Security and Privacy Issues in Bipolar Disorder Research

Panagiotis Rizomiliotis, Aggeliki Tsohou, Costas Lambrinoudakis, Stefanos Gritzalis

Dept. of Information and Communication Systems Engineering,

University of the Aegean, Samos GR-83200, Greece

email: {prizomil, agt, clam, sgritz}@aegean.gr

Abstract

Most mental illnesses, including bipolar disorder (BD), cause disability. BD is one of the world's 10 most disabling conditions, characterized by episodes of full-blown mania and major depression, with devastating consequences on the professional and social life of the patient. A major problem in BD diagnosis and treatment is the absence of objective criteria and lack of understanding of the underlying pathological mechanisms and symptoms linked to episodes. The need for a central repository that will maintain BD related data is therefore a prerequisite for triggering BD-research and address the aforementioned problem. Specifically, it will collect healthcare data for BD cases in Europe, phenotypical information (clinical, cognitive, electrophysiological, brain imaging and biochemical evaluations), genotype information, and other information like sleep activity, actimeter, speech characteristics etc.

Even though this approach is highly beneficial for medical research, the processing of medical data raises, by definition, security and privacy issues; protection of data confidentiality and integrity as well as inability to identify the patient. This paper presents an anonymity-preserving mechanism for disclosing electronic health care records to the research community without revealing the identity of the BD patient while taking into account local and international data protection legislation and other related ethical issues. Finally, we will identify the parts of the system where access control is required and will specify the rights that each user role should exhibit over the system resources.

Keywords: Bipolar disorder, mental research, biomarkers, access control, anonymization techniques

1. Introduction

Mental diseases constitute one of the most severe menaces against public health, as more than a quarter of the world population suffers or will suffer from a psychiatric disorder [18]. Most of the mental illnesses cause disability and, among them, bipolar disorder (BD), which is also known as the manic-depressive illness and is the most dangerous one. BD is characterized by episodes of full-blown mania, defined as periods of abnormally expanded or irritable mood, and major depression. These episodes have devastating consequences on the professional and social life of the patient. Alcohol and drug abuse and dependence, and social and professional isolation, are the most common complications of BD. Around 10-20% of the formerly hospitalized bipolar subjects die by committing suicide [19][20]. BD is one of the world's 10 most disabling conditions, as its lifetime prevalence is approximately 1% across all populations. All bipolar spectrum subtypes, bring the prevalence of all bipolar disorders to more than 4% of the general population.

Most research problems are still open in the area of BD diagnosis and treatment. Some of them include absence of objective criteria for BD diagnosis, reduction of the diagnosis time, understanding of the underlying pathological mechanisms, tracking patterns that lead to episodes, study of the episodes characteristics (frequency and duration), and the discovery of novel and innovative treatments.

In this paper we propose a model architecture (see Figure 1) with a two-fold contribution: a) supporting the objective selection of the most appropriate treatment for a BD and b) supporting research in the area of BD overcoming the interoperability issues resulting from the variety of different and heterogeneous clinical data records, as well as, the legal and ethical issues that arise from the processing of medical data.

