9. User Privacy in Big Data Analytics for eHealth: Data Privacy Model

Nidhi, Department of Business Science and Technology, Aarhus University, Denmark
nidhi@btech.au.dk
Albena Mihovska, Department of Business Science and Technology, Aarhus University, Denmark
amihovska@btech.au.dk
Ramjee Prasad, Department of Business Science and Technology, Aarhus University, Denmark
ramjee@btech.au.dk

ABSTRACT

Big data analytics can benefit the healthcare sector by incorporating improved situational analysis, database management, real-time decision making and new ways of diagnosis and treatment. However, its use opens critical security and privacy concerns. This paper surveys the open challenges of collecting and accessing health data, and the different types of possible breaches of privacy and security that are the key to the successful deployment of eHealth systems. To mitigate the privacy hindrance issue with the medical data, we propose an eHealth data privacy model, which will add transparency in to the personal data collection, aggregation, handling and storage. Transparency in healthcare sector have different interpretation at different level. We'll look at some different segments in the healthcare industry working to adapt to the call for transparency. The model builds upon the Information Accountability protocol for the transparency. The user will be the player and take decision on their data, how is to be used and shared.

Keywords—*Big Data; e-Health; e-Health Data; Data Security and Privacy*

INTRODUCTION

A typical eHealth system should be highly secured, responsive, and controlled and one, where the users' privacy and the protection of their personal data, remains intact. eHealth systems demand integrity, accessibility and availability along with interoperability, which is even more important with the colossal pool of data defining the infrastructure of today. Many everyday applications and services rely on the collection, storage, processing and analysis of data, often user-related, and often made available to different sectors irrespective of boundaries ranging from machine learning and engineering, to economics and medicine [1].

The amount of data generated in healthcare sector is increasing and will continue to increase with the technological enactments, creating room for new data handling and analysing techniques. Big data analytics provide tools to benefit healthcare for example; it provides customized medications, anticipated analytics, risk-intervention etc. [2]. It has marked a presence in handling and analysing data generated via the social media but it offers promising solutions for handling efficiently eHealth data (also, commonly referred to as "health big data"). Big data analytics includes data aggregation, processing, storage of eHealth data to make decisions and evolve new ways of treatment, keep the population healthy etc. [1].

The health-related data are usually stored and processed at distributed repositories at different levels [3]. There are numerous security and privacy concerns in moving the health data under the big data approach. The privacy of the patients and the safeguarding of their personal data are major issues in applying big data analytics to eHealth.

A recent survey published in [4], suggested that the lack of adequate security measures had resulted in numerous data breaches in the healthcare sector, exposing certain patients to economic threats, mental stress, and even social embarrassments. Sharing the patients' personal information without the user's official consent is one critical privacy breach for the healthcare sector. The authors in [5] have summarized the issue of connected

healthcare and requirement of appropriate protections to safeguard the privacy of the patients and for minimizing the medical error. Therefore, an appropriate equity is needed to maintain privacy and security of data and the patient's personal space in healthcare.

In this paper, we survey and analyse the privacy and security issues in healthcare when using big data analytics. Based on the investigation, we propose an eHealth Data Privacy Model for enabling transparency in the flow of data over the network and that only the data relevant to a particular health service provider would be delivered. The model is based on the concept of Information Accountability [6] which enable patients to decide the usage of their data on a shared platform. It advocates transparency in the data usage and enable one with the ability to track the appropriate use of data under the predefined rules.

The paper is organized as follows. Section II describes the current state of the art in the area. We elaborate the concept of eHealth data, and survey the associated threats and vulnerabilities, the potential attack zones and how data are transmitted and received in the network. Section III analyses the key factors and issues related to the flow of information in an eHealth scenario. We explore the different aspects of information security related to healthcare and the user. We highlight the issues in eHealth for data privacy and formulate the need for a data privacy model. In Section IV, we propose the health data privacy model and the related functionalities, modules, protocols and required networking. Section V highlights the issues and challenges involved in implementing the proposed model. Section VI concludes the paper.

