A Systematic Mapping Study on Patient Data Privacy and Security for Software System Development

Isma Masood
Department of software engineering
International Islamic University
Islamabad, Pakistan
ismamasood786@gmail.com

Saad Zafar
Faculty of Computing
Riphah International University
Islamabad, Pakistan
saadzafar@riu.edu.pk

***Abstract*-The exchange of Electronic Health Records (EHR) has increased threats to patient data privacy and security. The software systems developed for healthcare sector are required to explicitly address patient data privacy and security. A number of solutions have been proposed to incorporate these requirements into the software systems. However, there is no comprehensive study that synthesizes the different research initiatives according to any predetermined criteria. The main focus of this paper is to survey the various proposed solutions in the literature to incorporate patient data privacy and security into software systems. The proposed solutions are mapped against: (1) the software development stage for which the solution has been proposed, and (2) the established patient privacy and security principles. The existing literature has been surveyed using a systematic mapping study by phrasing two questions. In the mapping study, a total of 58 studies, dating from 2000 to 2011, were evaluated and mapped against the aforementioned categories.**

Keywords-Systematic mapping study; Electronic Health Records (EHR); Patient data privacy and security; Software system development.

I. INTRODUCTION

Health information and medical records contain sensitive personal information including diagnosis and testing information along with person's family history, genetic testing, history of diseases and treatments, history of drugs used, sexual orientation and practices, and testing for sexually transmitted diseases [1]. Nowadays, digitized health records are not only used for diagnosis and treatment but they are also used for other purposes like improving efficiency of the healthcare system, drive public policy development administration, conduct medical research, and to provide effective health services that can be tracked and evaluated [2,3].

Increasingly, the electronically shared information within healthcare sector is receiving new threats to patient data privacy and security. Threats to patient data privacy and security become a major cause of inaccuracies and improper disclosure of information, which threaten individual's personal life and financial well being [3, 4]. Therefore, many laws and policies in different countries have been implemented to protect patient data privacy and security especially for EHR [5].

To bridge the gap between different patient privacy rules, regulations and policies, Markle Foundation has proposed a set of principles under a *Common Framework* for uniform implementation of health information exchange across the health sector [9]. Markle Foundation works for advancement of health and national security through information and information technology in the United States of America. One of the major objectives of the Common Framework is to ensure patient privacy and seamless connectivity among various organizations related to the health sector. The privacy principles defined under the framework are described later in the paper.

A number of initiatives have been taken to propose effective integration these policies into software systems. However, effective implementation of *all* the policies and principles related to patient privacy and security into software systems remains a challenge.

Therefore, there is a room for new and improved solutions in this field. But before performing any new research, there is a need to synthesize the existing work in the area and to understand the need for improvement or to identify any new solution to an unresolved matter. Typically, a systematic literature review [SLR] is performed for this purpose. The idea of conducting SLR in the field of software engineering has been proposed by Kitchenham [6]. Often, a pre-requisite for conducting SLR is a mapping study, which is performed as an initial step to assess the feasibility of a complete SLR. In this paper we have conducted a mapping study as we could not find any SLR on the proposed solutions related to the Patient Data Privacy and Security in the field of software engineering. For this mapping study, we have followed the guidelines published in [7, 8].

We have presented the results of mapping study to identify available solutions on patient data privacy and security for software system development and have categorized these solutions against: (1) software development stages in software development cycle, and, (2) the well established policy principles for patient data privacy and security presented in [9]. Specifically, our mapping study addressed the following research questions (1) which solutions of patient data privacy and security have been

 ISBN: 978-1-61208-165-6

proposed for software system development? (2) Can we categorize these solutions using the Markle Foundation's Common Framework?

In Section II, we have described our systematic mapping process; in Section III, we provide explicit answers to our research questions; the discussion of the results is provided in Section IV; conclusion and the future work are given in the last section.

II. THE SYSTEMATIC MAPPING PROCESS

For our mapping study, we following the guidelines provided in [7, 8]. Accordingly, our mapping study was conducted in three stages. In Stage 1, we define the scope, the search strategy and the selection criteria. In the second stage primary studies were selected applying the search strategy and the selection criteria. Lastly, in Stage 3, the selected studies are classified into the different categories.

