Greece. E-mail: prizomil@aegean.gr.

treatment, die by committing suicide[2,3]. BD is a significant health problem worldwide as its prevalence is approximately 1% across all populations.

Despite the fact that BD is fairly common, diagnosis and management of individual cases is still frequently sub-optimal. This is partly due to a lack of detailed knowledge and understanding of BD and this is not helped by the fact that research in the disease is hampered by patient data being classified as sensitive and confidential. As a result knowledge that could result from subjecting data from large numbers of patients to techniques such as data analysis and data mining cannot be obtained. Achieving this in practice requires preserving anonymity and confidentiality when pooling data from large numbers of BD patients.

In this paper we propose a model to enable centralised collection of anonymised data from patients with BD. This data can be studied by researchers to further knowledge and understanding of BD and the new knowledge can be fed into an expert system that clinicians can use to assist them in patient management.

SYSTEM ARCHITECTURE

The architecture of the proposed system is shown in Figure 1.

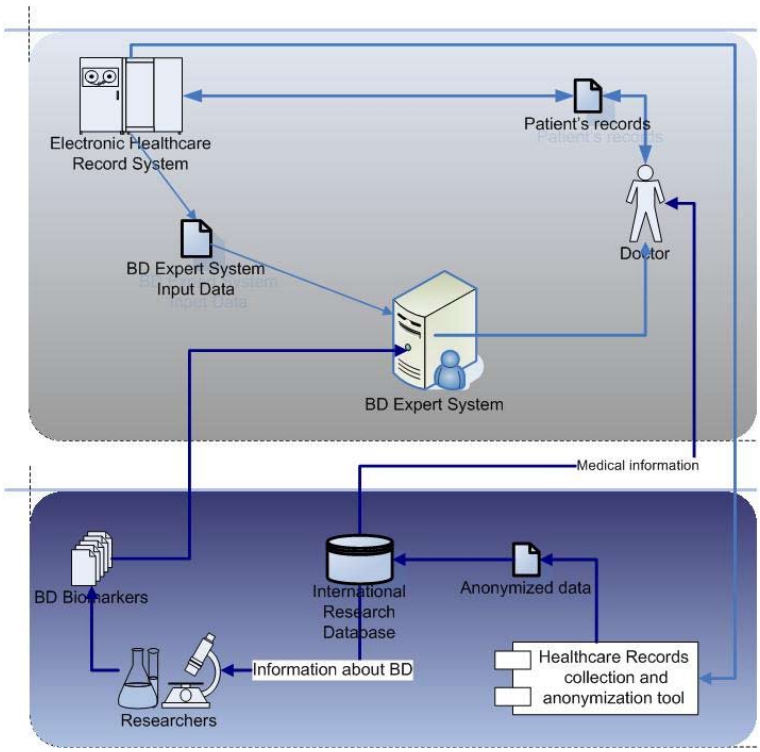

Figure 1. *Proposed architecture for an anonymised research database and expert system for bipolar disease (BD)*

The top half of the figure demonstrates how a doctor obtains help with patient management. The expert system processes data collected from the patient and information stored in the patient's electronic health record (EHR), including information relevant to BD. The output of the system is based on state-of-the-art knowledge about BD. The knowledge is codified in a special form known as a BD-biomarker. If necessary the doctor may also consult the research database, for example if he wants to obtain details on the case management of another patient similar to the one he is treating.

The bottom half of the figure shows how the model supports research into BD. A special tool is used to collect and anonymise data from different EHR systems. The collected data is used to update a Research Database. Authorised personnel can access to the Research Database and new knowledge gained from the database is formalised as electronic BD biomarkers. This is made available for interaction through the expert system presented in the top half of the diagram. The key aspects of the system with respect to anonymisation of data and meeting legal and ethical requirements will now be described in detail.

SECURITY AND PRIVACY REQUIREMENTS

In order to implement the BD research central repository, the processing of health data stored in geographically spread EHRs is inevitable. The communication between the different EHR systems and the repository of accumulated anonymised BD-related data system has to be carefully designed in order to guarantee the confidentiality, authenticity and integrity of data. The provided solution must be generic and flexible in the sense that it should address different systems ranging from current healthcare systems to legacy IT systems to allow interoperability between the EHRs and the repository and also with the BD research community systems. Finally access to the research data should be restricted only to authorised users.