STATE OF THE ART

The healthcare sector spans over a vast landscape demanding cooperation and the active participation of public and private bodies, the individual user along with innovations and initiatives from various fields including marketing, finance, education and many more. The eHealth's objective is to avail medical services and amenities accessible and available at a reasonable cost and available resources while maintaining the quality of care and productivity. In an eHealth scenario, both, the patient and the medical service provider will be connected for the health monitoring, routine check-ups and even emergency services, facilitated by real-time secure data exchange. The healthcare industry dominates the data volume per person per day ratio generated. To handle data, Big Data can make significant revolution without resulting into additional infrastructure. It is an emerging technology for the future generations, which can analyse wide variety of voluminous data. It enables the processing of high-volume, high-velocity, and/or high variety of data aiding optimized results, better and efficient analysis, improved decisionmaking etc.

A. EHEALTH DATA AND DATA FLOW STRUCTURE

In an eHealth scenario, human and associated data are the most valuable and vulnerable assets. The medical reports generated electronically, called Electronic Health Records (EHR) are documents containing the patient's personal details (i.e., that have been used for registration over the network, the personal social security numbers used for the medical insurance, the medical reports, the diagnosis reports, the discharge summaries etc.) The medical data represents the patient–doctor relationship (e.g., the information including the patient-identification, the medical history, the digital renderings of the medical images, the treatment received, dietary habits, sexual preference, genetic information, psychological profiles, employment history, income, and physicians 'subjective assessments of personality etc. [3], [7].) Figure 9-1 shows the flow of medical data within the healthcare system. The patient shares their medical history, symptoms and personal identification details to the primary health services' unit. The primary health service unit then registers the patient using a unique identifier and creates the patients' file, which is shared with restrictions with the billing unit (relevant treatment charges and genuine user detail) and organization's IT unit (billing details and treatments' summary). The primary health services are responsible for the various test reports, clinical data, laboratory activities etc. and accordingly communicate with the secondary health services, pharmacy and regional health centers using patients' identifier and hiding other background details. The primary health services also share the details with its employees, which are strictly service-based i.e. ground staff are only exposed to the details like medication timings, test routines while doctors get detailed medical history irrespective of personal details. Third party IT services are used to store the health

record files. They have an access to patients' personal data (identifiers) and medical data (contributed by health organizations) and make the same available on demand.

The information transfer and storage in organization's and third party's IT servers are critical [4] as even a single careless activity can expose a patient's details. The data of the patients can be used to improve the efficiency within the healthcare system, to drive the public policy development and the administration at a state and federal level, and in the conduct of the research to advance the medical science apart from the personal care [8][9].