A. Stage 1: Defining Scope, search strategy and selection criteria

We define the scope of the study as follows. The *population* of the study is selected as the set of articles addressing patient data privacy and security. As *intervention*, we selected any patient data privacy and security solution proposed for any of the software development cycle (e.g., requirements engineering, design, testing, etc.). The *outcome* of our study is a mapping of selected solutions to the patient data privacy principles found in [9]. Our search string for conducting the research was:

Patient AND *Data* AND (*Privacy* OR *Security*)

The research sources selected for our study were IEEE Digital Library, ACM Digital Library, Science Direct and Springerlink. To select relevant studies, we used the following inclusion and exclusion criteria.

Inclusion Criteria: A study contribution related to any stage of the software system development lifecycle. The study should also discuss *at least* one or more than one principles of patient data policy. For this purpose we read abstract, conclusion, introduction, or the full paper (if required).

Exclusion Criteria: Any study not related to the domain of software engineering, patient data privacy or security is not selected. The studies related to patient data privacy and security for images, sensor network and wireless transmission are also not included.

B. Stage 2. Selecting primary studies

In the first iteration, the search string was used at each resource. All references along with their abstracts were downloaded in Endnote [11] reference library. At this stage, we downloaded 4,670 references. In the second iteration, abstract of all reference were read and relevant studies which explicitly addressed the patient data privacy or security with contribution towards software system development were selected and placed in another library of selected papers. In this iteration, 120 studies were selected. We selected 93 papers from IEEE, 6 papers from ACM, 17 papers from Science Direct and 4 papers from Springerlink. In the third iteration, full texts of these 120 studies were downloaded. We read all the articles one by one and applied the inclusion and exclusion criteria and finally selected 51 studies in our third iterative phase. We placed our 12 doubtful studies in the pending folder. In the fourth iteration, we discussed these doubtful studies and decided to accept 7 studies and to reject 5 studies. The breakdown of the results from each of the source is presented in Table 1, whereas Table 2 shows the distribution of our four iterative phases and the number of studies which were retained in each phase. In Table 3, we summarize the most relevant publication channels.

TABLE 1. NO. OF STUDIES AT EACH RESOURCE

Resource	No. of studies	No. of selected studies	Percentage
IEEE	4,540	44	0.96%
ACM	74	6	8.1%
Science Direct	40	8	20%
Springerlink	16	0	0%
Total	4,670	58	1.2%

TABLE 2. NO. OF STUDIES AT ITERATIONS

1st iteration	2nd iteration	3rd iteration	4th iteration
4,670	120	51	58

TABLE 3. MOST RELEVENT PUBLICATION CHANNELS

Acronym	Type of publication	Percent
International Journal of medical informatics	Journal	13.7%
Information Technology in Biomedicine	Journal	6.8%
CCSW	Workshop	5%
ICBECS	Conference	3.4%

The IEEE Digital Library had yielded the most number of papers (4,670), followed by ACM (74), Science Direct (40), and Springerlink (16). It is noteworthy that the most relevant studies were found in Science Direct (20%) and the least were found in Springerlink (0%). ACM had 8.1% and IEEE Digital Library had 0.96% relevant studies, respectively. Most of the relevant studies were found in International Journal of Medical Informatics (13.7%). This was followed by Information Technology in Biomedicine (6.8%). The rest of the relevant studies were found in two conferences: Workshop on cloud computing security (CCSW) (5%) and International Conference on Biomedical Engineering and Computer Science (ICBECS) (3.4%).

As part of our inclusion criteria, we included studies from the year 2000 to 2011. For the year 2000 we did not find any relevant study. However, from the years 2001 to 2008 the number of relevant studies increased steadily with a sharp increase in the year 2008 (frequency=17). The only exception to the trend is the year 2009 where the total number was reduced to only 4. In 2010 the number was again increased to 10 studies showing a positive trend. Only one study was found to be relevant in the first quarter of

 ISBN: 978-1-61208-165-6

2011. This trend of number relevant studies per year is given in Table 4.

TABLE 4. PERCENTAGE OF STUDIES AT EACH YEAR

Years	Relevant Studies	Selected Studies	Percentage
2000	2	0	0%
2001	5	2	3.4%
2002	6	1	1.7%
2003	8	3	5.1%
2004	8	3	5.1%
2005	10	3	5.1%
2006	10	4	6.8%
2007	23	8	13.7%
2008	25	17	29.3%
2009	30	4	6.8%
2010	23	10	17.2%
2011	22	1	1.7%
Total	**172**	**58**	

