

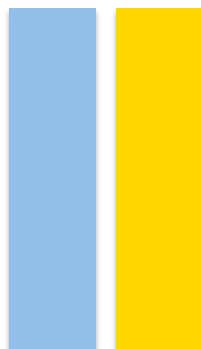
MESTRADO
INFORMÁTICA MÉDICA

Detection of illegitimate HIS access by healthcare professionals applying an audit trail-based detection model

Liliana Correia

M

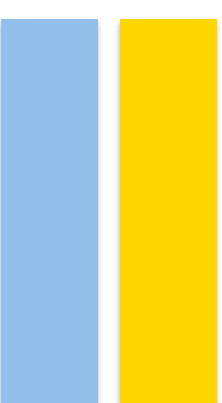
2019



SEDE ADMINISTRATIVA

FACULDADE DE MEDICINA

FACULDADE DE CIÊNCIAS





12^a ed

MIM

Detection of Illegitimate HIS access by healthcare professionals applying an audit trail-based detection model

Liliana Maria Barroso De Sá Cachada Correia

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

ORIENTADORES:

Pedro Pereira Rodrigues
Ricardo João Cruz Correia

SEP | 2019

*I dedicate this work to Ermelinda, my mum from heart,
a very special person who supported me in many different ways along this journey, and left us too early.*

Acknowledgments

In the first place, I would like to thank my advisors PhD Pedro Pereira Rodrigues and PhD Ricardo Correia for their guidance, challenges proposed and commitment along the course of this project. I am grateful for all their support, advice, criticism and friendship.

I also would like to thank Unidade Local de Saúde de Matosinhos (ULSM) for the opportunity given me to conduct this study within this prestigious institution. I would like to thank to Dr José Castanheira and Eng. Tiago Morais in particular for all their availability and contribution.

I want to thank to Dr. Rui Guimarães who conceded me some of his time to give me a background on data access laws and that was always available to help me with his knowledge.

I would like to acknowledge the importance of my employer HealthySystems during the project. I am grateful for the support of my colleagues and in particular, I would like to thank Duarte Ferreira whose contribution was very important.

I am grateful to all that in some way gave a little bit of their time to participate and contribute to this project.

Finally, I want to thank to my friends that always believed in me. A special thanks to my family for all the support, in particular to my husband Pedro Correia and my sons João Pedro and André that made this challenge possible and that always gave me strength to get here.

Abstract

Background: Complex health data management processes on healthcare institutions are typically carried out by many users, with different roles that have a certain relation with a patient in a certain moment which impose serious issues to demonstrate compliance with personal data protection laws. This study aims to create a platform that can help healthcare institutions in detecting illegitimate accesses in a short time and act preventively in future incidents.

Methods: Firstly, we gathered information through a literature review about data protection laws and guidelines that should be followed. We promoted discussions with experts and personalities responsible for data management and protection on the health sector in order to learn about their concerns. We interviewed healthcare professionals to learn about their routines and HIS access practices that could help us to describe scenarios of undue access and create rules to classify accesses as normal or suspicious. Secondly, undue access scenarios were described and use cases of activities on HIS were modelled and discussed with the experts to evaluate their applicability, followed by sequence diagrams to show systems interactions, data flow and checks that must be performed to identify illegitimate accesses. Then we matched variables to be analysed with the requirements of Ministers Council Resolution (MCR) nr.41/2018. Thirdly, we chose three use cases to implement, based on the availability of applicational logs, modelled the algorithms with activity diagrams and coded with JAVA programming language. Lastly, we evaluated the proof of concept comparing the total results with the number of accesses classified as suspicious. We used R Language to calculate absolute values, percentages and dispersion measures.

Results: We described eleven undue access scenarios, modelled use cases and sequence diagrams for each. Considering the logs available on the audit trail we implemented three use cases, having produced three algorithms applying the rules based on the information collected. Access classification by our system was in accordance to rules applied. “Check time of activity” use case had 63,8% of suspicious access being 55% of activity period shorter and 9,78% longer than expected. “Check days of activity” use case presented 2,27% of suspicious access and “EHR read access” 79%, the highest percentage of suspicious accesses.

Discussion: The results shown are the first picture of the accesses made during nine days in a system used on hospital stay, consulting and emergency context. A deeper analysis to evaluate algorithms sensibility and specificity should be done. Lack of more detailed information about professional routines and procedures, low quality of systems logs and few systems behaviour knowledge are some limitations of this study. Although we believe that this is an important step to go ahead in this field. Refining the algorithms, producing others for use cases left and creating a knowledge base to apply artificial intelligence for illegitimate access detection are for future work.

Keywords: HIS access; patient data access; undue access; data breach; illegitimate access detection.

Resumo

Introdução: Os complexos processos de gestão de dados em instituições de saúde, normalmente realizados por muitos utilizadores, com diferentes funções que têm uma certa relação com um paciente em determinado momento impõe sérios problemas à demonstração da conformidade com as leis de proteção de dados pessoais. O desafio é criar uma plataforma que ajude as instituições de saúde a detectar acessos ilegítimos num curto espaço de tempo e agir preventivamente em incidentes futuros.

Métodos: Inicialmente, realizou-se uma revisão de literatura sobre leis e diretrizes de proteção de dados que as instituições de saúde devem cumprir. Promoveu-se discussões com especialistas e responsáveis pela gestão e proteção de dados neste setor. Realizou-se entrevistas a profissionais da área para conhecer as rotinas e práticas de acesso aos Sistemas de Informação da Saúde, descrever cenários de acessos indevidos e criar regras de classificação dos acessos como normais ou suspeitos. Em seguida, descreveu-se cenários de acesso indevido, modelou-se casos de uso que permitissem a sua detecção e discutimo-los com especialistas para validar sua aplicabilidade. Seguiu-se a produção dos respetivos diagramas de sequência para mostrar as interações entre utilizadores e sistemas, fluxo dos dados e verificações que devem ser realizadas para identificar os acessos suspeitos. Cruzou-se posteriormente as variáveis a usar com as dos requisitos específicos da Resolução do Conselho de Ministros (RCM) nº 41/2018. Foram escolhidos três casos de uso a implementar com base nos logs aplicativos disponíveis, modelando os algoritmos com diagramas de atividade e codificando-os em linguagem de programação JAVA. Por fim, avaliamos a prova de conceito comparando os totais dos resultados obtidos com o número de acessos classificados como suspeitos, usando linguagem R para cálculo de valores absolutos, percentagens e medidas de dispersão.

Resultados: Descreveu-se onze cenários de acesso indevido, e modelou-se onze casos de uso e diagramas de sequência. Produziu-se três algoritmos aplicando regras baseadas nas informações prévias, classificando os acessos. No caso de uso “Tempo de atividade”, 63,8% foram classificados como acesso suspeito, sendo que 55% foi devido ao período de atividade ser inferior e 9,78% superior ao esperado. O caso de uso “Dias de atividade” apresentou 2,27% de acessos suspeitos e “acesso de leitura ao registo clínico eletrónico” 79%, a maior percentagem de classificação de acesso suspeito.

Discussão: Os resultados apresentados são a primeira fotografia dos acessos realizados durante nove dias num sistema utilizado para contexto de consultas, internamentos e urgência. É necessário realizar uma análise mais detalhada para avaliar a sensibilidade e especificidade dos algoritmos. Pouco pormenor sobre rotinas e procedimentos dos profissionais, baixa qualidade dos logs dos sistemas e pouco conhecimento do comportamento dos sistemas são algumas das limitações deste estudo. No entanto, consideramos que este foi um passo importante para avançar neste campo. Refinar os algoritmos, produzir outros para os casos de uso restantes e criar uma base de conhecimento para aplicar modelos de inteligência artificial são trabalho futuro.

Palavras-chave: acesso aos Sistemas de Informação na Saúde; acesso aos dados do paciente; acesso indevido; violação de dados; detecção de acesso ilegítimo.

Preamble

Fifteen years after I finished the Informatics/Applied Mathematics Degree, in which period I worked as project manager in training sector, I decided to take a break in my career, and start new challenges. I aimed to ingress in the IT sector and I looked up to enhance my knowledge by applying it in new areas. The Master Degree in Medical Informatics was a good opportunity to put it into practice.

During the two years that I worked in HLTSYS (HealthySystems), a company expert on the health sector that acts on cibersecurity, interoperability, data quality and data protection, and after the first year of master degree, I acquired a great interest about how health data collected on healthcare institutions is treated, accessed and protected. I also found that there is a gap in all these processes when Information Systems departments, in fact, do not have complete control of the actions performed on HIS. Who, when, how and why professionals access the electronic clinical records of patients is valuable information to ensure the security of sensitive data in healthcare institution guard.

Data breaches are a concern for all kinds of organizations. In healthcare institutions it is even more concerning because of the data type that they are responsible for. It may have a huge impact for people affected and their future generations. This topic is reported on the General Data Protection Regulation, on the Joint Commission International Accreditation Standards for Hospital and in other documents as well, reinforcing the importance for healthcare institutions to address this issue.

As soon as I realized this gap and the importance of a tool that could help identify problems related to data breaches, I proposed the scope of this project that I thought would be very interesting, valuable and innovative. In addition, it would oblige me to develop skills in areas such as data science, data privacy and software development on the health field. After discussing the topic with Professor Pedro Pereira Rodrigues and Professor Ricardo Correia, and after having their positive feedback, we defined the plan for this work, which began with an understanding of access to EHR.

I would like this work to contribute to fulfilment of legal requirements by healthcare institutions, and more than that, I hope that healthcare institutions can have an overview of what is happening with access to patients' EHR: who, when, how and for what purpose.

Scientific Results

The outcomes of the work described in this thesis were published in the following article:

L. Sa-Correia, R. Cruz-Correia, P. Rodrigues, "Illegitimate HIS access by healthcare professionals: scenarios, use cases and audit trail-based detection model", 2019 CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, Tunisia, 2019

At the date of delivery of this dissertation, the article was accepted on 23rd June 2019. The full article is presented in annex 7.1.

Table of Contents

Acknowledgments	vii
Abstract	ix
Resumo	xi
Preamble	xiii
Scientific Results	xv
Table of Contents	xvii
List of Figures	xix
List of Tables	xxi
Acronyms	xxiii
1 Introduction	3
1.1 State of the Art	4
1.1.1 Access to Electronic Health Record	4
1.1.2 Audit Trails	5
1.1.3 Data Breach Detection Systems	6
1.2 Research Question	7
2 Methods	11
2.1 Information gathering	11
2.2 Functional Analysis	12
2.3 Implementation	13
2.4 Evaluation	14
2.5 Study Authorizations	14
3 Results	17
3.1 Introduction	17
3.2 Information gathering	17
3.3 Functional Analysis	20
3.3.1 Scenarios, Use cases and Sequence diagrams	20
3.3.2 Logs Variables	32
3.4 Implementation	33
3.5 Evaluation	38
4 Discussion	47

5	Conclusion and Future Work	51
6	References	55
7	Annexes	61
7.1	Published article in HCIST2019 conference proceedings.	62
7.2	Interview Script	70
7.3	ULSM application for study authorization	72
7.4	ULSM study authorization	76
7.5	CHUP application for study authorization	79
7.6	CHUP study authorization	86

List of Figures

Figure 3.1 Use case 1 - Check failed access attempts	20
Figure 3.2 Sequence diagram 1 - Check failed access attempts	21
Figure 3.3 Use case 2 - Check wrong profile user	22
Figure 3.4 Sequence diagram 2 - Check wrong profile user	22
Figure 3.5 Use case 3 - Check professional off service access	23
Figure 3.6 Sequence diagram 3 - Check professional off service access	23
Figure 3.7 Use case 4 - Check time of activity	24
Figure 3.8 Sequence diagram 4 - Check time of activity	24
Figure 3.9 Use case 5 - Check days of activity	25
Figure 3.10 Sequence diagram 5 - Check days of activity	25
Figure 3.11 Use case 6 - Check inactive professional	26
Figure 3.12 Sequence diagram 6 - Check inactive professional	26
Figure 3.13 Use case 7 - Check simultaneous access	27
Figure 3.14 Sequence diagram 7 - Check simultaneous access	27
Figure 3.15 Use case 8 - Check access to Electronic Health Records (EHR) of healthcare institution employee	28
Figure 3.16 Sequence diagram 8 - Check access to EHR of healthcare institution employee	28
Figure 3.17 Use case 9 - Check remote access	29
Figure 3.18 Sequence diagram 9 - Check remote access	29
Figure 3.19 Use case 10 - Check EHR read access	30
Figure 3.20 Sequence diagram 10 - Check EHR read access	30
Figure 3.21 Use case 11 - Check access to EHR analysis based on a set of variables	31
Figure 3.22 Sequence diagram 11 - Check access to EHR analysis based on a set of variables	31
Figure 3.23 Activity diagram for UC4: Check time of activity	34
Figure 3.24 Activity diagram for UC5: Check days of activity	35
Figure 3.25 Activity diagram for UC10: Check EHR read access	37
Figure 3.26 Consecutive time of activity in minutes on Health Information Systems (HIS) by professional category	38
Figure 3.27 Consecutive time of activity in minutes on HIS by professional category related to healthcare delivery	39
Figure 3.28 Consecutive days of work by professional category	40
Figure 3.29 Consecutive days of work by healthcare delivery professional category	41

Figure 3.30Results of EHR read access	42
Figure 3.31Adjusted results of access analysis	43
Figure 3.32Results of access by professional category	44
Figure 3.33Results of access by professional category	44

List of Tables

Table 2.1	Interview script	12
Table 3.1	Interview results	18
Table 3.2	Motivations for access HIS with another professionals' credentials. Each interviewee could order up to 4 reasons based on the frequency.	19
Table 3.3	HIS access practices by professionals in general	19
Table 3.4	Relation between log sources and variables, and the UCs.	32
Table 3.5	Results for UC4: Check time of activity by professional category	38
Table 3.6	Results for UC5: Check days of activity	40
Table 3.7	Results of EHR read access	41
Table 3.8	Adjusted results obtained	42
Table 3.9	Number of results obtained by professional category	43
Table 3.10	Number of results obtained for professional categories directly involved in care delivery	44

Acronyms

AT Audit trail

CDTT Complementary Diagnosis Test and Therapeutic

CHUP Centro Hospitalar Universitário do Porto

CIDES Department of Health Information and Decision Sciences

CINTESIS Center for Research in Health Technologies and Information Systems

CSV Comma Seperate Values

DAR Data Access Responsible

DPO Data Protection Officer

DTT Diagnostic and Therapeutic Technician

EHR Electronic Health Records

FMUP Faculty of Medicine of the University do Porto

GDPR General Data Protection Regulation

HIPAA Health Insurance Portability and Accountability Act

HIS Health Information Systems

AI Artificial Intelligence

IDS Intrusion Detection Systems

IS Information Systems

ISD Information Systems Director

MOI Management of Information

MCR Ministers Council Resolution

OS Operating System

SAT Social Assistant Technician

UC Use Cases

ULSM Unidade Local de Saúde de Matosinhos

UML Unified Model Language



Introduction

1. Introduction

Since computers arrived to healthcare institutions and have become part of healthcare delivery, notable advances were made in providing care. HIS have proliferated in healthcare facilities and the amount of electronic health data has increased dramatically since then. A new paradigm was imposed to the practice of medicine and health research thanks to the quantity and quality of information available in this domain (Cimino and Shortliffe, 2006). EHR have been developed to integrate patient data and help healthcare professionals in making the best decisions (Cruz-Correia et al., 2007). Despite the benefits achieved with all this data collection, several security challenges arise and healthcare institutions have to face them daily, namely, ensuring patient privacy, and data integrity, confidentiality and availability (Ferreira et al., 2007; Fernández-Alemán et al., 2013).