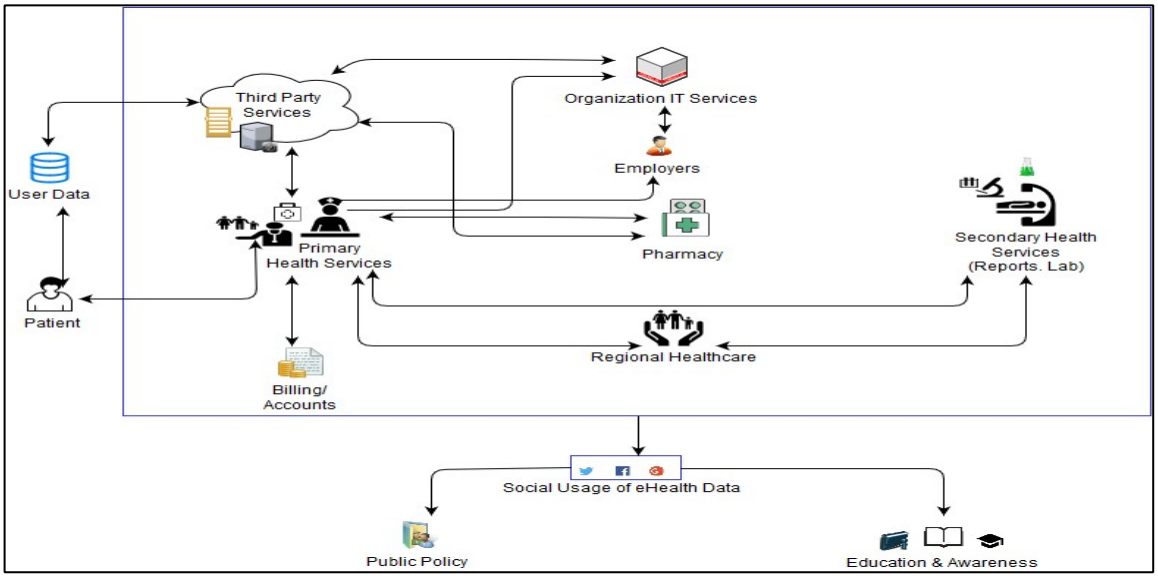

Figure 9-1 Data Flow in eHealth Scenario

B. DATA BREACHES

A medical data breach can lead to everything from an identity theft to billing fraud to blackmail, some breaches ultimately have little consequence on the patients affected. Whenever a medical data breach occurs, it signifies that there is a lack in security, while handling the information of the patient. In recent years, the survey [4], [10] recorded that 43 per cent of the total data breaches involved healthcare data. The healthcare breaches originated

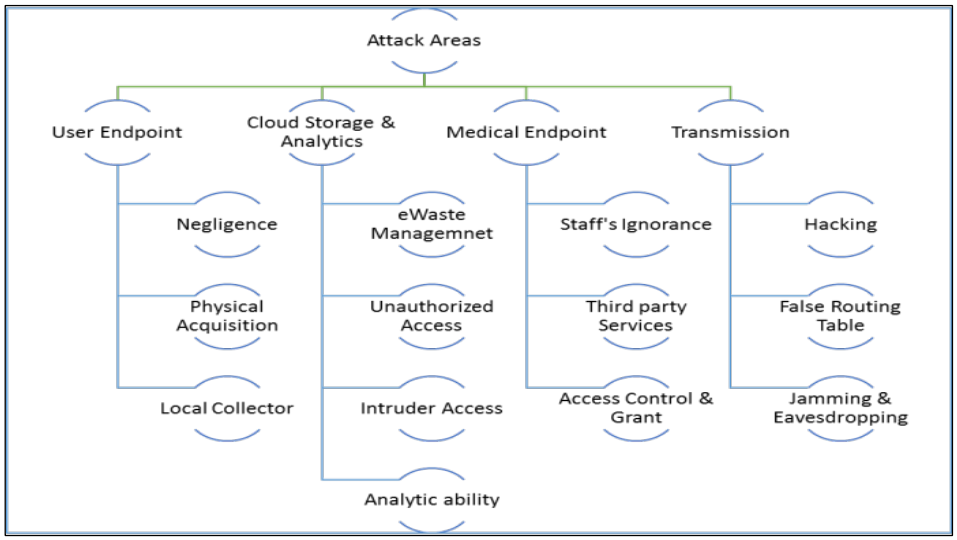

Figure 9-2 Attack Areas in Healthcare system

mostly from the service provider organization and/or a third party associate. It has been reported in [10], that about 90 percent of healthcare organizations had suffered data breaches, such as cyber attacks, employee mistakes, theft etc. Most common breaches are data and identity theft, unauthorized access, hacking the transmission, the loss of data in transmission, improper disposal, denial of service etc.

Figure 9-2 summarizes the vulnerable areas in each segment of the eHealth scenario cycle. The healthcare sector can be defined broadly into four sub-sectors, namely, theuserendpoint, the cloud storage and analytics, the medical endpoint and the tranmission. At the user end, the breaches mainly would be the result of ignorance or access to wearables, documents etc., unethically. At the storage and the analytics end, the unauthorized access, a compromised node/ employee, the lack of security measures and the improper disposal of data may cause medical breaches. The staff's negligence, unauthorised access, third-party service dependency are some of the root causes to breaches. Hacking or jamming the transmission of information also contribute to compromised and lost data.