Processing health data, by definition, raises several security and privacy issues, such as the protection of data integrity and confidentiality and the preservation of the patient's privacy. Health data belongs to a special category of personal data, commonly known as sensitive data. Legally all data contained in medical documentation such as electronic health records is considered as sensitive data. Consequently according to legislation it cannot be shared in a way that identifies a patient without the patient's explicit permission. Data can be anonymised by removing elements from it, but the challenge then becomes of how to best remove data to ensure that the patient cannot be identified whilst at the same time ensuring that the remaining data contains all the key elements necessary for research purposes including data mining.

ANONYMISING DATA FOR DATA MINING

A number of techniques have been proposed for modifying or transforming data in such a way so as to preserve privacy, but leaving it suitable for data mining. Some examples of these are:

The randomisation method: The randomisation technique uses data distortion methods in order to create private representations of the records[4,5]. In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. The randomisation approach is particularly well suited to privacy-preserving data mining of streams, since the noise added to a given record is independent of the rest of the data. The most common methods of randomisation are those of additive perturbations and multiplicative perturbations.

The k-anonymity model and l-diversity: The k-anonymity method is based on techniques such as generalisation and suppression according to which any given record maps to at least k other records in the data. The k-anonymity model was developed in order to prohibit the indirect identification of records from public databases, since combinations of record attributes can be used to exactly identify individual records. The l-diversity model was designed to handle some weaknesses in the k-anonymity model. Protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group. Thus, the concept of intra-group diversity of sensitive values is promoted within the anonymisation scheme[6].

Distributed privacy preservation: A partition is a division of a logical database or its constituting elements into distinct independent parts. The partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). There are applications where users wish to derive aggregate results from data sets partitioned across other individuals. While the individuals do not desire to share their entire data sets, they consent to limited information sharing. The overall effect of such methods is to maintain privacy for each individual entity, while deriving aggregate results over the entire data[7–9].

Downgrading Application Effectiveness: The output of applications such as association rule mining, classification or query processing may lead to violations of privacy and motivated research into downgrading the effectiveness of applications by either data or application modifications. Such techniques include association rule hiding[10], classifier downgrading[11], and query auditing[12].

Each one of the above techniques has its own advantages and disadvantages. Depending on the application the system designer has to choose the most adequate method. In this context, it is not a straightforward task to identify the most appropriate techniques for the anonymisation of medical data.

ANONYMISING MEDICAL DATA

Two specific systems that have been developed to anonymise medical data are:

- **The Scrub system**[13]
- **The Datafly system**[14]

The Scrub system[13] was designed for de-identification of clinical notes which usually occur in the form of textual data and contain references to patients, patients' family members, addresses etc. The Scrub system uses detection algorithms, based on several local knowledge sources, to determine when a block of text leaks information concerning the name, address or a phone number of a patient or a member of its family. This system was proposed in order to replace the traditional, and in most cases insufficient techniques, based on a simple "search and replace procedure".

The Datafly System[14] is one of the earliest practical systems for anonymisation and one of the first applications of privacy-preserving transformations. The system was designed in response to the concern that the process of removing only directly identifying attributes such as social security numbers was not sufficient to guarantee privacy. This work has a similar motive as the *k*-anonymity approach of preventing record identification, but it does not formally use a *k*-anonymity model in order to prevent identification through linkage attacks. The Datafly system, as well as most of its successors, proposes anonymity levels ranging from 0 to 1. An anonymity level of 0 results in Datafly providing the original data, whereas an anonymity level of 1 results in the maximum level of generalisation of the underlying data.

To enable collection of BD data in a centralised database we propose using the distributed databases model. All data is collected and stored in a local database maintained at the hospital or the clinic that treats the patient. The anonymisation process is applied to every distributed database and the anonymised data is then stored in the research repository. This will give research community members access to a full collection of anomymised clinical data. The major challenge with this approach is updating the centralised database. While it is relatively easy to add new records, special care is required for updating existing ones. One solution to this problem is the application of general purpose secure multiparty computation techniques borrowed from the cryptographic literature[15].