Concern for personal data protection has existed for long time and it has been on many Governments agendas. Laws such Data Protection Act (DPA, 1995) and Health Insurance Portability and Accountability Act (HIPAA) (104th Congress, 1996) date from 1995/1996. Nonetheless, in May 2018 the General Data Protection Regulation (GDPR) (Council of the European Union, 2016) came to reinforce the processes of personal data protection. The new GDPR, issued by European Commission, aims to standardize the laws related to personal data protection in Europe and brings up the reinforcement of the citizen rights and new obligations for companies. These obligations are the challenges that arise for healthcare organizations due to the great amounts of sensitive data that they process in various HIS and they need to demonstrate that they are compliant with the GDPR. Although the former law contains many of the issues addressed by the new regulation, this one brings a change of paradigm (Council of the European Union, 2016; SPMS, 2017). Healthcare organizations acquire greater responsibility in the way they protect the subjects' personal data. Data Protection Officer (DPO), the role created by GDPR, is responsible for data protection inside the healthcare institution, overseeing the strategy and procedures to ensure data protection and compliance with GDPR requirements.

One of the major obligations imposed to companies is the need to report any data breach or loss of information. "Personal data breach" is defined by the GDPR as a breach of security, leading to the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored or otherwise processed. As soon as the controller becomes aware that a personal data breach has occurred, the controller should notify the personal data breach to the supervisory authority without undue delay and, when feasible, not later than 72 hours after having become aware of it; unless the controller is able to demonstrate, in accordance with the accountability principle, that the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. Where such notification cannot be achieved within 72 hours, the reasons for the delay should accompany the notification and information must be provided as soon as possible (European Union, 2016).

Health data is very appealing for many reasons and so is the target of frequent attacks (Donovan, 2018). Despite it is known that there is a great concern about external attacks to HIS, studies show that there is, also, a low awareness of the potential risks associated to insiders malfeasance (Wikina, 2014; Valli, 2006; Morrow, 2016; Barrows and Clayton, 1996) .

Due to the large number of HIS in institutions, that process and manage great amounts of health data, and the complex data flow, healthcare institutions have serious issues to demonstrate that they are compliant with the GDPR. Audit trail (AT) can be very useful on this topic. They comply important GDPR requirements, such as audit and traceability of actions on patients' data (Gonçalves-Ferreira et al., 2018) and have been acquired by Information Systems (IS) services. However, these repositories are used in most cases as a forensic investigation tool, once a complaint occurs (Li and Oprea, 2016) and do not prevent or alert for possible undue accesses. Such occurrences may be data breaches and should be communicated to authorities and even to the affected person, in the terms of the law, as we saw previously (European Union, 2016). Moreover, if healthcare institutions have this knowledge, they can preventively act to avoid future incidents.

Healthcare typically implies complex data management processes by many different users, with multiple roles (e.g. physician, head of department, nurse, researcher, auditor) of many patients with whom they have momentarily a relation in a context of healthcare delivery, clinical research, education, among others, are every day implied in healthcare delivery environment. Distinguishing non-legitimate access from legitimate is a very hard task in this environment. Focused on this problem, we propose the development of a platform to detect suspicious activity on HIS through an audit trail-based detection model, which will provide grounds for the DPO to effectively and efficiently investigate potential data breaches inside the healthcare organization.

1.1 State of the Art

1.1.1 Access to Electronic Health Record

Personal data protection is a Portuguese governmental concern since long time. The Law nr. 67/98 from 26th October 1998, Portuguese data protection law, has been updated and rectified throughout the past years, being presented in its most recent version as Law nr. 103/2015, from 24th August. (lei, 1998). This legislation was recently reinforced by the new GDPR, which brings a shift of paradigm on responsibility attributions about legal compliance and also about the creation of heavy sanctions that have aroused organizations' interest to this issue (European Union, 2016). In the particular case of health data, the complexity of the problem increases greatly. In fact, if we consider that when someone unduly accesses to another subject's personal data, it can cause serious damages to that subject, putting at risk his future. But the debauchery of a subject's clinical records can bring harmful consequences not only to the individual himself, but also to the coming generations. Health data is especially sensitive, because it concerns the deeper intimacy of people, who have rights protected by the legal system: privacy, secrecy and confidentiality (Guimarães et al., 2018). The Law nr. 26/2016 from 22nd August, in the Article 38, determines that, relatively to undue access to nominative data (AR, 2016):

“1 - Any person who, with the intention of improperly accessing nominative data, falsely declares or testifies before an organ or entity referred to in article 4, paragraph 1, has a direct, personal, legitimate

and constitutionally protected interest justifying access to the information or documents intended, shall be punished with imprisonment for up to one year or with a fine.

2 - The attempt is punishable.”

The Article 3 of the same document defines a nominative document as a “*document containing personal data defined in accordance with the legal regime for the protection of personal data, in terms of the legal regime for the personal data protection;*”.

In its turn, personal data is all kind of data that refers to a subject including sensitive data, such as health data (European Union, 2016). In this respect, it is the responsibility of healthcare institution system to prevent unwarranted access of third parties to clinical procedures, thus fulfilling the requirements of the legislation governing the personal data protection, as can be seen from STA0394/18. Currently, the reality of information in most of public healthcare providers, in what concern to health records, is characterized by extensive kilometers of paper-based medical records scattered across the hospital’s various floors, and also outside hospitals, on guard duty companies specialized in the custody of such assets; dozens of repositories and databases, with clinical records, including Complementary Diagnosis Test and Therapeutic (CDTT) imaging; thousands of microfilm reels without optical search and with files in cardboard sheets; and so on. Such is the information complexity and dispersion on the custody of healthcare institutions that makes the responsibility to prevent undue access a challenge without precedent.

The scope of this study is precisely about the detection of improper accesses to EHR. An EHR is a repository of electronically maintained information about an individual’s health status and healthcare, such that it can serve the multiple legitimate uses and users of the record. Typically, an EHR used to be a record of care provided when a patient was ill, but nowadays, as health care is evolving to encourage healthcare providers focusing on the continuum of health and healthcare, it aggregates information about illness, recovery and wellness as well (Cimino and Shortliffe, 2006). Healthcare providers present a vast diversity of IS, some of them are administrative and billing, others are for image processing and others are pure EHR. This means that healthcare providers keep at their responsibility lots of sensitive information about the past, the present and the future of the data owners in many different ways and formats. The dispersion of patients’ clinical information is well known to all those who deal with HIS. It is also known that as more professionals access to HIS more is the risk of information breach. In fact not only are the professionals working inside the hospitals that access to EHR, there is also the risk of remote access by third party companies. Although healthcare institutions are taking the first steps in traceability and audit compliance, they are not yet equipped with tools that enable them to more effectively identify illegitimate accesses or potential inappropriate accesses to EHR. How can we know who accesses, with what authorization, on what date, with what equipment, what was done, what information was accessed, what resulted from that action, in what context, etc.? How can we monitor thousands of hits per day and identify which are inappropriate?

1.1.2 Audit Trails

A log is a collection of data generated by an event on an information system each time an action is executed by a user or by the system itself. Log information usually contains legal required data, such as date and time, the user account, source and destination request, action performed, data accessed, among other data that can be included (Halamka et al., 1997). Audit trail is a repository of events information

that aggregates logs of one or more sources, which allows the traciability and auditing of actions on HIS. They can be used for various purposes such as access management, monitoring of employee behaviour or computer/information system failures (Cruz-Correia et al., 2013). They provide a historical record of users activity based on a sequence of events and act as proofs of compliance and operational integrity. Audit trails can also identify areas of non-compliance by providing information for audit investigations. Currently, audit trails are essential in healthcare organizations to demonstrate compliance with specific security requirements. The HIPAA and the GDPR require the monitoring healthcare data access by organizations (who views the EHR) as well as data integrity (data cannot be modified without detection), to protect the security of patient data in HIS (European Union, 2016; 104th Congress, 1996; Barrows and Clayton, 1996). Healthcare institutions store massive data of systems activity in these repositories and most of the time it is used for forensic audits when a complaint occurs and do not prevent or alert for possible undue accesses (Li and Oprea, 2016). However, this information could be very useful to act preventively and avoid some kind of security incidentes.

1.1.3 Data Breach Detection Systems

The cyber attacks are a great concern of enterprises nowadays, and are continuously evolving and becoming each time more sophisticated. The community for cybersecurity is working hard to protect companies against threats and breaches (Coburn, 2018).

There are a variety of security defences such as firewalls, Intrusion Detection Systems (IDS), anti-virus software, and other systems to keep the information safe, log analysis systems, and more. They use sophisticated algorithms for detection of anomalous and malicious activities that can lead to breaches. In this context, several algorithms are applied based on detection of anomalies or outliers in a dataset. The definition given by Grubbs in 1969 for outlier was (Grubbs, 1969): “An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs”. Markus Goldstein extended this definition and assigned to anomalies two important characteristics (Goldstein and Uchida, 2016): (1) “Anomalies are different from the norm with respect to their features” and (2) “They are rare in a dataset compared to normal instances”. Anomaly detection algorithms are used in various domains like intrusion detection, fraud detection, Data Leakage Prevention, among others (Goldstein and Uchida, 2016).

Intrusion detection is used for monitoring of network traffic and server applications, through the analysis of great amounts of data in almost real-time applying outliers detection algorithms that should detect intrusion attempts and exploits (Goldstein and Uchida, 2016; Ng et al., 2015).

Fraud detection typically implies the analysis of log data to detect misuses of a system or suspicious events indicating fraud. In particular for detection of fraudulent accounting, algorithms analyse financial transactions, and credit card payments logs can be used to detect misused or stolen credit cards. Fraud detection is becoming more important in this area due to the increase of internet payment systems (Delamaire et al., 2009; Goldstein and Uchida, 2016; Issa and Vasarhelyi, 2011).

Data Leakage Prevention is used to detect loss of sensitive data at an early stage focusing on near-real-time analysis such that it serves as a precaution method. In this context, the logs of databases access, file servers and other information sources are analysed in order to detect uncommon access patterns (Goldstein and Uchida, 2016; Hauer, 2015).

There are diverse algorithms of anomalies detection that can be divided in different types: (i) by Classification consisting of two phases (training phase and test phase) and can be divided into multiclass or single class (Issa and Vasarhelyi, 2011); (ii) by Clustering which assumes that anomalous objects arise clustered in smaller clusters away from the centroid cluster, or not belonging to any cluster (Issa and Vasarhelyi, 2011); (iii) distance from the nearest neighbour (k-NN) in which it is assumed that the normal data has a dense neighbourhood, and that the anomalies are away from their neighbours (Ramaswamy et al., 2000), (iv) Statistical algorithms related to probability calculations (Obikee et al., 2014), (v) Information Theory Algorithms (Costa, 2014), (vi) Algorithms of spectral methods (Jolliffe, 1986; Sun et al., 2008). For each type of algorithms are presented various methods, which can be combined to get the best expected result (Lobato et al., 2016).

Some methods that apply the Classification algorithms are Neural Networks, Bayesian Networks, Support Vector Machines and Rule-Base method. The Cluster-Based Local Outlier Factor (CBLOF) and the Local Density Cluster-Based Outlier Factor (LDCOF) methods are two examples of Clustering algorithms' applications. Anomalies detection by the nearest neighbour (k-NN) presents two approaches, the global perspective with the application of the k-nearest neighbour (k-NN) method and the local perspective with the Local Outlier Factor (LOF). In the statistical methods approach, parametric methods such as models based on normal distribution, regression or mixed parameters, or non-parametric methods such as histograms or Kernel functions can be used. The information theory methods use Kolmogorov's theories of Complexity, Entropy and Relative Entropy. In the case of anomaly detection by spectral methods, Principal Component Analysis (PCA), Compact Matrix Decomposition (CMD) or Single Value Decomposition (SVD) are the most used. These are powerful algorithms that are used independently or combined and have great results in threats detection (Lobato et al., 2016). However, the systems mentioned before are generally deployed to detect external intrusions and do not address the problem of insider threats.

1.2 Research Question

The problem we want to address is whether it is possible to detect improper access by authorized staff to electronic health records by identifying standards outside of what is considered the normal routine of healthcare professionals, through systems logs analysis. So, the research question we want to answer is:

Can illegitimate HIS access by healthcare professionals be detected applying a audit trail-based detection model?



Methods

2. Methods

2.1 Information gathering

To gather information about the legitimacy of access to health data in the beginning of the project it was done a literature review focused on laws, guidelines and technical documents about HIS. Then we discussed the theme of undue health data access on meetings with the Information Systems Director (ISD) and DPO from healthcare institutions and with an expert jurist on data access laws that is also Data Access Responsible (DAR) of a hospital. The conversations took place in two separate moments. The first meeting was with a jurist expert on personal data access laws where we discussed what is an illegitimate access and its consequences. After this meeting we discussed the theme with the ISD and DPO from healthcare institutions where we discussed how is managed the access to HIS. From this discussion we pointed out some concerns about undue access to HIS data.

Secondly, in order to better understand the HIS access routines by professionals and thus describe scenarios of undue access to health data, a survey based on interviews with eleven questions of qualitative and open answers, to professionals from healthcare institutions was conducted. The information collected was about work routines and HIS access practices: (1) professional category of the interviewed, (2) consecutive hours and days of work in order to define limits for existing activity on HIS by users, including exceptional shifts; (3) HIS used by each professional's category and (4) how they access them, to identify misused profiles or misconfigured permissions; (5) computers used for HIS access; (6) the permissions for remote access in order to understand who can access from outside the healthcare institution and for what purpose; (7) type of information accessed, to match with profile permission; (8) access limitations felt during their activities and how they get over them, to identify abnormal accesses motivated by task accomplishment like credential sharing; (9) motivations for EHR access that can explain the use of another professional's credential, (10) practices adopted by professionals in general and (11) other experiences that the interviewed wish to report so that we could detected different scenarios from the ones we initially previewed at moment of discussion with the experts. Table 2.1 shows the topics of the questions applied (See also interview script in the annex 7.2).