C. *Stage 3. Classifying selected Studies*

In the next stage, we divided our studies according into three categories. In the first category, we classified the studies according to the research approach used in the selected primary studies. We divide the research approaches according to the classification proposed by Weiringa et al. [10]. The *validation research* is used for those novel techniques that have not been implemented and are validated through experiments in a lab-like environment. The *evaluation research* is used to evaluate the techniques that have been implemented in practice. This research type explores how well the technique has been implemented. In the *solution proposal* either a novel solution is proposed or an existing solution is extended significantly. The *philosophical papers* propose either a conceptual framework to structure concepts into a new taxonomy. On the other hand *opinion papers* express personal opinion of the authors about a technique and the *experience papers* explain the experience of the authors of how a technique has been implemented in practice.

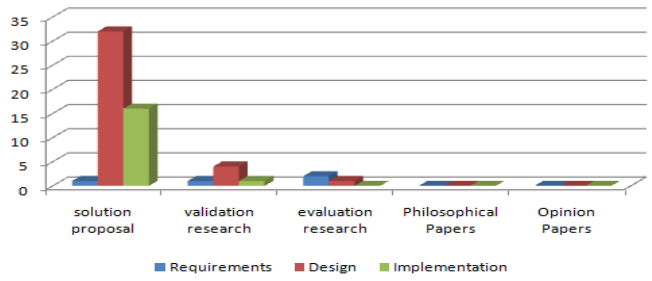

Figure 1: Mapping of studies according to research types

TABLE 5. RESEARCH TYPE AND SOFTWARE DEV.PHASE FACETS

Context	Solution	Validation	Evaluation	Total
Req.	1	1	2	**4**
Design	32	4	1	**37**
Imp.	16	1	0	**17**
Ver.	0	0	0	**0**
Maint.	0	0	0	**0**
Total	**49**	**6**	**3**	**58**

Table 5 shows the distribution of research type facet of the selected studies. An overwhelming majority of research approaches in the selected primary studies proposed a new solution (*f*=49). The next approach used the most was validation research (*f*=6) followed by evaluation research (*f*=3). However, we did not find any study that could be classified into any of the other research type categories. The results of this classification are summarized in Figure 1.

We also classified the studies on the basis of different stages of software development. Specifically, we grouped the software development stages into: *requirements*, *design*, *implementation*, *verification*, and *maintenance*. The breakdown of the classification of the selected studies is given in Table 5. The majority of selected primary studies addressed the Design phase of the software development (*f*=37), followed by the Implementation phase (*f=17*), while some of the studies were classified under the Requirements phase (*f*=4). We did not find any study related to software Verification and Maintenance phases.

Our next categorization was based on the Markle Foundation's privacy principles [9]. The first principle of (1) *Openness and Transparency* mandates that there should be an overall policy of openness regarding personal data. The individuals should be aware of the nature stored data, its location and its access control policy. The (2) *Purpose Specification and Minimization* principle requires that the data collection purpose should be defined at the time of collection and its use should be limited to the intended purpose. Under the (3) *Collection Limitation* principle the personal health information must only be collected lawfully and with the knowledge and consent of the concerned individual. The (4) *Use Limitation* principle states that personal data must not be disclosed, made available or used in any manner other than the specified purposes. The (5) *Individual Participation and Control* principle requires that individuals have the right of access and control over their stored personal information. The (6) *Data Integrity and Quality* states that only the relevant data is stored and that the data is always accurate, complete, and current. The (7) *Security Safeguards and Controls* requires there should be reasonable security safeguards against the risks of loss of data or unauthorized access. The accountability of entities responsible for keeping and maintaining the personal data according to stated principles is covered under the (8) *Accountability and Oversight* principle. Lastly, the (9) *Remedies* principle states that there are adequate legal and financial remedies to address any security breaches or privacy violations.

Table 6 shows the distribution of studies according to the aforementioned privacy principles. As reflected in the data

 ISBN: 978-1-61208-165-6

shown in the table, we found many single studies that address multiple privacy principles. The most coverage was given to the Use Limitation principle (*f*=38). This was followed equally by the Individual Participation and Control, and Security Safeguard and Control principles (*f*=24). After them the most covered principle was Data Integrity and Quality principle (*f*=16), followed by the Purpose Specification Principle (*f*=14). The next principle covered the most was the Accountability and Oversight principle (*f*=13), whereas, the Remedies and Collection Limitation had the least coverage with a frequency of 3 and 1, respectively.