C. THREATS TO PATIENTS' PRIVACY AND DATA SECURITY

A threat scenario is defined by the motives, resources, accessibility and technical capability. The threats to the privacy of the patient are a major outcome of the illegitimate usage of data either by internal/ external agents or by a third party agent in the data flow chain. The authors in [7], [8] summarized various threats to the user integrity and life. The threats impose different level of risk depending on the motive, the attack zone, the sensitivity and the mitigation and prevention strategies. The threats were categorized as the threats arising from accessing the patients' data inappropriately either by internal or external sources, and as the threats arising from data exposure over the network.

INFORMATION IN HEALTHCARE

The increased use of Web-based services has significantly raised the bars for the privacy concerns of the users. Published research has summarized the user content among a wide range of users, which include students, employed, senior citizens etc. The disclosure of data is user-dependent, i.e. a user agreement is needed in eHealth to verify the consolidation of the user to disclose personal data for research and development and/or other healthcare related needs. The current security and privacy in health data was summarized into the following subcategories [11-17].

A. DATA ACCESS AND SECURITY IN EHEALTH:

The healthcare institutions appeal to have security measures to govern the data access. Some of the common steps taken in that direction include access control systems, intrusion detection systems, policies etc. In [11], the authors have used a game theoretic approach to model the optimal levels of access. Remodeling the existing access control policies in the healthcare scenario is challenging apart from being highly expensive. In healthcare, we have individuals having different sets of roles, dependent data streams, independent data systems, dynamic configurations etc.

Data security requires concrete frameworks and defined protocols to identify, solve and mitigate the related security issues as stressed in [3], [5]. The emergence of ubiquitous access to patient data via mobile devices has exposed the vulnerabilities of the patients even further.

B. AUTHORIZED DATA DISCLOSURE AND INTEGRITY:

Attributing public health, the privacy policies should be made strongest when it comes to individual and communal interests. For each solution proposed, it should be carefully outweighed how much data gets disclosed and at what span [12]. Health services should be available on demand, which requires a full-time data protection.

Healthcare systems are getting more and more vulnerable to cyber security incidents nowadays. Factors like voluminous data generation; extensive usage of IT technologies to connect patients and healthcare utilities; data exposure over the network; diversified nature of healthcare systems; outdated applications and systems; poor

security algorithms; expansion in devices with enhanced capabilities and many more have contributed to the exponential increase in the number of incidents in the healthcare [16], [17].

C. EHEALTH AS A CRITICAL DATA PLATFORM

The above mentioned factors along with the data breaches makes the healthcare sector critical. A healthcare platform deals with asset classification and requirement to form a base layer of the healthcare system and the components may include index services of the user and/or service provider, registration proofs, identifiers etc. The data privacy, security and integrity involves network elements and storage (internal/external clouds). Access is determined using identifiers. Availability is the crucial among all as it can cost even life of the patient in case of emergencies.

D. CHALLENGES IN THE HEALTHCARE DATA

The authors of [14], [16] and [19] suggested that to maintain the users' privacy and in order to establish a balance in the economic constraints, quality of service and care and availability are the main challenges in healthcare. They advocated on the efficient and effective solutions for privacy maintenance at affordable and operative costs.

E. TRANSPARENCY IN HEALTHCARE

In healthcare sector, transparency is needed at every sector and individual end-point. Transparency has its individual definition from patient, doctor, healthcare organization, payers and providers [20]. From patients' perspective, transparency is needed in data acquisition and its usage and the cost for quality and services. The data collected from patients are in silos hence it becomes more important to check how data moves into the network both online and offline. With each bit of data comes an individual role and responsibility of data managers in healthcare sector. Data are subjected to limited access grants to protect patients' confidentiality.