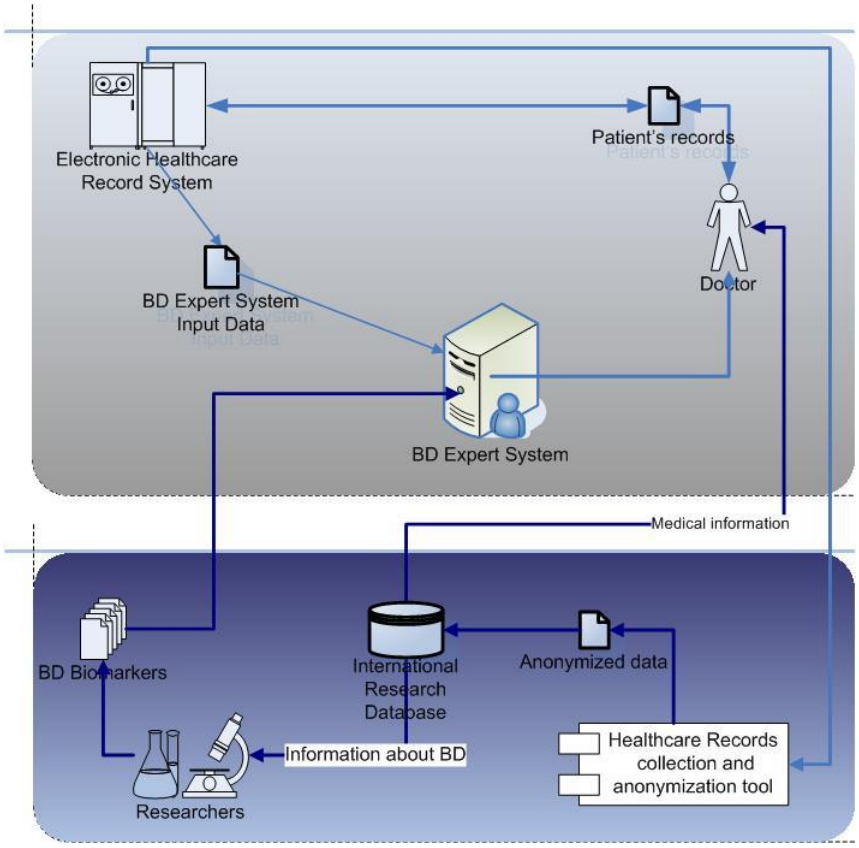

Figure 1: The Proposed Architecture

The support of the most appropriate treatment for the BD patient is outside the scope of this paper. We suffice to say that it utilizes an expert system that processes data collected from the patient and information stored at the patient's EHR (including genotypic information and cognitive assessment data). The output of the system is based on the state-of-the-art knowledge about BD disease, knowledge that is being codified in a special form known as BD-biomarker. If necessary the doctor may also consult the research database in order to have an up-to-date knowledge of the latest improvements and results in the area of BD. On the other hand, for supporting BD research an anonymization tool is used for collecting and anonymizing data from different EHR systems. The collected data are used to update the Research Database of BD cases. The authorized researchers can consult the International Research Database, according to the access control policy enforced. New scientific results are formalized as electronic BD biomarkers suitable for interacting with the Expert System presented above.

The paper is structured as follows: In the next section we highlight the most important legal and ethical issues that affect the computer mediated BD. Section 3 lists the security and privacy requirements that have been identified, while sections 4 and 5 present the proposed anonymization mechanism and access control policy respectively. Section 6 concludes the paper.

2. Legal and Ethical Issues of Computer-Mediated Bipolar Disorder

The proposed architecture should comply with the principles and rules laid down in the European data protection legislation. The European Data Protection Directive is technology-neutral and therefore can be well adapted to future challenges [13]. Health data belong to a special category of personal data, commonly known as sensitive data (Art. 8§1 Data Protection Directive). Their processing is in principle forbidden, although the Directive contains specific exceptional provisions for their processing. The main goal of the proposed architecture is the facilitation of BD-research with respect to the protection of patient privacy and legal compliance. Given that the Article 29 Working Party [1] considered that "*all data contained in medical documentation, in electronic health records and in EHR systems should be considered to be sensitive data*" [14], special scrutiny will be paid to all kind of data that are intended to be included in the central research repository. All privacy principles must be respected during both the collection and the processing of the data.

Another major issue refers to the rights of the patients with bipolar disorder. The European Commission has recently proposed a Directive on the application of patients' rights in cross-border healthcare [16], where data protection issues are discussed. This, as well as, the initiative of the European Economic and Social Committee on patient's rights [17] should be also taken into account.

3. Security and Privacy Requirements

In order to implement the BD research central repository, the processing of health data stored in geographically spread EHR is inevitable. 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. According to the legal and ethical issues, presented above, the processing of BD – related complete medical records is forbidden, since they reveal the patient's identity. The data that will be available to the research community must not reveal the identity of the person who has been diagnosed with BD disorder. For that purpose, an anonymization process is required and an anonymization tool, that will comply with the local and international data protection regulation, shall be employed. The anonymization of patient's data from different EHR systems is a strong prerequisite, and at the same time a very demanding task. Currently there are several alternative approaches but there is no universally accepted solution.