Table 2.1: Interview script

Type of Information	Question
Professional	Professional Category
Routines of professionals	Number of hours worked on a regular shift
	Number of hours worked on a exceptional shift
	Number of consecutive days worked in regular weeks
	Number of consecutive days worked in exceptional weeks
Information System (IS) access	Information systems accessed for task accomplishment
	Access done by credentials (username/password) (yes, no)
	If yes: the username is the same for all IS (yes, no)
	If not: how is the authentication done
	The user uses allways the same computer to work (yes, no)
Remote access	A computer is shared by many professionals (yes, no)
	Possibility of remote access (yes, no)
Data accessed	If yes, indicate the reasons for remote access
	IS accessed contain health data (yes, no)
	If yes, indicate which IS contain health data
	There are constraints to access data needed (yes, no)
Reason for access using credentials of another professional (select by frequency order)	If yes, how is it solved
	Need to access data that own credentials do not allow
	Give information to family, friends and acquaintances
	Curiosity
	Help other professionals on their tasks
Actions that may originate undue or suspicious access	Identification of other reason
	Sharing of credentials
	Use credentials of inactive professionals
	Access to data of patients without context
	Execute tasks demanded by other professionals
	Usurpation of credentials
	Identification of other actions that are known

In order to conduct the interviews inside the healthcare institutions we applied for a study authorization in two healthcare institutions that presented different care settings (primary, secondary and tertiary care). So we apply for study authorization on ULSM and Centro Hospitalar Universitário do Porto (CHUP). We got positive answer from ULSM on time but for CHUP we got positive answer four months latter. As we could not wait for this decision, we decided to choose a convenience sample of volunteers to interview from three public hospitals from North Portugal including ULSM, based on the volunteer's availability, so that it included at least two representatives per profession (physicians, nurses, administrative staff, others) and type of care (primary, secondary and tertiary). Twelve (12) interviewees were from ULSM and were done in ULSM facilities and sixteen (16) were volunteers from other hospitals and were done outside the healthcare facilities. The interviews were applied between November 2018 and February 2019.

2.2 Functional Analysis

In the functional analysis phase, based on the information collected previously, we wrote scenarios for each situation identified as illegitimate access, or that could become an illegitimate access to EHR by professionals of healthcare institutions. For each scenario a use case was designed and described in which

way it is, or could turn, an illegitimate access. In the end of this process, the use cases produced were discussed and validated by the ISD and DPO of ULSM. Next step was to produce a sequence diagram for each use case to show systems interactions, the data flow and checks that must be performed to identify the suspicious accesses. For requirement specification it was used Unified Model Language (UML) (Booch et al., 1996).

From the use cases and sequence diagrams modelled, an analysis was performed to identify variables that would be necessary to trace, in order to detect the suspicious accesses, and were then matched with Ministers Council Resolution (MCR) nr 41/2018 document, taking advantage of mandatory data from systems logs (Conselho de Ministros, 2018).

2.3 Implementation

In the implementation phase we developed a proof of concept for three use cases modelled, producing algorithms for illegitimate access detection. The selected use cases for implementation were chosen taking into account the logs' data collected at the moment of implementation and that were available to be analysed. For this reason we have decided to implement UC4: Check time of activity; UC5: Check days of activity and UC10: EHR read access, and leave more ahead the development of use cases that depended on information which was not yet collected or that needed previous treatment. The representation of algorithms was done using activity diagrams, applying UML specification.

For development of the algorithm it was used the logs available on audit trail HS.Register installed on ULSM. We used the logs of Obscare system because it was the only applicational logs ready to use. However, Obscare is used on hospital stay, consulting and emergency context, which includes the different contexts of care. Before the development it was done a research on a demo of HS.Register that intended to find out the type of data that could be found on the logs and to have a reference of the variables to use on the development. The rules applied for decision and classification of activities by professionals on HIS were based on the information about the routines of healthcare professionals and the type of data available on Obscare's logs.

The program language that we used to develop the prototype of suspicious access detection was JAVA. This option was based on its characteristics once JAVA is multiplatform, and so, it is independent of the Operating System (OS), and with this programming language it is not necessary to rewrite the program code when installing it in another OS, since the conversion is done by the Java Virtual Machine in a quasi-automatic process, which, commercially speaking, is very beneficial (Arnold et al., 2005). The connection to Elasticsearch, technology used by HS.Register, was done using the Elasticsearch API for this purpose.

The information obtained was saved on a Comma Seperate Values (CSV) document, classifying accesses as "suspicious access" or "normal access". This report file was the outcome of the implementation phase.

2.4 Evaluation

The proof of concept developed were tested in real environment using logs of Obscare system that were being collected by the audit trail HS.Resgister on ULSM. Our prototype requested the data to the audit trail between 23rd and 31st July 2019. The data obtained was analysed in order to find any kind of erros on dates and calculations, inconsistencies and misclassifications. For each dataset, it was removed duplicated records and it was analysed the impact of N/A existence. For the dataset of use case UC4: Check time of activity we produced a summary table by professional category with the metrics: (1) total of results, (2) number of professional with identification code, (3) minimum time of activity, (4) 1st and 3rd Quartiles and median values of time of activity, (5) mean of time of activity, (6) maximum time of activity; (7) standard deviation and (8)number of classification as suspicious activity. This table was illustrated with a boxplot graph showing the time of activities by professional category. For the dataset of use case UC5: Check days of activity we produced a summary table by professional category with the metrics: (1) total of results, (2) minimum days of activity, (3) 1st and 3rd and median days of activity, (4) mean of days of activity, (5) Maximum of days of activity; (6) standard deviation and (7)Number of classification as suspicious activity. This table was illustrated with a boxplot graph showing the consecutive days of activity on HIS by professional category. For dataset of use case UC10: EHR read access we produced a summary table by date with the metrics: (1) total of results, (2) total of results without professional identification, (3) total of results with null patient identification, (4) total of suspicious access classifications, (5) total of suspicious access classifications without professional identification and (6) total of suspicious access classifications with null patient identification. The table was illustrated with a line graph showing the comparison between total of access results and the ones classified as suspicious, by date. We used a boxplot graph to show the comparison by dispersion measures. We showed the results in a table comparing the accesses by professional category. We showed the relation with a bar graph and also showed the comparison between total results and results classified as suspicious by professional category with a boxplot graph.

2.5 Study Authorizations

This study on the field required prior authorization from the institutions where the intervention took place. An application for study authorization was made to ULSM (see annex 7.3) and CHUP (see annex 7.5). The scope of the study auctorization applied to ULSM included all phases of the project and the scope of the study auctorization applied to CHUP was the interviews to professionals on the institution that access to HIS. Both ULSM and CHUP authorizations were conceded (see annexes 7.4 and 7.6).



Results

3. Results

3.1 Introduction

The complexity of the health data workflow is such, that turns very hard to track accesses and acknowledge the illegitimate ones when they happen. Audit trails can be very useful to detect undue accesses but in most cases they are used as a forensic investigation tool when a complaint occurs (Li and Oprea, 2016), as seen previously.

In order to identify patterns and establish rules that may help the identification of unauthorized access of professionals working in healthcare institutions, we firstly conducted some discussions with experts on HIS, an ISD and a DPO from a hospital, and a jurist expert on health data access laws.

A survey was also applied to professionals, from Healthcare facilities, who have access to HIS, that aimed to acquire a better knowledge about the professionals' routines and practices of HIS access.

3.2 Information gathering

Discussion with experts

From the information provided by ISD, DPO and the jurist we could identify some major concerns: (1) failed attempts to login IS ; (2) the correspondence between the professional category and the access profile; (3) type of information that professionals access; (4) the legitimacy of an access by a professional (the purpose of the access); (5) remote access; (6) access by inactive professionals.

Interviews Results

Professional category of interviewees

28 professionals of six different categories from the three healthcare settings (primary, secondary and tertiary care) that have access to HIS were interviewed: 6 physicians (21%), 12 nurses (43%), 1 diagnostic and therapeutic technician (DTT) (4%), 2 nutritionists (7%), 1 psychologist (4%), 1 pharmaceuticals (4%), 1 Social Assistant Technician (SAT) (4%) and 4 administrative (14%).

Table 3.1: Interview results

	Professional Category								
	Nr (%)	Physician	Nurse	DTT	Nutritionist	SAT	Psychologist	Pharmaceutic	Administrative
	28(100%)	6(21%)	12(43%)	1(4%)	2(7%)	1(4%)	1(4%)	1(4%)	4(14%)
Shifts Duration									
<i>Normal duration in hours</i>									
Primary care		7/8	7/8	7/8	7/8	7/8	7/8	7/8	7/8
Secondary/tertiary care		7/8	7/8	7/8	7/8	7/8	7/8	7/8	7/8
<i>Shortest duration in hours</i>									
Primary care		6	–	–	–	–	6	–	–
Secondary/tertiary care		6	6	–	–	–	–	–	–
<i>Longest duration in hours</i>									
Primary care		12	12	–	–	–	12	–	12
Secondary/tertiary care		12	12	12	–	–	–	–	12
<i>Normal period of consecutive days</i>									
Primary care		5	5	5	5	5	5	5	5
Secondary/tertiary care		5	5	5	5	5	5	5	5
<i>Exceptional period of consecutive days</i>									
Primary care		6	7	–	–	–	–	–	8
Secondary/tertiary care		6	8	6	–	–	–	6	8
Information Systems Accessed									
BSimple	4(14%)	2	2						
Clinidata	4(14%)	1	1	1					1
HP-HCIS	3(11%)	1	1	1					
J.One	14(50%)	3	11						
PCE	8(29%)	6	1					1	
PEM	6(21%)	6							
SClinico	24(86%)	6	12		2	1	1		2
SiiMA	4(14%)	1	1	1					1
SINUS	5(18%)		1		2	1			1
SONHO	8(29%)	2			2	1			3
Others	8(29%)	1	2	1		1		1	2

Work routines (work hours and days)

All interviewees confirmed the schedules usually vary between 7 and 8 hours daily, 35 and 40 hours weekly, and 5 consecutive days per week. Normally they make exception shifts lasting 12 hours a day, which may reduce the normal shifts down to 6 hours per day that week, and the emergency head physician can work shifts up to 24 hours a day.

Characterization of access to HIS

Physicians, nurses, Diagnostic and Therapeutic Technician (DTT), nutritionists, psychologists, pharmaceutics and social assistant technicians have access to many IS with health data, although they have different level of patient data access permissions. Administrative usually do not have access to patient health data. Accordingly to the interviewees, the most relevant HIS daily used by these professionals are described in table 3.1.

As we can see on table 3.1, almost every professionals, 24 in 28 (86%), access to Sclinico, which is connected to SONHO and SINUS. In *Others* category we have grouped the following systems: Gestcare CCI, Obscare, RNU, SGICM, Synapse FujiFilm and AIDA.

According to all interviewees and ISD, access to HIS that are managed internally by healthcare institution itself, is done by username and password credentials, unique for each user. Username includes the staff identification number or professional association number and the password is defined by each user.

23 professionals interviewed share computers during the shifts as well as they do not always access from the same computer. Usually just the heads of department have their own computer to work. All professionals can access from any computer on the hospital using their credentials.

Remote access

About remote access, just one professional confirm to access remotely by VPN. According to ISD, VPN access is granted by a special authorization by the institution Administration Council, and it is necessary for tasks accomplishment, such as management activities, hospital encounter coding by coder physicians, and others. Nonetheless remote accesses are always vulnerabilities because they are open doors to get into the HIS.

Issues on information access

Nurses and DTT stated that they have constraints related to patient data access that makes their job harder. This problem is sometimes solved by asking to head of department or to patient's physician the necessary information. When that is not possible, they, simply do not access to information needed or, sometimes, another professional with more permissions provides his or her credential so that the work can be done without restrictions.

Reasons for access with credentials of other users

Professionals were asked about motivations for access HIS with another professional credential. The answers obtained were ordered by the occurrence frequency (see Table 3.2).

Table 3.2: Motivations for access HIS with another professionals' credentials. Each interviewee could order up to 4 reasons based on the frequency.

Motivation	Order of frequency for access motivation				Not answered
	1st	2nd	3rd	4th	
Professional need to access information that own credentials do not allow	24	0	0	4	–
Give patient information to relatives, friends or acquaintances	0	15	3	1	9
Support other professionals	2	4	1	0	21
Curiosity	0	1	2	3	22
Other	–	–	–	–	28

Table 3.3 presents the answers obtained when professionals were asked about the knowledge that they had related to HIS access practices by other professionals in general. It also includes other situations of HIS accesses that could be suspicious of being undue accesses that were not in the predefined list.

Table 3.3: HIS access practices by professionals in general

Situation	Frequency
Predefined list	28
Sharing of credentials	10
Use of credentials that belonged to a professional that is not on service on the healthcare institution	14
Patient's data access of a patient without encounter context	18
To help other professional on his or her tasks	18
Other added by the interviewee	3
There are professionals that share credential to get help from others professionals or to help other professionals on their tasks. Medical students are an example of this practice, or the administrative that uses physicians' credentials to help them with some tasks	1
Professionals that use information available in systems to confront other professionals about their personal life	1
Patients that have been blackmailed because someone unduly access to his or her EHR	1

3.3 Functional Analysis

Based on the gathered information we could describe eleven scenarios that can represent undue or, at least, suspicious accesses. From the described scenarios it was possible to model a use case for each situation and the respective sequence diagram showing the interactions between systems and the data flow. Then we matched the use cases with the specific requirements from Ministers Council Resolution nr 41/2018 which provides guidelines for technology to be compliant with GDPR (Conselho de Ministros, 2018).

3.3.1 Scenarios, Use cases and Sequence diagrams

Scenario 1 - Failed access attempts: *A user tries to access a HIS and is not succeeded. The user can try a second chance and succeed or try over again without success.*

According the professionals interviewed and the ISD, credentials composed by username and passwords are so frequently used that it is not easy to forget them. Consecutive failed login may indicate that something is wrong with the access and must be tracked (Conselho de Ministros, 2018) (see figure 3.1).