TABLE 6. CLASSIFICATION OF STUDIES ACCORDING TO PRIVACY PRINCIPLES

Principle	Req.	Design	Impl.	Total
Openness	2	5	1	8
Purpose Specification	1	9	4	14
Collection Limitation	1	0	0	1
Use Limitation	1	23	14	38
Individual Participation and Control	1	17	6	24
Data Integrity and Quality	1	13	2	16
Security Safeguards and Control	2	17	5	24
Accountability and Oversight	2	9	2	13
Remedies	0	2	1	3

III. RESEARCH QUESTIONS

Based on the above data, we now answer our two research questions.

RQ1: Which solutions of patient data privacy and security have been proposed for software system development?

In our mapping study we found 58 relevant primary studies. Out of these studies 63% of the studies were related to the Software Design. While 27% of the studies contributed towards Software Implementation and only 6% aimed at Software Requirements. Therefore, we can conclude that the most research is being conducted on how to effectively design software systems related to the requirements of patient data privacy and security. Similarly, there is also significant focus in the research community on how to effectively implement the patient data privacy and security requirements. Surprisingly, much less studies are focused on requirements analysis and specification phase of software development (see Figure 2).

RQ2: Can we categorize these solutions using the Markle Foundation's Privacy Principles [9]?

The mapping of selected studies against the Markle Foundation's Privacy Principles is given in Figure 3. As discussed earlier, a single study was often mapped against multiple principles. But we found the solutions in the studies mapped reasonably well against the privacy principle. It is important to note that the Use Limitation was covered in 41.4% of the studies, followed by Individual Participation and Security Control principles with 41.4% studies. The other two principles covered in the selected studies were Data Integrity and Quality, and Purpose Specification with 27.6% and Purpose Specification 24.1%, respectively. The coverage of rest of the principles was not very significant.

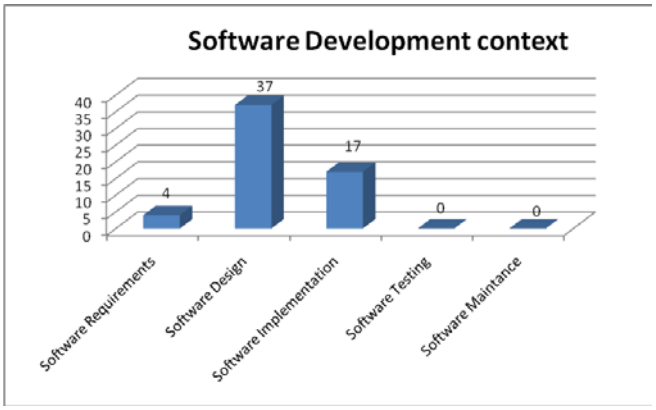

Figure 2: Mapping of studies according to software development context

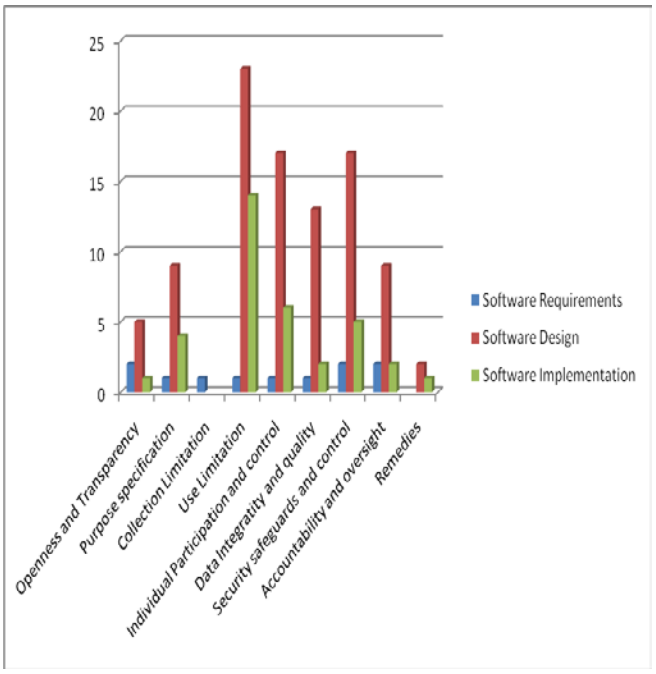

Figure 3: Mapping of studies against privacy principles

IV. DISCUSSION

The amount of personal information stored and exchanged by the health information systems is increasing by the day. With the increase in the volume of data the concern about the patient data privacy and security is also

increasing. The data stored about the patient include sensitive information like history of diseases and treatments, history of drugs used, sexual orientation and practices, results of sexually transmitted diseases, etc. As a result, a number of rules, regulations and best practices have been proposed to ensure that the stored data does not violate individual's privacy and that the data is never use inappropriately. Consequently, there has been a steady increase in research community to ensure that the software systems deployed must effectively integrate all the requirements related to patient data privacy and security.