F. DATA ANALYTICS

"Big Data" in health is defined as a voluminous complex and distributed data set, which imposes difficulty for conventional technologies in analysing and maintaining the information [19].

In order to safeguard the user privacy and tackle issues with interoperability and data repositories, we need a data model, which will be transparent, secure and able to analyse to produce the desired results. The data model should be able to manage widely distributed and scattered data. The data scheme should address user privacy and data sharing within agreement.

PROPOSED MODEL FOR DATA TRANSPARENCY

Health data are sensitive and demand appropriate and authenticated usage. While implementing digital data records, the required security measures for data storage, access and monitoring should also be put in place [14], [15]. Our proposed data privacy scheme allows for transparency in data-handling. It reinforces mechanisms to mitigate the illegitimate (unauthorised access, modification, disclosure to unintended user etc.) use of data [13], [21] and to best exploit the benefits gained from sharing the health data.

In our proposed data privacy model, the patients set the usage authority for their information and decide the extent, to which data can be aggregated with others. The data is shared under well formulated set of rules, guidelines and policies.

A. COMPONENTS AND PARTICIPANTS

The proposed model is shown in Figure 9-3. The following agents are the entities in the eHealth scenario;

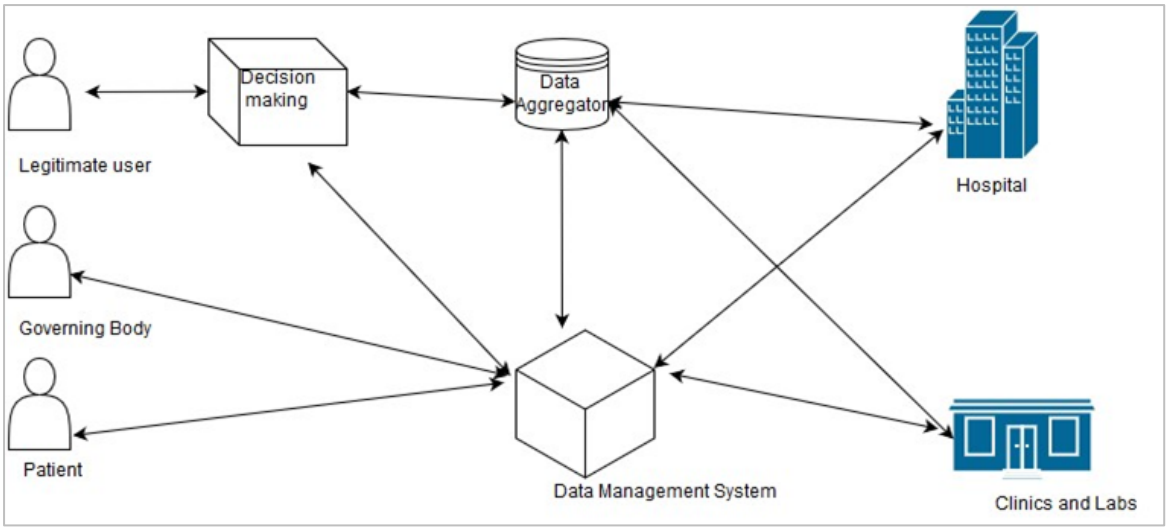

Figure 9-3 Proposed Data Privacy Model

Patient: Patients are the data owners. They can view the records of accesses made at any time and can submit their queries on unauthorized accesses.

Healthcare Utility: These are the healthcare authorities, responsible to aggregate the patients' personal data and medical history, test reports, medication briefs etc. These are often referred to as Data Providers. They are responsible to provide data to the data aggregator, which maintains, stores and analyzes using big data analytics and policies imposed by a system manager. They also maintain data logs to manage risks.

Governing Body: These system managers make and set policies to regulate the shared eHealth system and avoid misuse. These are responsible for the data integrity and log authenticity.

Data Users: Data users would make use of the aggregated data. Healthcare professionals, approved researchers, and government studies are examples of data users. Data users will be able to access specific log entries regarding their own access to patient information. They will be able to review the entries when they receive an inquiry requesting that they justify why they needed to access the relevant information in the given situation.