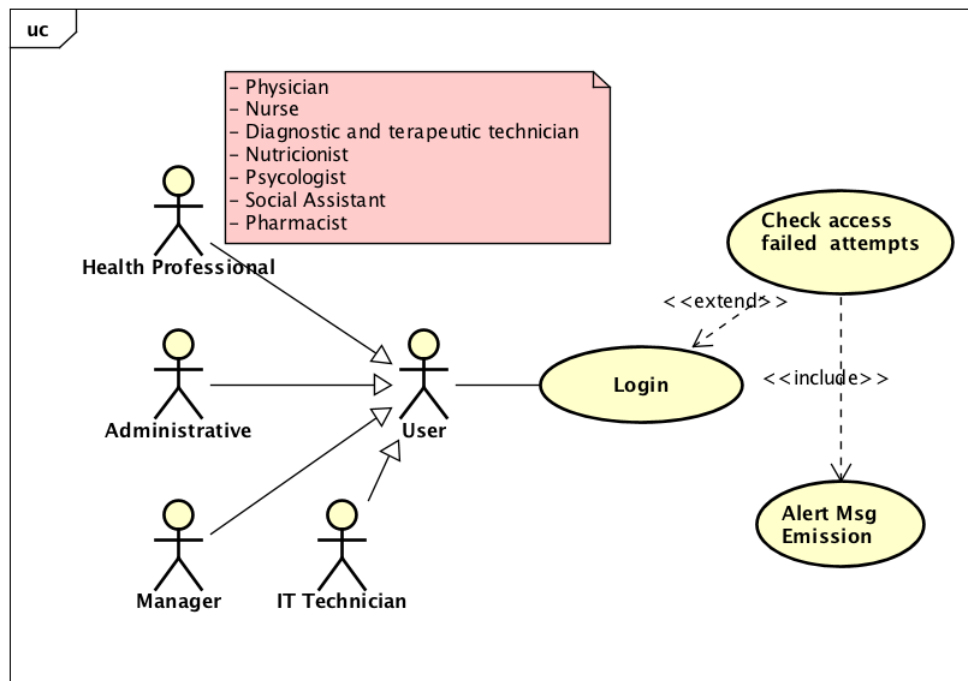


Figure 3.1: Use case 1 - Check failed access attempts

Sequence diagram 1 describes the sequence of actions and the data flow for UC1 (see figure 3.2).

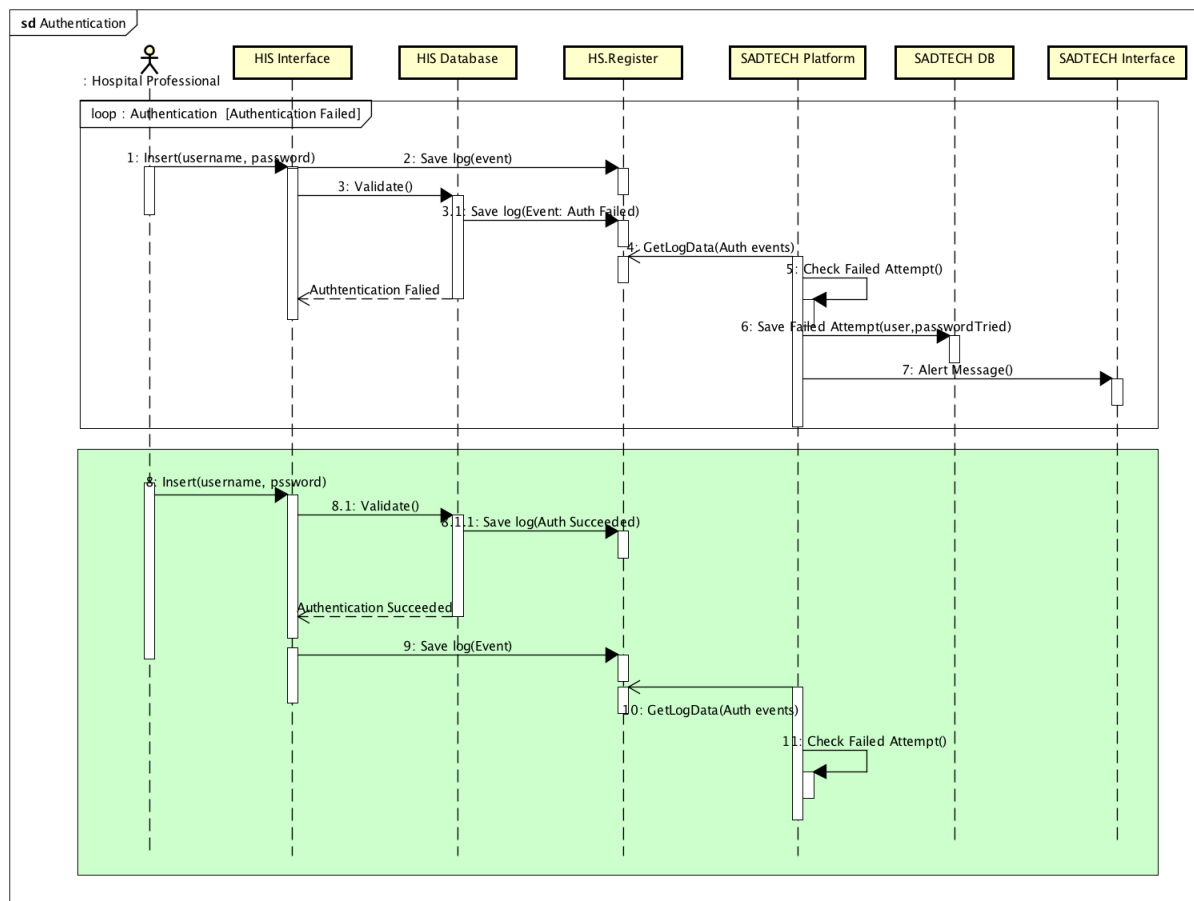


Figure 3.2: Sequence diagram 1 - Check failed access attempts

In this sequence diagram, we have both failed and succeeded workflows. The white area represents the behaviour of the systems when the user fails the authentication. The green area represents the behaviour of the systems when the authentication is well succeeded.

Scenario 2 - Wrong profile user: *IT department set a less restricted profile to a user because he needs to access more information than usual to accomplish his tasks for that month. The administrative logs into HIS with a physician profile. The user profile does not match to the user's professional category.*

In this case, a user may need information that usually can't access to accomplish tasks. This need may legitimate the access (AR, 2016) and ISD confirmed that it can be done assigning user another profile or customizing profile access permissions. However, when the job is completed if the access permissions are not restored, the access becomes illegitimate since then (see figure 3.3). The permissions or profile must be restored.

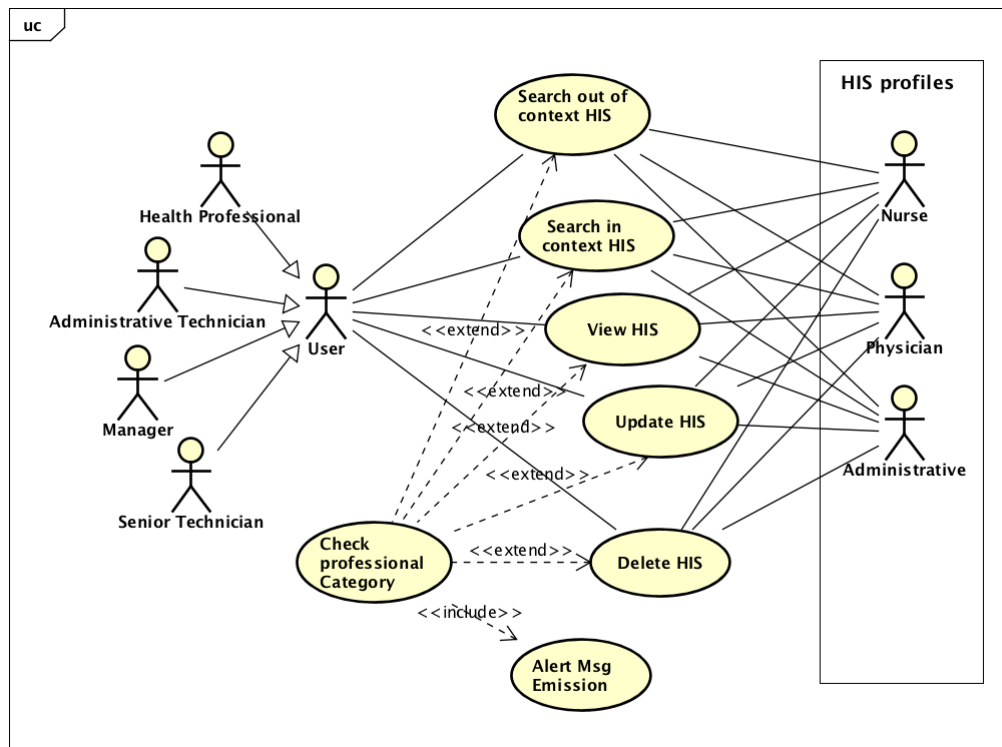


Figure 3.3: Use case 2 - Check wrong profile user

Sequence diagram 2 describes the sequence of actions and the data flow for UC2 (see figure 3.6).

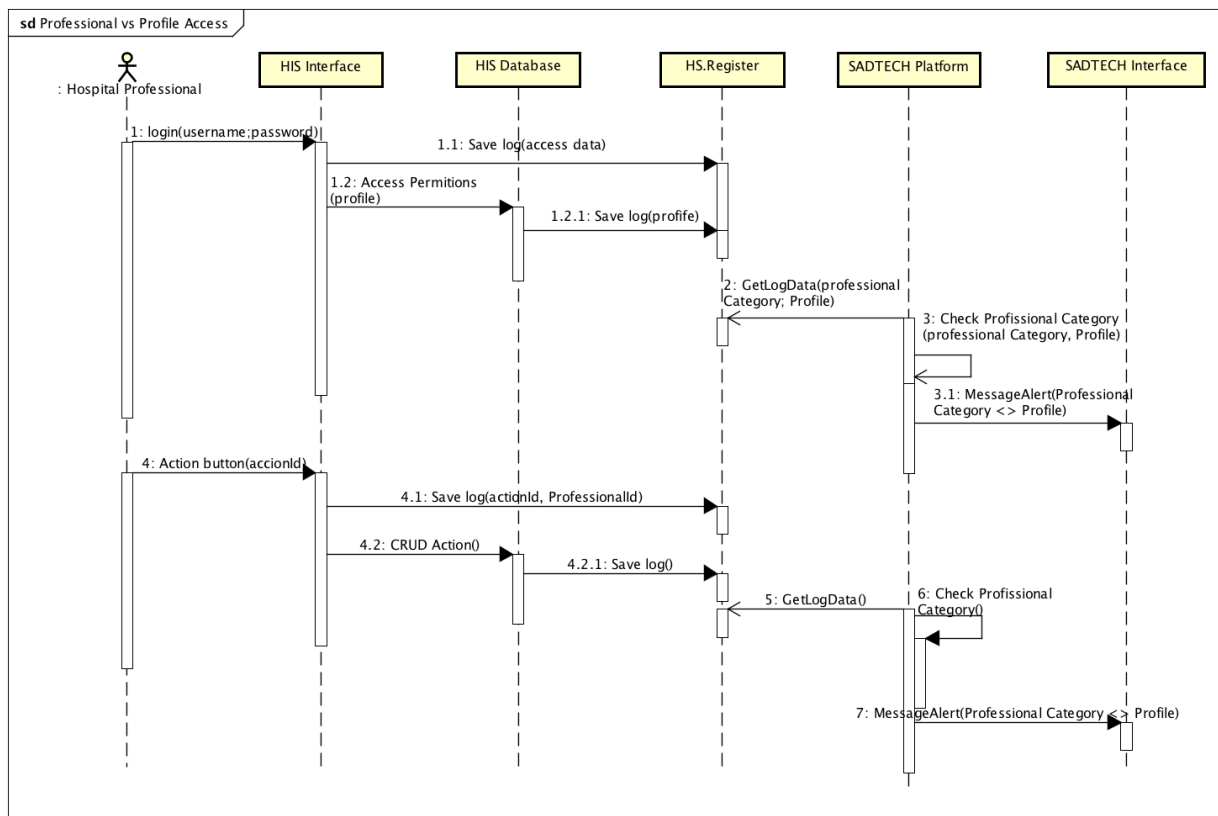


Figure 3.4: Sequence diagram 2 - Check wrong profile user

Scenario 3 - Professional off service access: *A user logs into HIS inside the healthcare institution and access to patients EHR but his presence is not registered on the attendance system.*

A user must access HIS when is in service performing his or her duties and, thus, their HIS activity must match the time period of the user presence registry. If they do not match, this may indicate that the user is accessing the systems outside the professional activity, and in this case access is undue (see figure 3.5). Although it is possible that some exceptions occur, this should not happen and must be monitored (Conselho de Ministros, 2018).

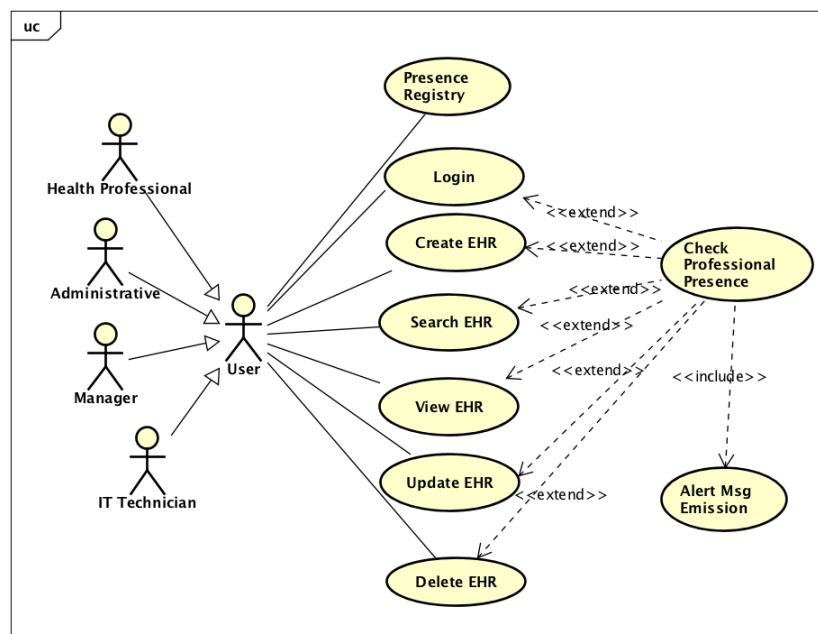


Figure 3.5: Use case 3 - Check professional off service access

Sequence diagram 3 describes the sequence of actions and the data flow for UC3 (see figure 3.6).