The motivation behind our study was to investigate the feasibility for conducting a complete Systematic Literature Review. Here we cover the breadth of patient data privacy and security presented in the literature. The subsequent SLR studies can investigate the depth based on the results presented in our work.

The steady increase in the related primary studies from the year 2001 to 2010, with a few possible exceptions, indicates a growing interest in this significantly important research area (see Table 4). Similarly, the need of implementation of patient data and security requirements is reflected from the fact that most of the selected studies are concerned about the Design and Implementation of the privacy related requirements and less attention is paid to critically important phases of Requirements Analysis and Specification, Verification and Maintenance. This notion is further reinforced by the fact that the most common research approach used in the primary studies is Solution Proposal, with much less studies on validation and evaluation research. Likewise, we did not find any study based on experience reports, philosophical papers, or opinion papers.

Perhaps, not surprisingly the most importance is given to the Use Limitation, Individual Participation and Security Control principles. However, less coverage is given to the rest of the privacy principles, without which any software system cannot effectively implement a complete set of patient data privacy and security requirements.

We identify the following two limitations of our study: (1) some studies may have been missed due to the diverse use of the terms used in the search string; and (2) studies published in English language were selected in the search.

V. CONCLUSION AND FUTURE WORK

In this study, we have presented initial findings on solutions available for patient data privacy and security to develop software system. On this topic, we found 58 papers published in the years from 2000 to the first quarter of 2011. We have mapped these solutions against principles of privacy policy to cover all aspects of patient data privacy and security. A large number of studies focused on Software Design as compared to Software Implementation and Software Requirements while, no study found on testing and maintenance. The Use Limitation principle along with Individual Participation and Control, and Security Safeguard and Control had most coverage in the selected studies. Our future work includes performing in-depth Systematic Literature Review on various aspects of Patient Data Privacy and Security identified in this study.

REFERENCES

[1] U S. Congress, Office of Technology Assessment," Protecting Privacy in Computerized Medical Information" OTA-TCT 576. Washington, DC, US Government Printing Office, Sept 1993.

[2] A. Appari and M.E. Johnson., "Information Security and Privacy in Healthcare: Current State of Research."*International Journal Internet and Enterprise Management*, vol. 6, pp. 279-314, Oct. 2010.

[3] L.Gostin., "Health Care Information and Protection Privacy : Ethical and Legal Considerations" in ETATS-UNIS, 1997, pp. 683-690.

[4] C. H. Liu, Y. F. Chung, T. S Chen, and S. D Wang, "The Enhancement of Security in Healthcare Information Systems." *International Journal of Medical System*",pp. 1-16, Nov. 2010.

[5] M.Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Lalecil., "A Survey and Analysis of Electronic Healthcare Records Standards. "*Journal ACM Computing Surveys*, vol.37, pp. 277-315, Dec. 2005.

[6] B. Kitchenham and S.Charters., "Guidelines for performing systematic literature reviews in software engineering", Technical Report, EBSE-2007-01, Keele University, 2007.

[7] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson., "Systematic mapping studies in software engineering.", in 12th International Conference on Evaluation and Assessment in Software Engineering (EASE), pp. 71-80 , June. 2008.

[8] W. Afzal, R. Torkar, and R. Feldt., "A systematic mapping study on non-functional search-based software testing", in 20th International Conference on Software Engineering and Knowledge Engineering (SEKE), 2008.

[9] Markle Foundation, Connecting for Health Common Framework. January 10, 2011. <www.connectingforhealth.org>

[10] R.Wieringa, N.Maiden, N.Mead, and C.Rolland, "Requirements engineering paper classification and evaluation criteria: a proposal and a discussion", *Journal Requirements Engineering* . vol. 11, pp. 102–107, Dec. 2005.

[11] T.Reuters,"EndNote-Your smater refrence assistant"Internet. June 5, 2010. <http://www.endnote.com/>

 ISBN: 978-1-61208-165-6