The privacy preservation techniques that is the most appropriate for the proposed architecture is the *Distributed Privacy Preservation*. In the proposed scenario, even though several entities (hospitals, clinics, individual doctors) do not desire to share their entire data sets, they are willing to give their consent to limited information sharing. At the same time, there are entities wishing to derive aggregate results from

data sets partitioned across other individuals. To achieve this, distributed algorithms for k-Anonymity can be used, combining previous proposed solutions[7–9], in order to maintain k-anonymity across different distributed parties. It is assumed that the data record has both sensitive attributes and quasi-identifier attributes. The solution uses encryption on the sensitive attributes which can be decrypted only if therefore are at least k records with the same values on the quasi-identifiers[9]. Thus, k-anonymity is maintained. The issue of k-anonymity is also important in the context of hiding identification in the context of distributed location databases[7,8]. In this case, k-anonymity of the user-identity is maintained even when the location information is released.

ACCESS CONTROL

The access to the research database must follow a role-based access control (RBAC) policy. Ideally it should define specific roles that will be authorised to access the research database, associating specific access privileges to each role. Examples of some of the roles and privileges are as follows:

- *Health Researchers* (e.g. clinicians, scientists, pharmaceutical companies, etc) who study BD. Their need is met by granting access on all available medical data (patient medical history, treatments, pharmaceutical substances etc) related to their field of interest.
- *Doctors* providing health care services to BD-patients. In cases where the diagnosis or/and treatment of the patient is not straight forward, the doctor will be able to obtain guidance from the expert system and also may be able to access previous similar cases that have been treated by other clinicians.
- *BD-Biomarker administrators* will be responsible for 'representing' new knowledge about the disease in a structured format, known as a BD-biomarker. The biomarkers are then utilised by the expert systems for supporting clinicians in the optimal management of patients with BD.

DISCUSSION

In this paper we have proposed a model that can be used to support research to further knowledge and understanding of bipolar disease and also aid clinicians in managing individual patients using the best current available knowledge. The former objective is achieved by providing a centralised system architecture that enhances access to BD related data. To meet the requirements for data privacy and confidentiality, legal and ethical requirements the model includes an anonymisation process and a role-based access control policy. The implementation of the proposed system will enhance the secure interoperability and seamless communication of BD health data between clinicians, health researchers, and those responsible for creating the knowledge in the Expert system. This can them be used by clinicians to aid them in patient management.

REFERENCES

1 Andrews G, Totov N. Depression is very disabling. *Lancet*. 2007; **370**: 808–9.

2 Harris E, Barraclough B. Suicide as an outcome for mental disorders: a meta-analysis. *The British Journal of Psychiatry* 1997; **170**: 205–28.

3 Bostwick JM, Panktratz V. (2000). Affective disorders and suicide risks: a reexamination. *Am J. Psychiatry* 2000; **157**: 1925–32.

4 Agrawal R, Srikant R. Privacy-preserving data mining. *Proceedings of the ACM SIGMOD Conference, 20*00.

5 Agrawal D. Aggarwal C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. *ACM PODS Conference, 20*02.

6 Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-Diversity: Privacy Beyond k-Anonymity. *ICDE* (2006).

7 Bettini C, Wang XS Jajodia S. Protecting privacy against location based personal identification. *Proc. of Secure Data Management Workshop*, Trondheim, Norway, 2005.

8 Gedik B, Liu L. A customizable k-anonymity model for protecting location privacy, *ICDCS Conference, 20*05.

9 Zhong S, Yang Z, Wright R. Privacy-enhancing k-anonymization of customer data, In *Proceedings of the ACMSIGMOD-SIGACT-SIGART Principles of Database Systems*, Baltimore, 2005.

10 Verykios VS, Elmagarmid A, Bertino E, Saygin Y, Dasseni E. Association Rule Hiding. *IEEE Transactions on Knowledge and Data Engineering* 2004; **16**(4).

11 Moskowitz I, Chang L. A decision theoretic system for information downgrading. *Joint Conference on Information Sciences, 2000*.

12 Adam N, Wortmann JC. Security-Control methods for statistical databases: a comparison study. *ACM Computing Surveys*, 1989; **21**: 515–56.

13 Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. In: Cimino, JJ, (ed.), Proceedings, *Journal of the American Medical Informatics Association. Washington*, DC: Hanley & Belfus, Inc, 2006, pp. 333–37.

14 Sweeney L. Guaranteeing anonymity while sharing data, the Datafly system. *Proc AMIA Annu Fall Symp.* 1997, pp. 51–55.

15 Goldreich O, Micali S, Wigderson A. How to play the mental game. *Proceedings of the 19th Annual ACM Symposium on Theory of Computing*, ACM, 1987, pp. 218–29.