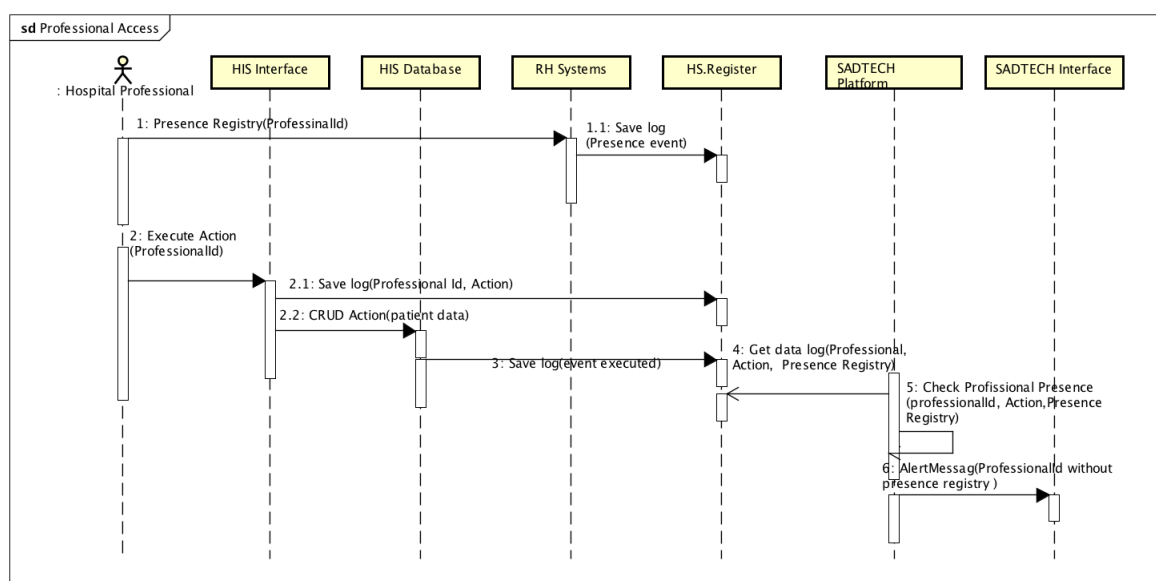


Figure 3.6: Sequence diagram 3 - Check professional off service access

Scenario 4 - Time of activity: *A professional uses his credentials during his shift to accomplish his tasks. In the end of his shift another user uses his credential to access a HIS, to take a look at a patient EHR.*

For this scenario we propose counting the period of time of users' actions on HIS, analysing all consecutive period of credentials activity, in terms of daily hours. If the added time it is longer or shorter than the shift duration should be, it gives information of suspicious users' activity (see figure 3.7). This scenario has the advantage of not depending on the presence's system. If audit trail does not have the necessary information from presence system or if this system is down, it is still possible to detect such accesses this way.

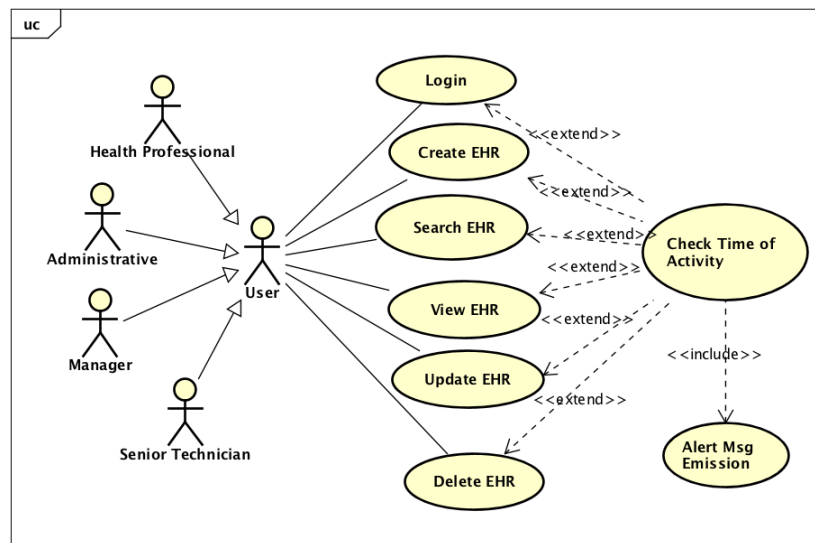


Figure 3.7: Use case 4 - Check time of activity

Sequence diagram 4 describes the sequence of actions and the data flow for UC4 (see figure 3.8).

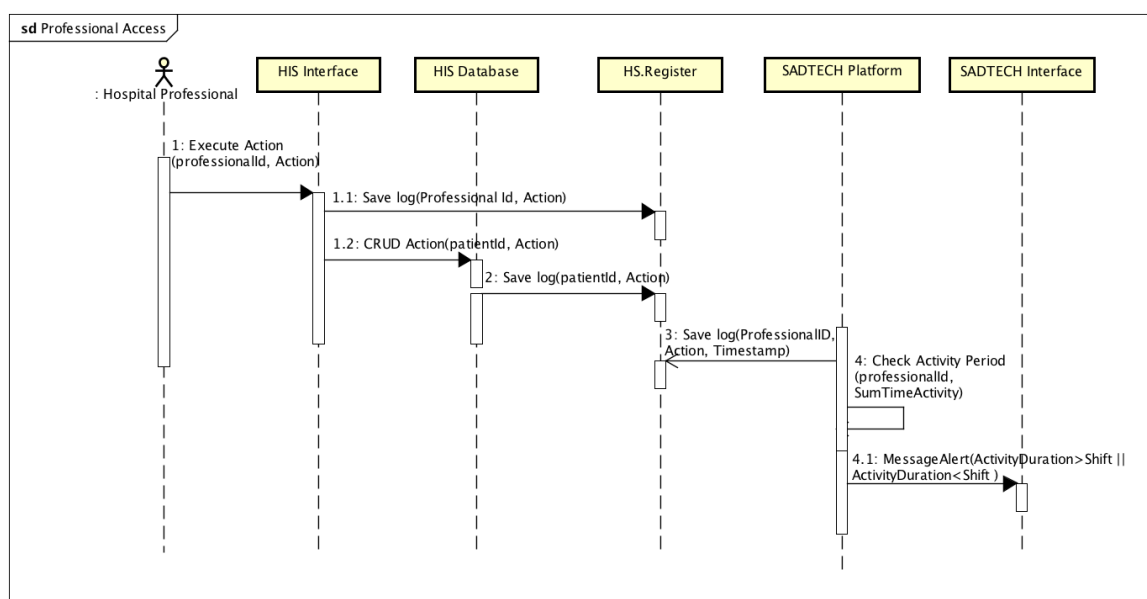


Figure 3.8: Sequence diagram 4 - Check time of activity

Scenario 5 - Days of Activity: *A professional uses his credentials during his shift to accomplish his tasks. In the end of work week, when he is off, another user uses his credential to access a HIS, to take a look at a patient EHR.*

For this scenario we propose counting the consecutive days of activity of users' actions on HIS, analysing the number of consecutive days of credentials activity. If the added number of days is longer then should be, it gives information of suspicious users' activity (see figure 3.9). This scenario has the advantage of not depending on the presence system, too. Again, if audit trail does not have the necessary information from presence system or if this system is down, it is still possible to detect such accesses this way.

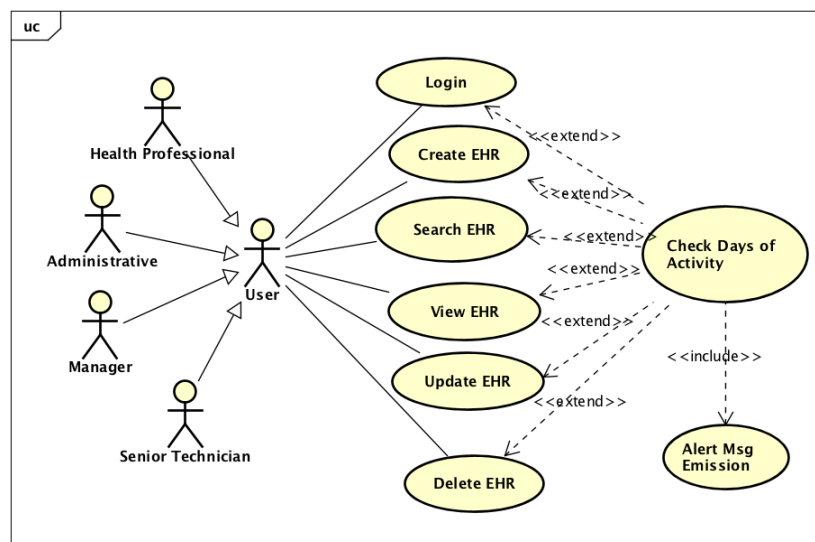


Figure 3.9: Use case 5 - Check days of activity

Sequence diagram 5 describes the sequence of actions and the data flow for UC5 (see figure 3.10).

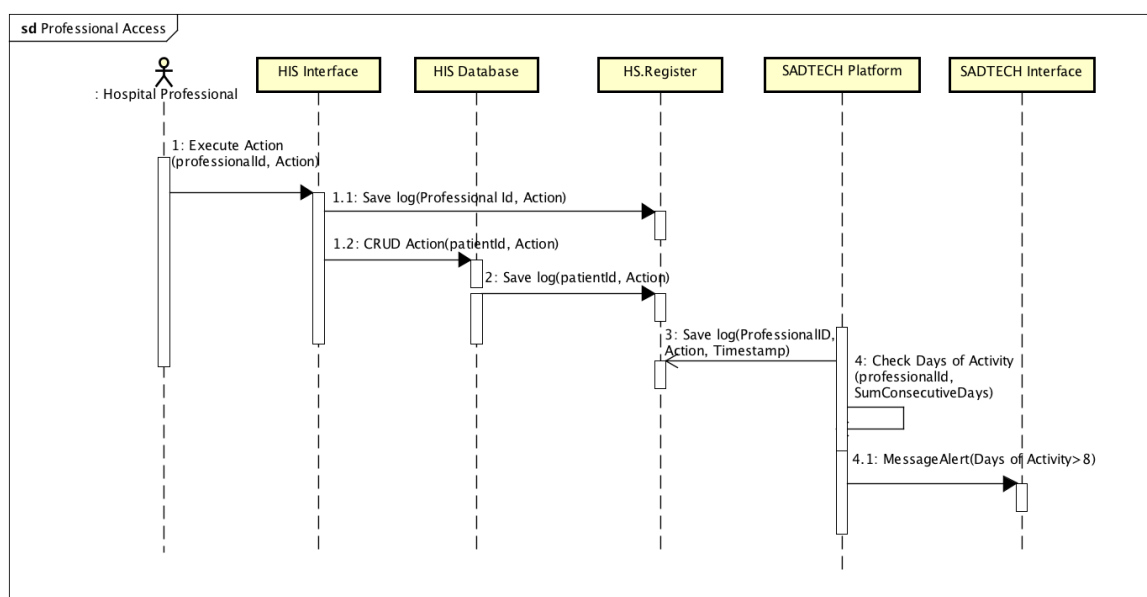


Figure 3.10: Sequence diagram 5 - Check days of activity

Scenario 6 - Access of an inactive professional: *A user access to HIS with someone's credential, but the credential's owner is no longer working on the healthcare institution. Despite the credential has constant and recent activity, there is no registry on the presence system neither on the payroll system of that professional (the owner of the credential used).*

Active credentials of users that are no longer healthcare institution employees, is a critical issue reported by ISD, and which motivated the first fee applied to a Portuguese Hospital for GDPR noncompliance (Séneca, 2018). It was also reported by 3 professionals interviewed that credentials of professionals that are no longer employees at the institution are still in use by other users. For this type of usage, we propose to detect actions of a user that does not have any registry on presence or payroll systems during 3 months period (see figure 3.11) (Conselho de Ministros, 2018).

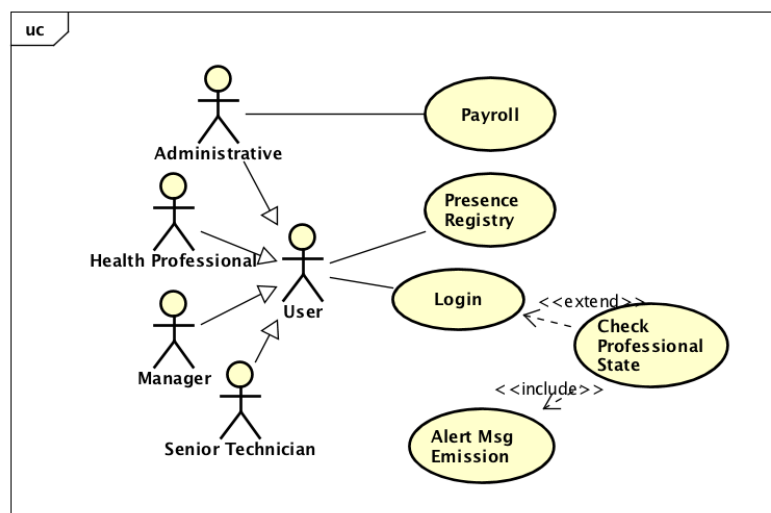


Figure 3.11: Use case 6 - Check inactive professional

Sequence diagram 6 describes the sequence of actions and the data flow for UC6 (see figure 3.12).

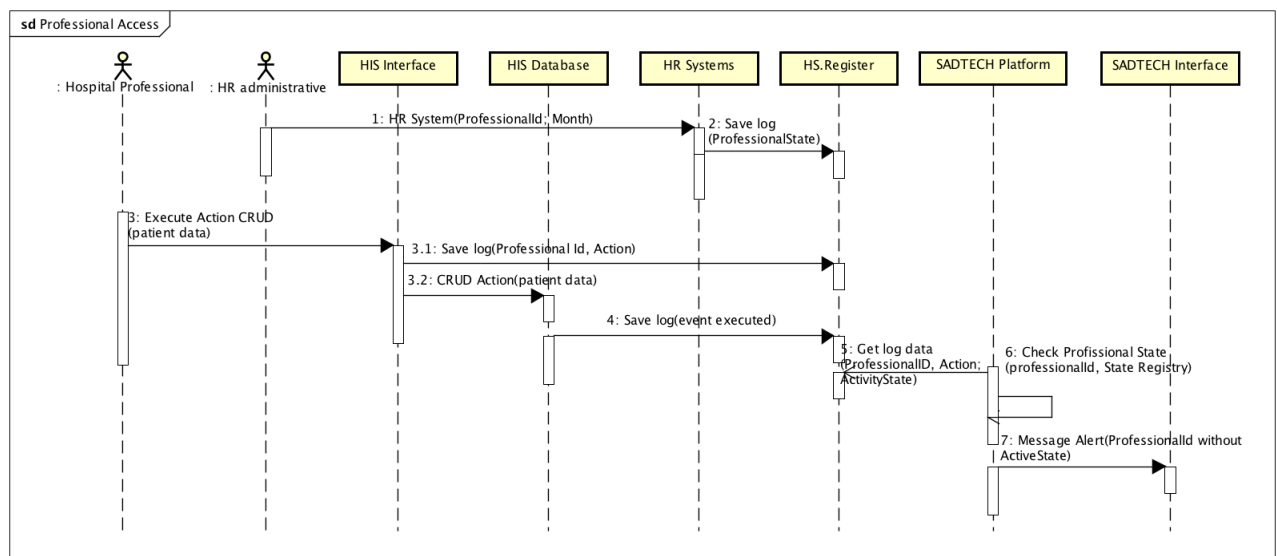


Figure 3.12: Sequence diagram 6 - Check inactive professional

Scenario 7 - Simultaneous access: *A user accesses to a HIS and another user accesses to HIS at the same time with same credentials from another computer. The computer may be from the same department or not.*

This scenario shows the possibility of two different people using the same credential, which results in bad conduct (see figure 3.13). However, according to all interviewees it is possible for same professional to access a system from one computer and for a valid reason access again from another computer while the other session is still active, accounting that the physical proximity of computers used allowed it.