Data Aggregator: It collects information from the data providers. It works based on a policy set that considers the data provider and owner and allows aggregation of unobjectionable data.

Data Management System: It acts as the main governing body that maintains interoperability among various participants as well as the services in the eHealth scenario.

The Data Management System regulates and governs each of the related entities, to ensure that there is no hindrance to the patient's privacy. It sets usage policies that allow for inflow and out-flow of data for research-based applications, for maintaining logs, and for making the logs available for the patients to review anytime.

The data management system is responsible for establishing default policies and maintaining logs and other information for the governing body. The data aggregator and data management system coordinates while maintaining the retrieval policies and recording the logs events. The data management system also sets policies for the smooth functioning among the clinics, hospitals, laboratories and other medical end utilities. The data aggregator sends and receives data requests and responses from the laboratories and other users responsible for generating viable health information. The legitimate user sends requests to the decision making body to execute the request. The decision making body on receiving the request from a legitimate user, generates a query data and sends it to the data aggregator for approval.

B. DESCRIPTION AND WORKING PRINCIPLE

The proposed data protection model, shown in Figure 3, allows for transparency on how the patient's data are used within the healthcare system. The model guarantees no data usage without the users' consent or agreement. The model uses the secure key management scheme as proposed by the project MAGNET and MAGNET Beyond [23]. It explains a new key agreement protocol based on elliptical curve cryptography and personal public key infrastructure. The patients have control over the data access by intended or third-party users. The Governing Body, responsible for regulatory policies have power to grant/deny any request to access the user-data without hampering their privacy.

The model creates log for all the successful and unsuccessful accesses, which can serve as a database to validate future requests. The patients can refer to the log to check their access details periodically. The data management system monitors and controls all accesses together with the system manager, a patient and data aggregator. In the case of an unauthorized request, patient can go for inquiry and ask justification. The data management system is responsible to answer the user's inquiry.

The basic working principle is explained through the flowchart in Figure 9-4. The model works based on the policies that take into consideration the interests and concerns for the patient as well as the healthcare utilities. The following section describes briefly the Model's working;

GOVERNING POLICIES

Healthcare Utility: Data providers share their data under predefined set of usage and collection policies. These policies are user-friendly and do not interfere in the services provided. Policies would only govern the data usage and access for the healthcare development and facilitate research.

Patient: The user controls the personal data shared on the network for maintaining health and disease control by setting up policies, which let them, decide when, how and by whom the data is to be accessed. The policies also govern the amount of visibility depending upon the usage, purpose and motive. The users can invoke filters at their choice.

Governing Body: the system managers are the government bodies, which set up certain rules and regulations to maintain a social balance and to restrict privacy hindrance to the individual.

DATA COLLECTION AND AGGREGATION

In the model, a data aggregator collects information from the data providers and the data owners. Only the permissible data would be stored and analysed further for other purposes intending to develop an overall eHealth scenario.

REQUEST TO ACCESS DATA

When a data user executes a query in the system, the query service retrieves a policy for the data user. This can include rules regarding, which data they can access, how they can use data, and required de-identification of the results. If they are permitted to perform the query, the retrieved rules are then applied to filter the result set, removing restricted information. The information access request is logged, and the policy versions used to determine the access request is stored with the context-aware log entry.

MAINTAINING ACCESS RECORDS

The logs produced in an accountable system can contain sensitive information themselves and must be appropriately protected, including restricting who can view these logs and for what purpose.