Furthermore, the communication between the different EHR systems and the repository of accumulated anonymized BD-related data system should be very carefully designed in order to guarantee (as much as possible) the confidentiality, authenticity and integrity of data. Also, 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 allowing the interoperability between the EHRs and the repository and also with the BD research community systems. Finally, the access to the research data should be restricted only to authorized users. The characterization of user categories and the privileges of each category should be described in an access control policy.

[1] Under Article 29 of the Data Protection Directive, a Working Party on the Protection of Individuals with regard to the Processing of Personal Data is established, made up of the Data Protection Commissioners from the Member States together with a representative of the European Commission. The Working Party is independent and acts in an advisory capacity. The Working Party seeks to harmonize the application of data protection rules throughout the EU, and publishes opinions and recommendations on various data protection topics.

4. The Anonymization Process

The topic of privacy has been traditionally studied in the context of cryptography and information-hiding. In recent years, because of the widespread proliferation of electronic data maintained by private organizations, corporations and hospitals, data mining has been viewed as a severe threat to privacy.

Privacy-preserving data mining has a plethora of applications in surveillance which are naturally supposed to be "privacy-violating" applications. The key is to design methods which continue to be effective, without compromising security. Typically, such methods reduce the granularity of representation in order to reduce the privacy, by applying some form of transformation on the data. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy, which in some applications is unacceptable.

A number of techniques have been proposed for modifying or transforming the data in such a way so as to preserve privacy. Some examples of such techniques are as follows:

- *The randomization method:* The randomization method is a technique for privacy-preserving data mining in which noise, or more precisely pseudorandom bit strings, is added to the data in order to mask the attribute values of records [6][7]. Techniques are designed to derive aggregate distributions from the perturbed records and for that data mining techniques have be developed in order to work with these aggregate distributions. The randomization 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 randomization 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 generalization and suppression according to which any given record maps onto 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 anonymization scheme [8].
- *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 [1][2][3].
- *Downgrading Application Effectiveness:* The output of applications such as association rule mining, classification or query processing may lead in violations of privacy and motivated the research in downgrading the effectiveness of applications by either data or application modifications. Such techniques include association rule hiding [9], classifier downgrading [10], and query auditing [11].

Each one of the above techniques has advantages and disadvantages. The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. 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 anonymization of medical data.

Next, we present two characteristic examples of systems that have been proposed for the anonymization of medical data:

- The Scrub system [4] 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 [5] is one of the earliest practical systems for anonymization 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 generalization of the underlying data.

Medical data play a crucial role in the progress of BD research. In order to satisfy all the privacy concerns raised and to hurdle all the obstacles introduced by the data protection laws and regulations, an anonymization process has to be applied before clinical data become available for processing by the research community.

Traditionally, data mining algorithms have been applied in centralized collections of data. Admittedly, the accuracy of any data mining task may be also increased when data are collected from various locations. In the case of BD data the distrusted databases model is unavoidable. All data are collected and stored in a local database maintained at the hospital or the clinic that the patient's doctor works. While the natural choice would be to maintain this distributed structure, there are some facts that must be considered due to the anonymization process. As we have already mentioned, the preservation of privacy comes with a certain cost. Since the cryptographic protocols, that are applied, inevitably increase the computational cost, there are applications that cannot tolerate this overhead and for which the adoption of privacy preserving techniques is prohibitive. Motivated by above, we propose a centralized research repository. More precisely, the anonymization process is applied to every distributed database and the anonymized data are stored to the research repository. Thus, the research community members will have access only to this collection of anomymized clinical data. Of course this approach has several drawbacks. The most important debit concerns the update procedure of this centralized database. While it is trivial to add new records, special care is needed for the update of the existing ones. One solution for this problem is the application of general purpose secure multiparty computation techniques borrowed again from the cryptographic literature [12].

The family of 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. More precisely, distributed algorithms for k-Anonymity can be used, combining the solutions proposed in [3], [1], [2], in order to maintain k-anonymity across different distributed parties. The work in [3] discusses the extreme case in which each

site is a user which owns exactly one field from the data. 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. 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 [1][2]. In this case, k-anonymity of the user-identity is maintained even when the location information is released.