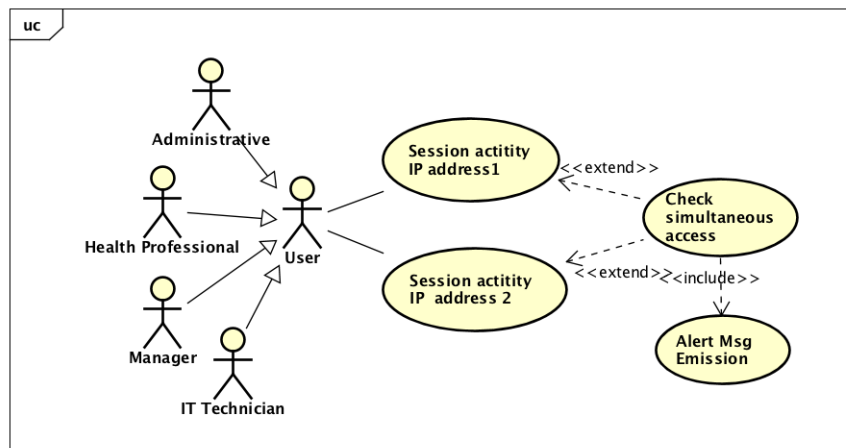


Figure 3.13: Use case 7 - Check simultaneous access

Sequence diagram 7 describes the sequence of actions and the data flow for UC7 (see figure 3.12).

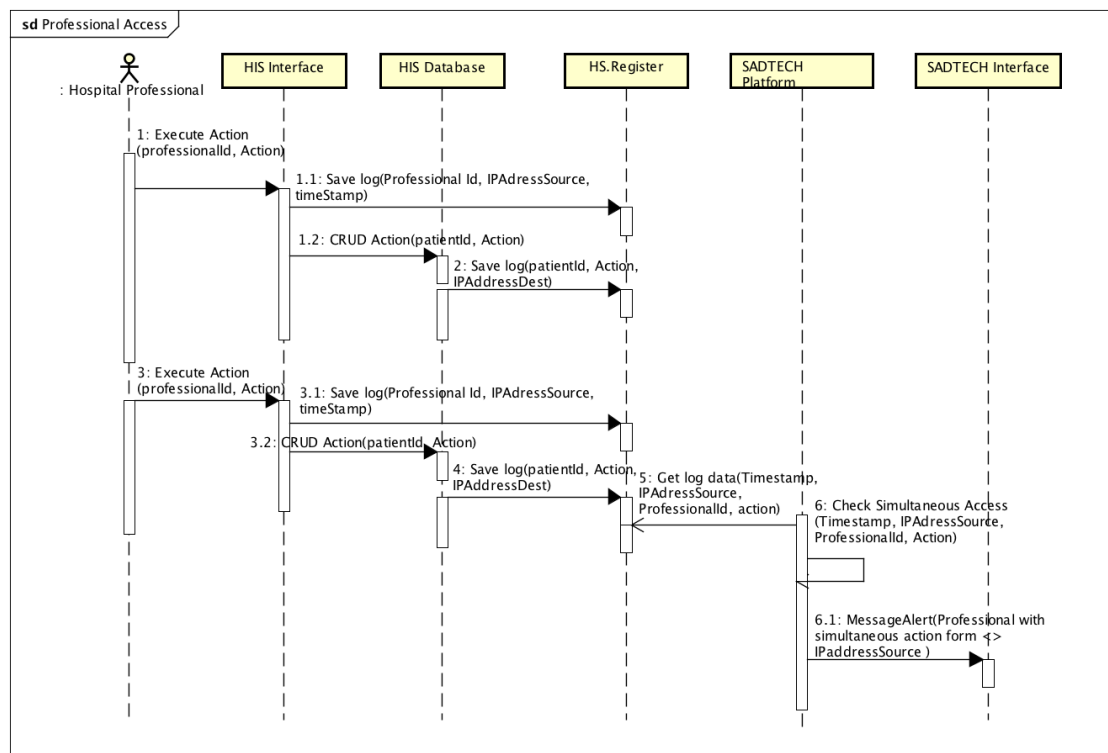


Figure 3.14: Sequence diagram 7 - Check simultaneous access

Scenario 8 - Access to EHR of healthcare institution employee: *A patient has an appointment in the same hospital where she works. The physician does not fill the EHR fields because he wants to be sure that no one sees this patient's information. The physician is aware that it happened in the past and can bring harm to the patient, who is also employee in that hospital.*

This scenario shows that there are critical issues concerning patient data privacy, confidentiality and research data quality related to patients which are simultaneously employees on same healthcare facilities. Thus, we propose to check the accesses to employees EHR and produce a report to send the affected person with the information of who accessed to his or her health data (see figure 3.15).

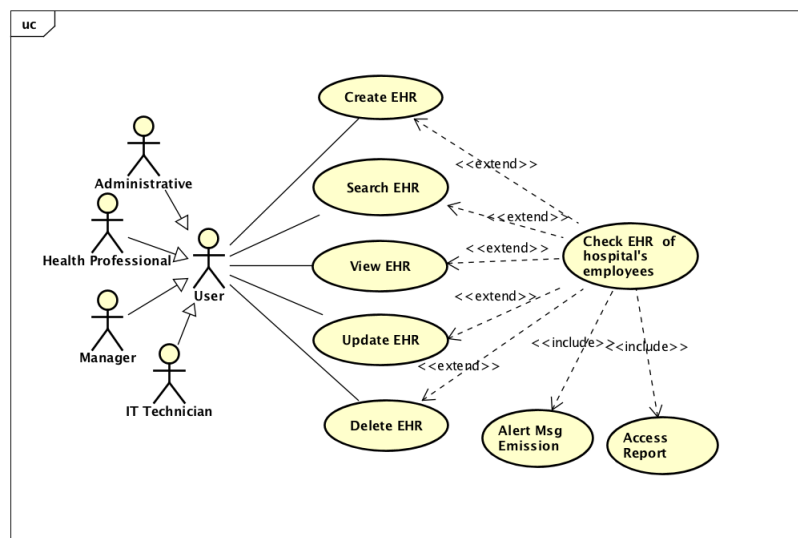


Figure 3.15: Use case 8 - Check access to EHR of healthcare institution employee

Sequence diagram 8 describes the sequence of actions and the data flow for UC8 (see figure 3.16).

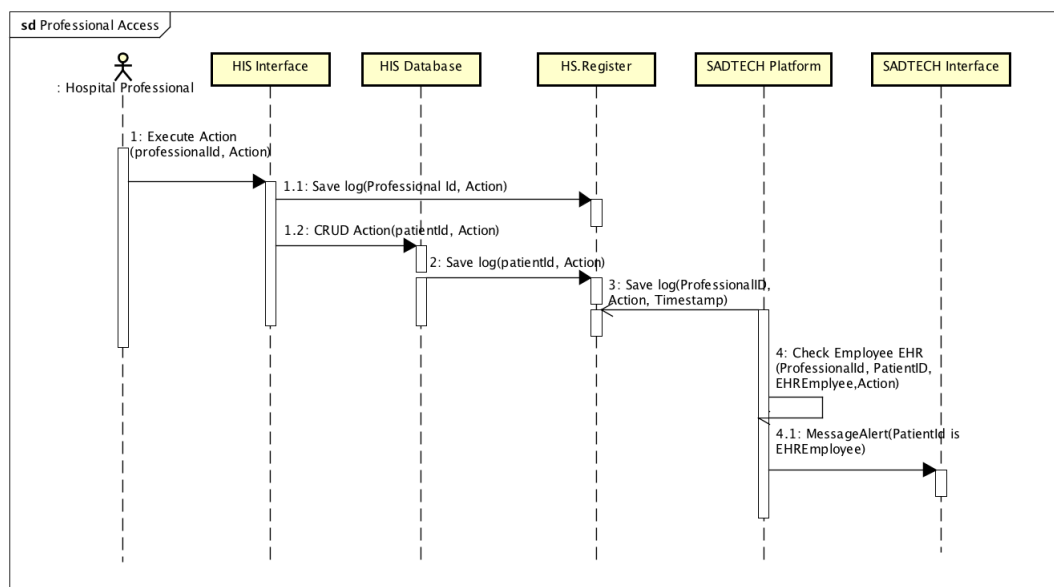


Figure 3.16: Sequence diagram 8 - Check access to EHR of healthcare institution employee

Scenario 9 - Remote access: *An employee of healthcare institution, that has a VPN access authorization, accesses late at night and sees some EHR during a short period of time.*

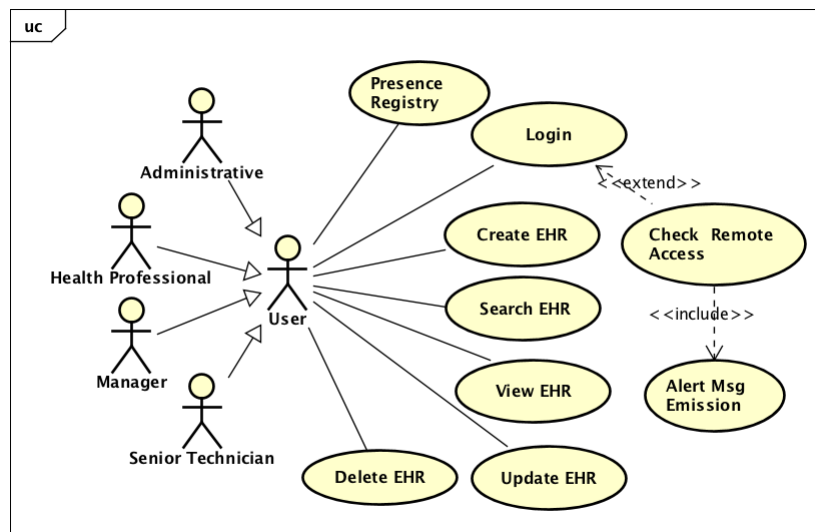


Figure 3.17: Use case 9 - Check remote access

This scenario describes potential risks of VPN accesses. According the ISD they are granted through a special authorization, and they are necessary for task accomplishment, such as management tasks, encounter coding by coder physician and others, as seen on interviews, but in fact, they are potential vulnerabilities. It is an open door to get into the HIS and must be monitored. This will allow to identify the VPN accesses on use and the ones that are active without usage, and also to track the activity with this type of access (see figure 3.17).

Sequence diagram 9 describes the sequence of actions and the data flow for UC9 (see figure 3.18).

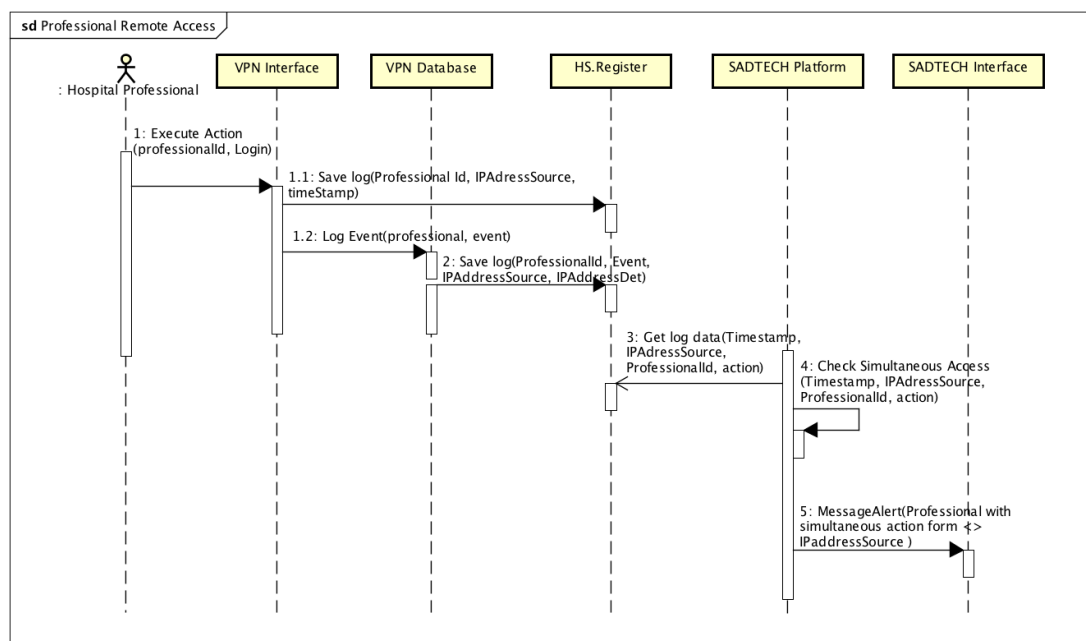


Figure 3.18: Sequence diagram 9 - Check remote access

Scenario 10 - EHR read access: *A professional access to a patient EHR. Why does he access? What are the evidences of the healthcare delivering of that professional.*

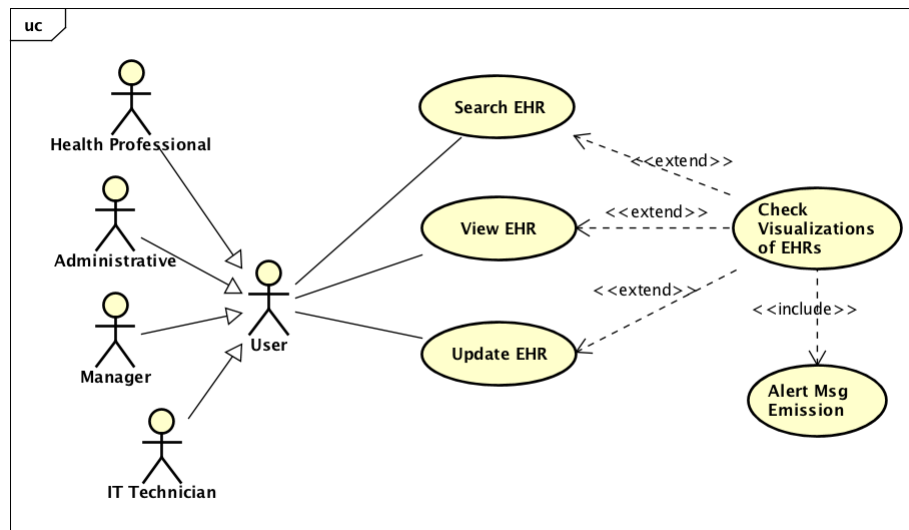


Figure 3.19: Use case 10 - Check EHR read access

The visualization of an EHR without healthcare delivery purpose is a great concern for ISD and DPO. The lack of evidences that can justify such access is already spotted as an issue to solve (see figure 3.19). Management of Information (MOI) 11.5 of JCI for hospitals certification points out this type of access as a requirement that should be addressed to mitigate problems related to data breaches (Joint Commission International, 2017).