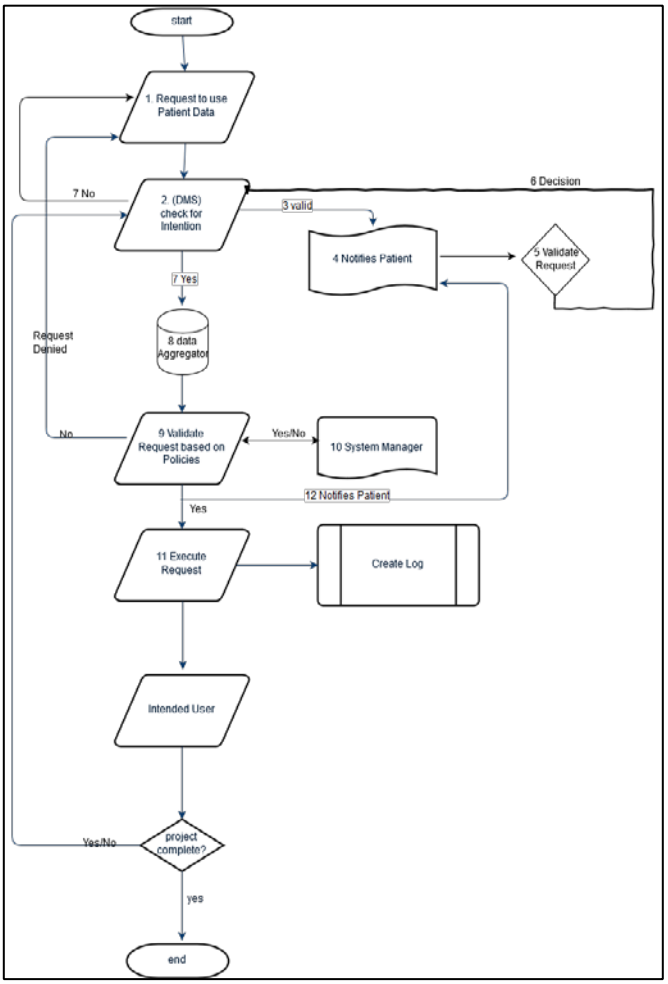

Figure 9-4 Working Principle of Proposed Data Privacy Model

IMPLEMENTATION

The implementation of the proposed model involves the formulation of rules and policies, data collection and aggregation, request to access data and creation and maintenance of access record-logs.

The governing policies are framed at three levels or working groups, comprising the healthcare utility, the patients and at the governing body as explained in the previous section. Each of the entities have their own requirements and accordingly set policies. For example, the doctors may share patients' symptoms and behavior for expert comments but the system would hide the personal data and identification; similarly, the patients would participate in guiding the usage of their data. The policies are generally depicted in the Open Digital Rights Language (ODRL) [22], which encourages the adoption of open international specifications for expressing policies in language. For data collection, aggregation and access, we will implant filters that will prevent restricted data access and maintain context-aware log entry. The model also incorporates a context aware security management scheme, which allows having virtual identities and including various agents to ensure trust, privacy, security and disclosed information. It authenticates uncompromised nodes and make decision on what to be shared keeping the patient aware of it.

CHALLENGES:

The proposed eHealth data privacy model would have to comply with the required scalability, heterogeneity and performance metrics assuring the data storage in the knowledge of the user and the concerned authorities.

The successful implementation of the proposed data privacy model in eHealth Big Data Analytics use cases requires to be overcome the challenges imposed by big data alone and also the adverse effects in healthcare sector.

Analyzing random as well as discrete data in eHealth will be complex and it will be difficult to maintain the exponential growth in the operation and computation time. The log maintenance will be difficult. Access grant against the query while maintaining its privacy will be a challenge. How an information is accessed, queried or stored including their log records will be a challenge. The major challenge for the implementation of the model is to accumulate and correlate the information coming from heterogeneous data sets.

CONCLUSION

The paper proposed a data privacy model in healthcare using big data analytics, added transparency in the data handling and accessing. The proposed model triggers the privacy issues in data aggregation and allowed access by maintaining logs and seeking due consents from the users who are the data owners. It also promotes the data sharing and risk mitigation in healthcare.

Future work will incorporate solutions to the imposed challenges in system scalability, interoperability, heterogeneity of data sets and implementation challenges.

REFERENCES

[1] Wang, W. and E. Krishnan, Big data and clinicians: a review on the state of the science. Journal Medical Informatics, 2014. 2: p. e1.