5. Access Control

The access to the research database should follow a role-based access control (RBAC) policy. Ideally it should define specific roles that will be authorized to access the research database, associating specific access privileges to each role.

The proposed list of roles follows:

- *Health Researchers* (academia, health care organizations, laboratories, pharmaceutical) who study BD. Their need is to grant 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 need to access additional information on similar cases that have been monitored by other HISs throughout Europe.
- *BD-Biomarker administrators* will be responsible for 'representing' new knowledge about the disease in a structured form know as BD-biomarker. The biomarkers are then utilised by the expert systems for supporting the doctors in the selection of the optimal treatment for the patient.
- *System – Data Base Administrators*

6. Conclusions

In this paper we have proposed a solution for addressing some problems of the bipolar disorder unresolved issues. We made an attempt to assist the research community of bipolar disorder by providing a centralised system architecture that will enhance the access to BD related data. The system is also based on the support of the most appropriate treatment for the BD patient utilizing an expert system BD-biomarkers. Besides these elements, our model includes an anonymization process and a role-based access control policy as a consequence of the significant security, privacy, legal and ethical requirements that exist from the processing of the BD-related data. The implementation of the proposed system will enhance the secure interoperability and seamless communication of health data between a) health researchers, b) doctors, and c) BD-Biomarker administrators, and others.

7. References

[1] Bettini, C., Wang, X. S., Jajodia, S.(2005). Protecting Privacy against Location Based Personal Identification. *Proc. of Secure Data Management Workshop*, Trondheim, Norway.

[2] Gedik, B. & Liu, L. (2005). A customizable k-anonymity model for protecting location privacy, *ICDCS Conference*.

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

[4] Sweeney, L. (1996). 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, 333-337.
[5] Sweeney, L. (1997). Guaranteeing Anonymity while Sharing Data, the Datafly System. *Journal of the American Medical Informatics Association*.
[6] Agrawal, R., Srikant, R. (2000). Privacy-Preserving Data Mining. *Proceedings of the ACM SIGMOD Conference*.
[7] Agrawal, D. Aggarwal, C. C. (2002). On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. *ACM PODS Conference*.
[8] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). l-Diversity: Privacy Beyond k-Anonymity. *ICDE*.
[9] Verykios, V. S., Elmagarmid, A., Bertino, E., Saygin, Y., and Dasseni, E. (2004). Association Rule Hiding. *IEEE Transactions on Knowledge and Data Engineering*, 16(4).
[10] Moskowitz, I., Chang, L.(2000). A decision theoretic system for information downgrading. *Joint Conference on Information Sciences*.
[11] Adam, N., Wortmann, J. C. (1989). Security-Control Methods for Statistical Databases: A Comparison Study. *ACM Computing Surveys*, 21(4).
[12] Goldreich, O., Micali, S., Wigderson, A. (1987). How to play the mental game. STOC '87: *Proceedings of the 19th Annual ACM symposium on Theory of computing*, ACM, pp. 218-229.
[13] European Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Communication on future networks and the internet*. COM (2008) 594 final, Brussels, 29.09.2008, p. 10.
[14] Article 29 Data Protection Working Party, *Working document on the processing of personal data relating to health in electronic health records* (EHR), WP131, 15.02.2007, p.7.
[15] Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use, *Official Journal of the European Communities*, L121/34 – L121/44.
[16] European Commission, *Proposal for a directive of the European Parliament and of the Council on the application of patient's rights in cross-border healthcare*. COM (2008) 414 final, Brussels, 02.07.2008.
[17] European Economic and Social Committee, *Opinion on patient's rights*, SOC/221, Brussels, 26.11.2007.
[18] Andrews, G, Totov, N. (2007). Depression is very disabling. *Lancet*. 370: 808-809
[19] Harris, E., Barraclough, B. (1997). Suicide as an outcome for mental disorders. A meta-analysis. *The British Journal of Psychiatry*. Mar; 170: 205-228.
[20] Bostwick, J.M., Panktratz, V. (2000). Affective disorders and suicide risks: a reexamination. Am J. Psychiatry. Dec; 157(12): 1925-1932.