Sequence diagram 10 describes the sequence of actions and the data flow for UC10 (see figure 3.20).

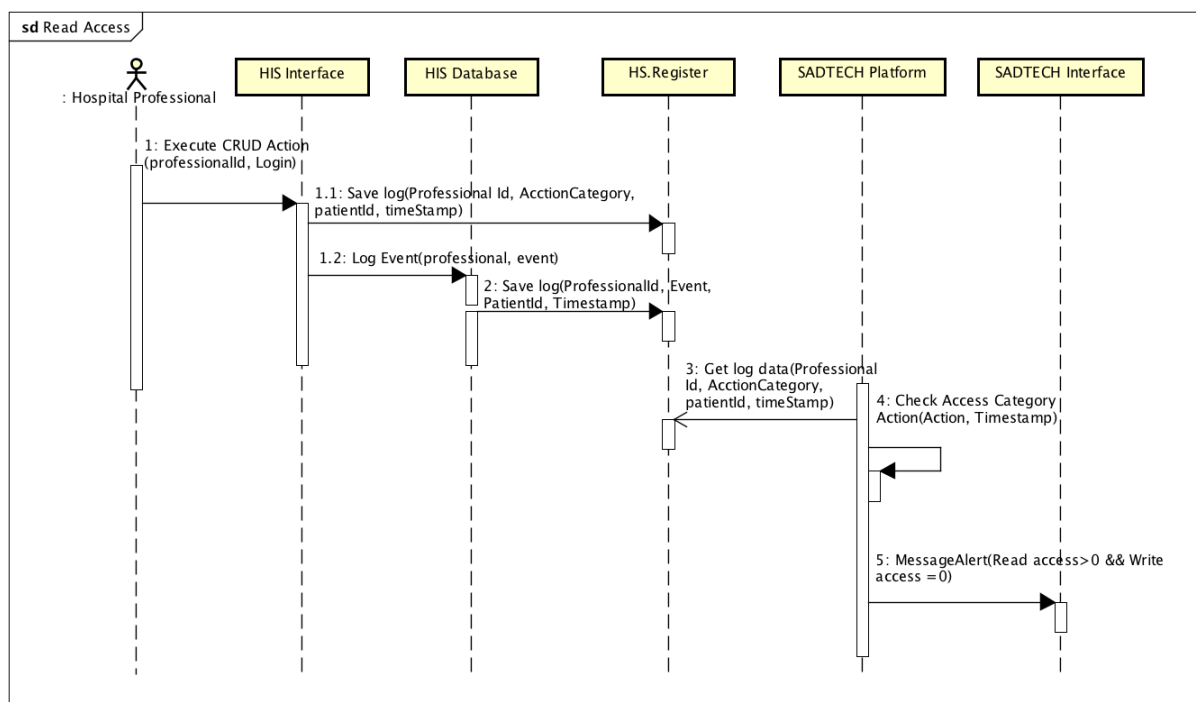


Figure 3.20: Sequence diagram 10 - Check EHR read access

Scenario 11 - Access to EHR analysis based on a set of variables: *A professional access to a patient EHR. Why does he access? What is his role? Is he responsible for that patient? Is he giving a second opinion? Has he been following the patient? Did he get Complementary Diagnosis Test and Therapeutic (CDTT) results? Is he prescribing? Does the patient have an appointment? Has he been admitted at the hospital? What was the motivation for the access to occur?*

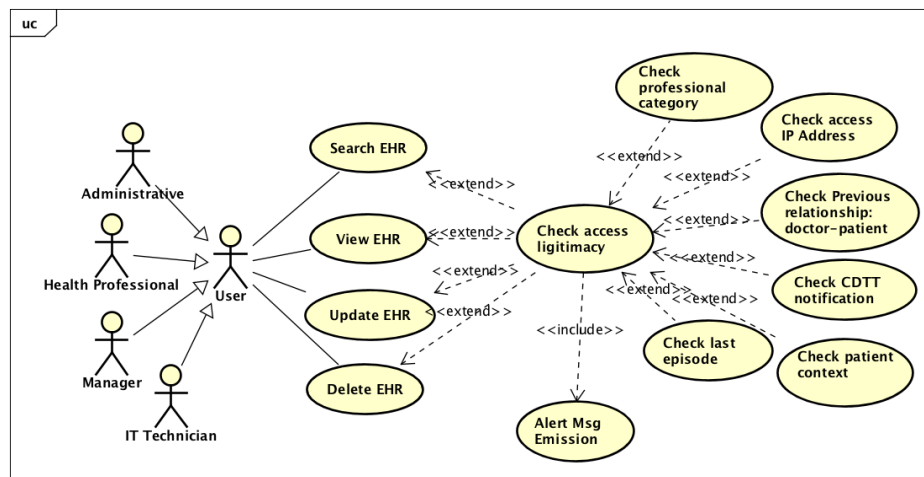


Figure 3.21: Use case 11 - Check access to EHR analysis based on a set of variables

Going deeper on the last use case presented, this one describes what should be analysed when an action occurs on the EHR. How can we know that is not an undue access? This is another concern for ISD and DPO, to find out if an access apparently normal is in fact legitimate or not. In this case we do not focus just on actions of EHR reading but we want to explore all possible actions made by users. We propose to perform a context analysis, using different variables' information that can influence decisions about the access legitimacy and present the probability of being undue (see figure 3.21).

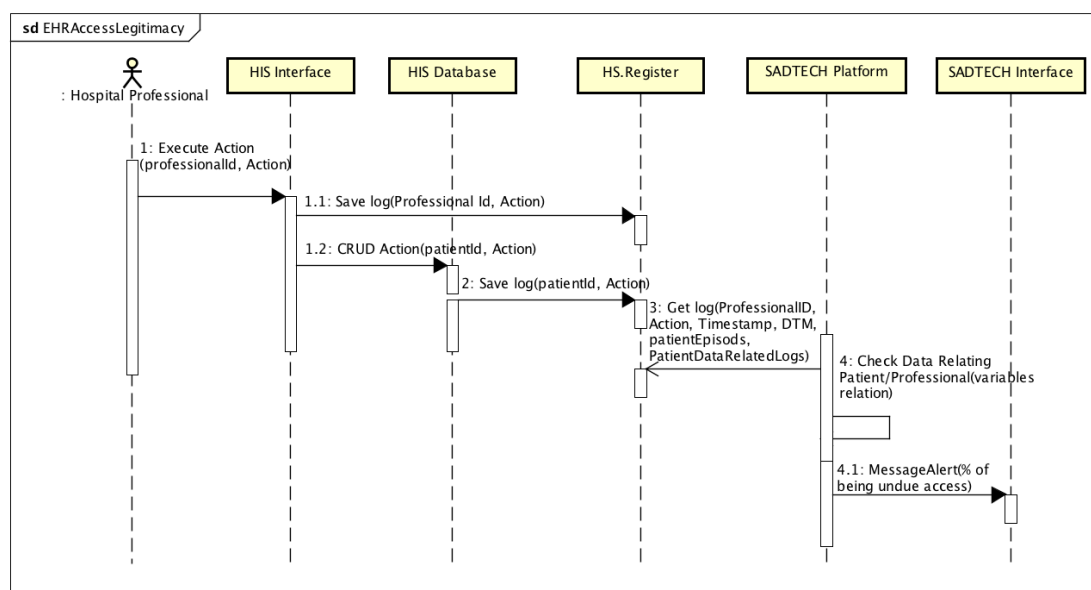


Figure 3.22: Sequence diagram 11 - Check access to EHR analysis based on a set of variables

3.3.2 Logs Variables

Based on the use cases produced we identified 34 relevant variables needed by the algorithms of undue access suspicions, of which 14 are required to comply with the MRC nr. 41/2018. It is necessary logs from three different type of sources: (1) system logs, (2) firewall logs, (3) application logs (from HIS, presence registry system and payroll system). Table 3.4 shows that just two use cases need information from the Human Resources Systems to be implemented: UC3 - Check professional off service and UC6 - Check inactive professional. All the others need information from applications that must produce mandatory data on their logs if they are compliant with legal requirements. The last column calculates the relevance of each variable to all eleven use cases and the last row shows the number of variables that are necessary to implement each use case with the wanted detail.

Table 3.4: Relation between log sources and variables, and the UCs.

Log Source	Variables	UC1	UC2	UC3	UC4	UC5	UC6	UC7	UC8	UC9	UC10	UC11	Variable Usage
Systems logs (Logon Session Events)	security id	x		x	x	x	x	x					54.5%
	account name*	x		x	x	x	x	x					54.5%
	timestamp*	x		x	x		x	x					45.5%
	logon type code*	x		x	x	x	x	x					54.5%
	source IP*	x						x					18.2%
Applicational logs	professional id	x	x	x	x	x	x	x	x		x	x	90.9%
	professional category		x		x	x			x		x	x	54.5%
	account name*	x	x	x	x	x	x	x	x		x	x	90.9%
	account profile		x						x		x	x	36.4%
	timestamp*	x	x	x	x	x	x	x	x		x	x	90.9%
	patient id		x						x		x	x	36.4%
	context		x						x		x	x	36.4%
	action type			x			x		x		x	x	45.5%
	action id*	x	x	x	x	x	x	x	x		x	x	90.9%
	source IP*	x	x					x	x		x	x	54.5%
	last CDTT notification date								x		x	x	27.7%
	last patient episode date								x		x	x	27.7%
	previous patient-physician interaction								x		x	x	27.7%
Presence system registry	professional id			x			x						18.8%
	timestamp entry registry			x			x						18.8%
	timestamp exit registry			x			x						18.8%
Payroll system	professional id						x						9.1%
	month processed						x						9.1%
Firewall logs	timestamp*	x		x			x	x		x			45.5%
	account name*	x		x			x	x		x			45.5%
	action type*	x		x			x	x		x			45.5%
	protocol						x	x		x			27.7%
	source IP*	x		x			x	x		x			45.5%
	destination IP*									x			9.1%
	source port*									x			9.1%
	destination port									x			9.1%
	path									x			9.1%
Total of variables per UC		14	9	16	9	9	19	15	13	9	13	13	

* Mandatory variables required on the Ministers Council Resolution nº 41/2018

3.4 Implementation

Considering the wingspan of the work involved for implementation of all use cases and taking into account the scope of the thesis, we decided to select three use cases to implement. In order to choose which use cases were to be implemented, we reflected on the type of logs available to work at that moment, how fast we could obtain results and the practical application of information obtained after development. UC4: Check time of activity, UC5: Check days of activity and UC9: Check EHR read access were chosen. The first reason for this choice was the fact that the data from applicational logs was enough for the development as it was. It did not required preparation of new information or to collect information from other sources that were not being collected yet.

Algorithm proposal for use case UC4: Check time of activity

The use case UC4: Check time of activity intend to detect the activity of professionals on HIS with a period shorter or longer than the expected for a shift.

According to the results of interviews, shortest shift has about 6 hours and the longest shift may have 24 hours. However, this applies only to the emergency head shift physician once a week. So we took 12 hours as a superior limit for longest shift.

We also know that the access to patients' EHR may be done in the beginnig of the shifts and then at the end of the shift, at least. So we considered that an interval less then 6 hours between access of same professional indicates that he or she is in continuous activity.

The continuous activity time is added. If there is an interval superior to 6 hours we considered that the professional was off and the continuous activity time starts counting again.

If the activity time of what we considered a period of work is less then 5 hours or greater then 12 hours, the access is classified as "suspicious access" and an alarmistic is launched, else the access is classified as "normal access".

In the end of the process a CSV report is generated with information of the accesses:

- 1) professional Id number
- 2) professional category
- 3) start period timestamp
- 4) end period timestamp
- 5) number of consecutive hours of activity
- 6) access classification

The algorithm is described on the activity diagram for UC4: Check time of activity (see figure 3.23).

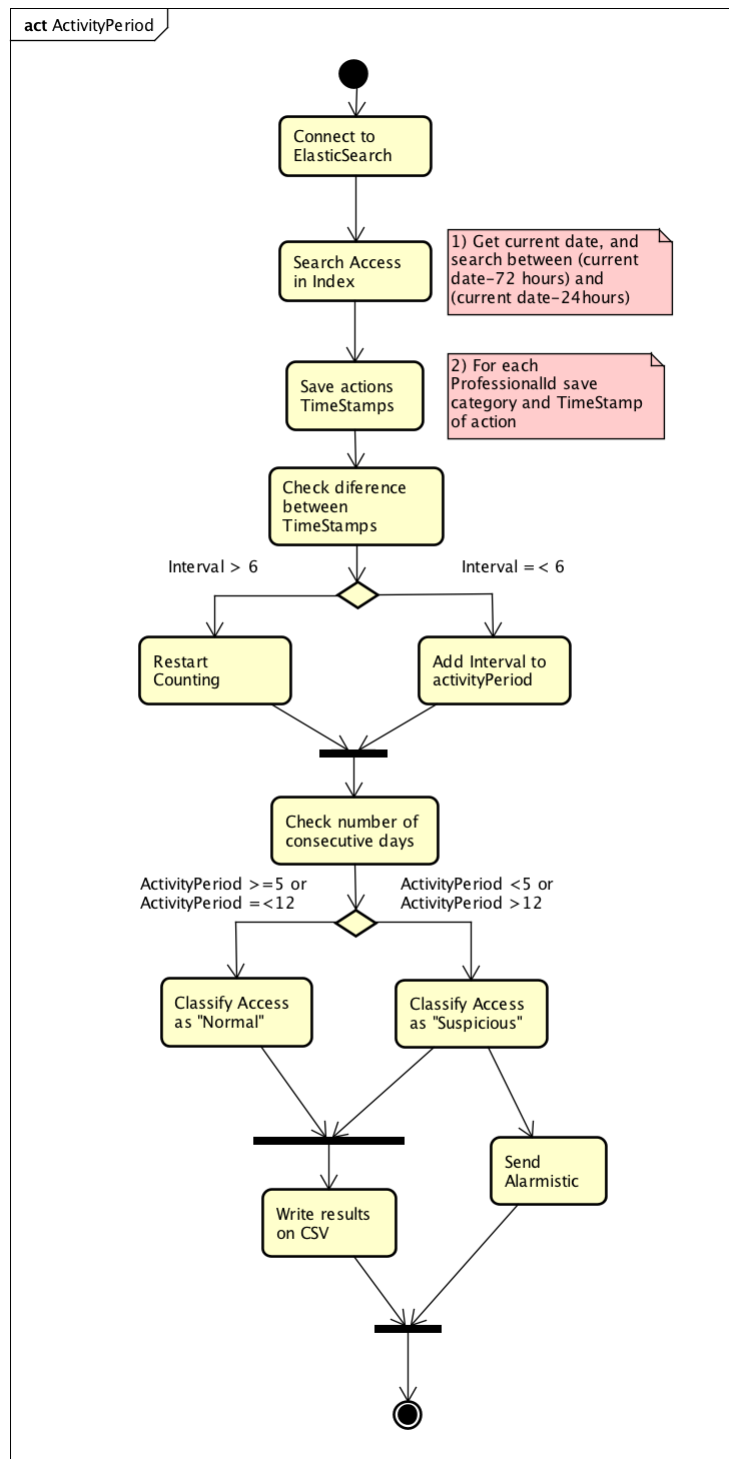


Figure 3.23: Activity diagram for UC4: Check time of activity

Algorithm proposal for use case UC5: Check days of activity

Again, applying the results of interviews, days of work are usually up to 8 consecutive days. The algorithm we propose for use case UC5: Check days of activity, counts the consecutive days of work in a period of 9 days. If there is a gap of at least one day between two consecutive days of accesses, the activity period starts counting. In the end of the period analysis, if the sum of the days with activity of

a professional is greater than 8 days, the access is classified as “suspicious access” and an alarm is launched, else it is classified as “normal access”.