[2] Huser, Vojtech, and James J. Cimino. "Impending challenges for the use of Big Data." International Journal of Radiation Oncology• Biology• Physics (2015).

[3] Ball, Marion J., and Jennifer Lillis. "E-health: transforming the physician/patient relationship." International journal of medical informatics 61.1 (2001): 1-10

[4] "Data Breaches In Healthcare Totaled Over 112 Million Records In 2015." Ed. Dan Munro. N.p., n.d. Web. 28 Sept. 2016.

[5] Terry, Nicolas. "An eHealth diptych: the impact of privacy regulation on medical error and malpractice litigation." American journal of law & medicine 27 (2001).

[6] Weitzner, Daniel J., et al. "Information accountability." Communications of the ACM 51.6 (2008): 82-87.

[7] Hodge, James G. "Health information privacy and public health." The Journal of Law, Medicine & Ethics 31.4 (2003): 663-671.

[8] Campos, Maria João Magalhães Pereira. "Identity in eHealth-from the reality of physical identification to digital identification." (2012).

[9] J. J. Rodrigues, I. de la Torre, G. Fern´andez, and M. L´opez-Coronado, "Analysis of the security and privacy requirements of cloud-based electronic health records systems," Journal of medical Internet research, vol. 15, no. 8, 2013.

[10] By Greg Slabodkin. "Survey: No Cure In Sight for Healthcare Data Breaches." Information Management RSS. N.p., 2016. Web. 14 Sept. 2016.

[11] Zhao, X., and Johnson, M.E. (2008) ―Information Governance: Flexibility and Control through Escalation and Incentives,‖ Workshop on the Economics of Information Security, Hanover, NH

[12] Wilkowska, Wiktoria, and Martina Ziefle. "Privacy and data security in E-health: Requirements from the user's perspective." Health informatics journal 18.3 (2012): 191-201.

[13] Dimitra Liveri, Anna Sarri, Christina Skouloudi and ENISA. "Security and Resilience in EHealth: Security Challenges and Risks." Security and Resilience in EHealth: Security Challenges and Risks. European Union Agency for Network and Information Security (ENISA), 2015. Web. 13 Sept. 2016.

[14] N. H. Shah and J. D. Tenenbaum, "The coming age of data-driven medicine: translational bioinformatics' next frontier," Journal of the American Medical Informatics Association, vol. 19, no. e1, pp. e2–e4, 2012.

[15] L. P. Garrison Jr, "Universal health coverage–big thinking versus big data." Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research, vol. 16, no. 1 Suppl, p. S1, 2013.

[16] Appari, Ajit, and M. Eric Johnson. "Information security and privacy in healthcare: current state of research." International journal of Internet and enterprise management 6.4 (2010): 279-314.

[17] NRC National Research Council (1997) ―For the Record: Protecting Electronic Health Information

[18] J. Feigenbaum, A. D. Jaggard, and R. N. Wright, "Towards a formal model of accountability," in Proceedings of the 2011 workshop on New security paradigms workshop. ACM, 2011, pp. 45–56.

[19] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, "Information accountability," Communications of the ACM, vol. 51, no. 6, pp. 82–87, 2008.

[20] healthcatalyst. "3 Best Practices for Payer-Provider Collaboration to Improve Patient Care." Health Catalyst 3 Best Practices for PayerProvider Collaboration to Improve Patient Care Comments. N.p., 2016. Web. 10 Oct. 2016.

[21] R. H. Sloan and R. Warner, "Developing foundations for accountability systems: Informational norms and context-sensitive judgments," in Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies, ser. GTIP '10, ACM. New York, NY, USA: ACM, 2010, pp. 21–26.

[22] "Community & Business Groups." ODRL Community Group. N.p., n.d. Web. 27 Sept. 2016.

[23] Prasad, Ramjee, ed. My personal adaptive global NET (MAGNET). Berlin: Springer, 2010.