In the end of the process a CSV report is generated with information of the accesses:

- 1) professional Id number
- 2) professional category
- 3) start analysis period date
- 4) end analysis period date
- 5) number of consecutive days worked
- 6) access classification

The algorithm is described on the activity diagram for UC5 - Check days of activity (see figure 3.24).

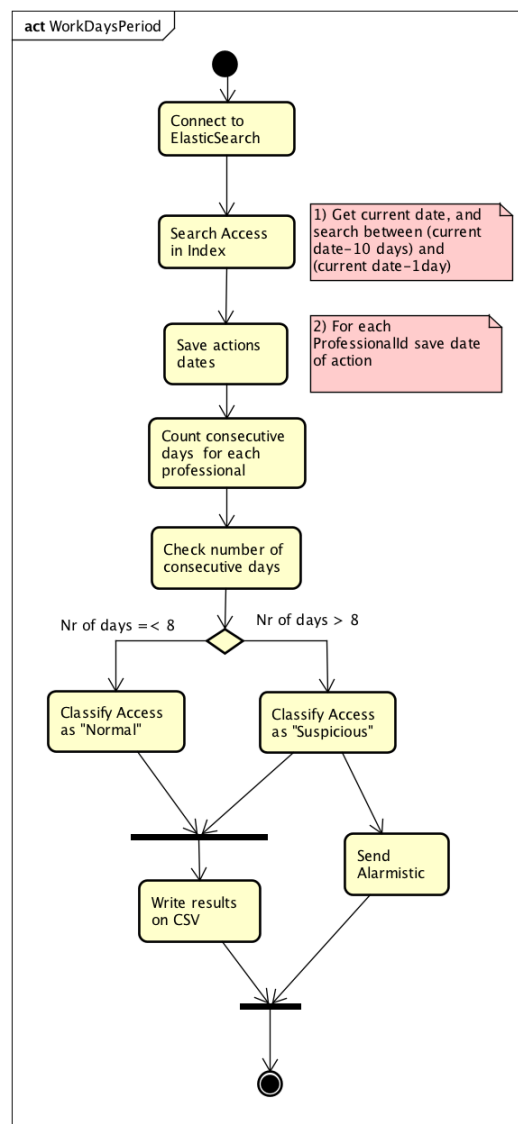


Figure 3.24: Activity diagram for UC5: Check days of activity

Algorithm Proposal for use case UC10: Check EHR read access

The use case UC10: Check EHR read access intend to detect the access made by professionals to visualize the EHR of a patient. According to results of interviews, the longest shift may have 24 hours. We also know that the update of patients' EHR is sometimes done with some delay and the professional may have to take a look at the EHR to confirm if all the necessary registries are done.

So, we decided to analyse on a certain date if there were "search" actions of a professional on a patient EHR, which we consider the action as "read". Then we confirm if there were any "create" or "update" action in the 24 hours before to 24 hours after the analysed date, which we consider the action as "write".

The date of analysis is the date two days prior to the current date, so 48 hours prior the current date. Which means that the analysis to preform is 72 hours before the current date (analysis start date) and 24 hours before the current date (analysis end date). For each professional that access a patient EHR on the date analysis it will be checked the type of access (read or write) made. If on the date of analysis there is "read" action and write "action", the access will be classified as "normal access". On the other hand, if there is a "search" action without a "create" or "update" action, it will analyse if there are any action of this type in the previous 24 hours or in the next 24 hours. If there is at least one action of "create" or "update" in this period, the access is classified as "normal access", else the access is classified as "suspicious access" and an alarmistich is launched.

In the end of the process a CSV report is generated with information of the accesses:

- 1) professional Id number
- 2) professional category
- 3) patient Id number
- 4) date of analysis
- 5) EHR Visualized on date of analysis (y: yes ; n: no)
- 6) EHR Updated on date of analysis (y: yes ; n: no)
- 7) EHR Updated 24 before date of analysis (y: yes ; n: no)
- 8) EHR Updated 24 after date of analysis (y: yes ; n: no)
- 9) access classification

The algorithm is described on the activity diagram for UC10: Check EHR read access (see figure 3.25).

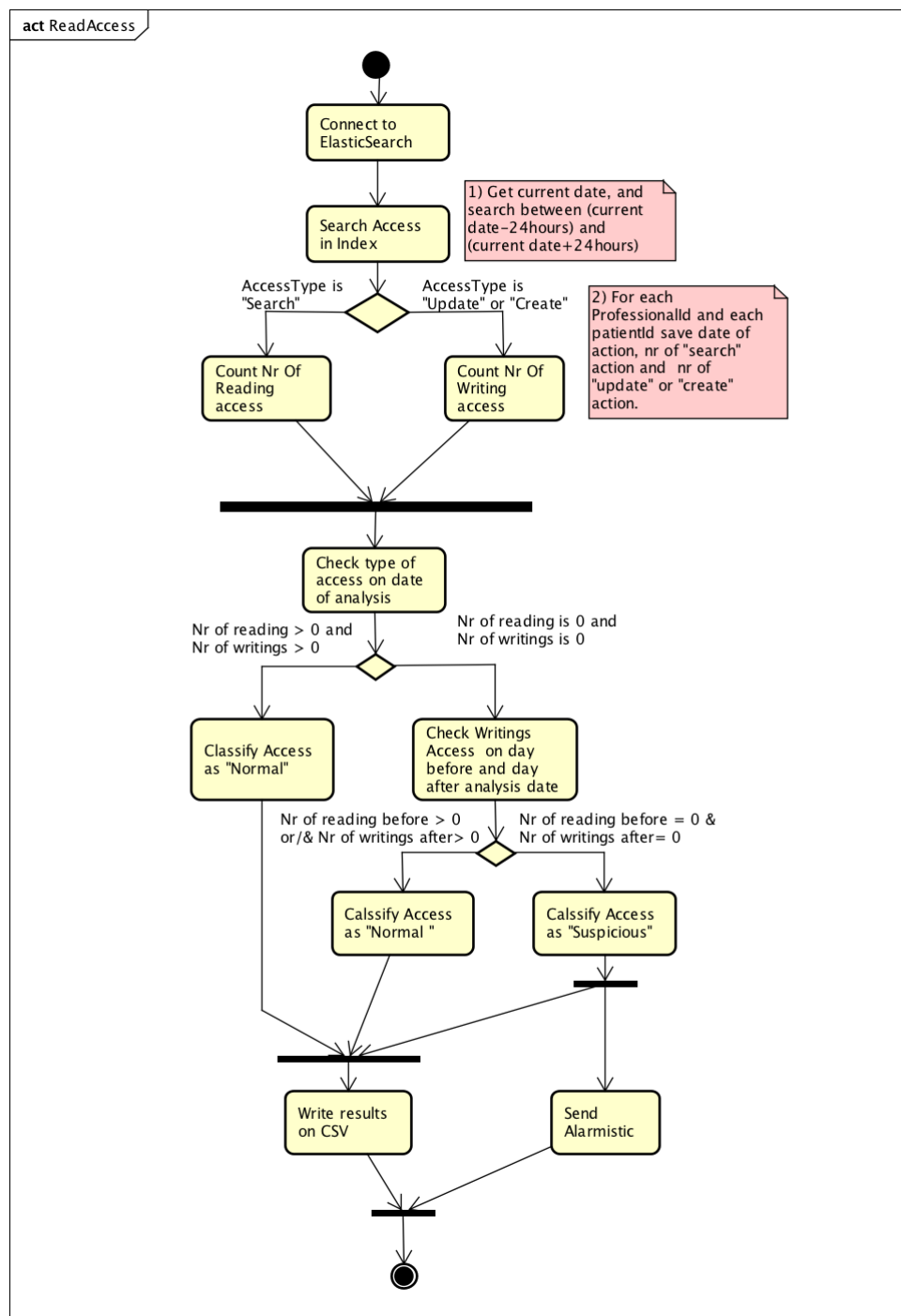


Figure 3.25: Activity diagram for UC10: Check EHR read access

3.5 Evaluation

The algorithms were evaluated during nine days, between the 23rd and the 31st July, using logs of Obscare system from ULSM. Obscare system's logs were the only applicational logs that were being collected on audit trail HS.Register at the moment of tests. This system is being used on hospital stay, consulting and emergency context.

Use case UC4: Check time of activity

The first algorithm add the time of activity for each professional and classifies the access as suspicious if a professional accesses a period of time less than 5 hours (<300 minutes) and greater then 13 hours (>780 minutes), otherwise the system classifies the access as normal. In the referred period, which counts nine days, we got 276 results, after removing the duplicated ones, of which 176 were classified as suspicious.

Table 3.5: Results for UC4: Check time of activity by professional category

Professional Category	Nr results	Professional identified	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Standard Deviation	Suspicious access
No identified	6	0	4305	4305	4305	4308	4305	4328	6.12	6 (100%)
Admin Sirai	5	5	4.00	4.00	6.00	6.40	8.00	10.00	2.60	5 (100%)
Admin VCOBSGYNV3,Create users	1	1	2.00	2.00	2.00	2.00	2.00	2.00	0	1 (100%)
Administrative	56	56	1.00	96.75	338.00	307.54	364.25	713.00	210.83	24 (43%)
Nurse	108	105	0.00	50.25	210.50	335.45	361.75	4320.00	194.41	71 (66%)
Nurse,Admin VCInt,Physician, Development team	1	1	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1 (100%)
Nurse,Administrative,Admin Physician,Admin Sirai, Admin_Backoffice	4	4	0.00	2.25	8.00	13.00	18.75	36.00	16.31	4 (100%)
Management, Admin, VCOBSGYNV3,Creat Users,Physician	1	1	137.00	137.00	137.00	137.00	137.00	137.00	0.00	1 (100%)
Indicators Sirai,Admin Sirai	1	1	0.00	0.00	0.00	0.00	0.00	0.00	0	1 (100%)
Physician	28	28	0.00	74.5	230.00	314.60	399.00	1543.00	345.88	21 (75%)
Physician_Specialist,Physician	65	65	0.00	124.00	338.00	436.40	669.00	1431.00	365.68	41 (63%)
	276	267	0.00	67.00	266.00	422.80	448.50	4320.00	766.37	176 (63%)

The boxplot graph (figure 3.26) gives a better look of the behaviour of professional in terms of time of activity on HIS by category.

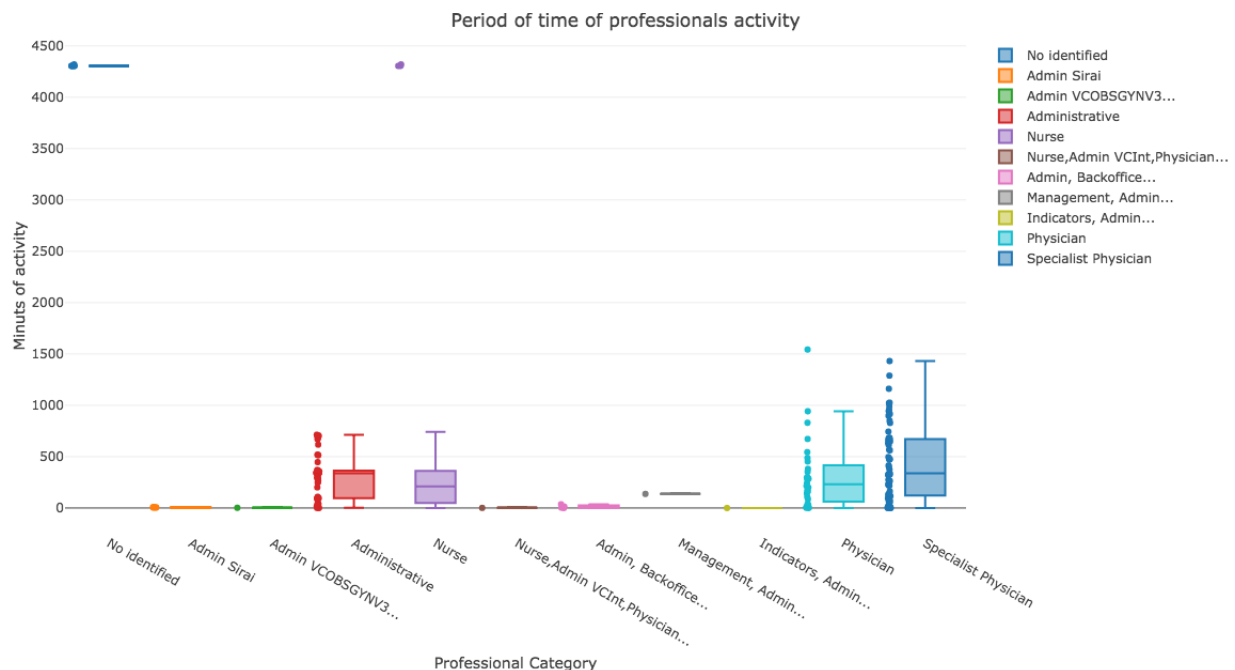


Figure 3.26: Consecutive time of activity in minutes on HIS by professional category

From the data presented we can see that there are two outliers with different behaviours. For the category “No identified” we have all the cases with a duration completely distinct from the others, but all the 6 occurrences have the same behaviour. This happens because this category is related to automatic processes that are running in parallel. On the other hand we have an outlier for “Nurse” category similar to “No identified” category, because there are some automatic processes associated to “Nurse” category. We can also see that categories with management and research competencies have activities with very short duration and few occurrences.

Taking a close look to categories that are more related to healthcare delivery and removing the automatic processes from “Nurse” category we can see that the results presented (figure 3.27) generally do not exceed the superior limit fixed as suspicious access. However the categories “physician” and “specialist physician” have some results that exceed that limit. It is also possible to see that there are many accesses that do not go over the inferior limit being responsible for most of the suspicious access classifications.

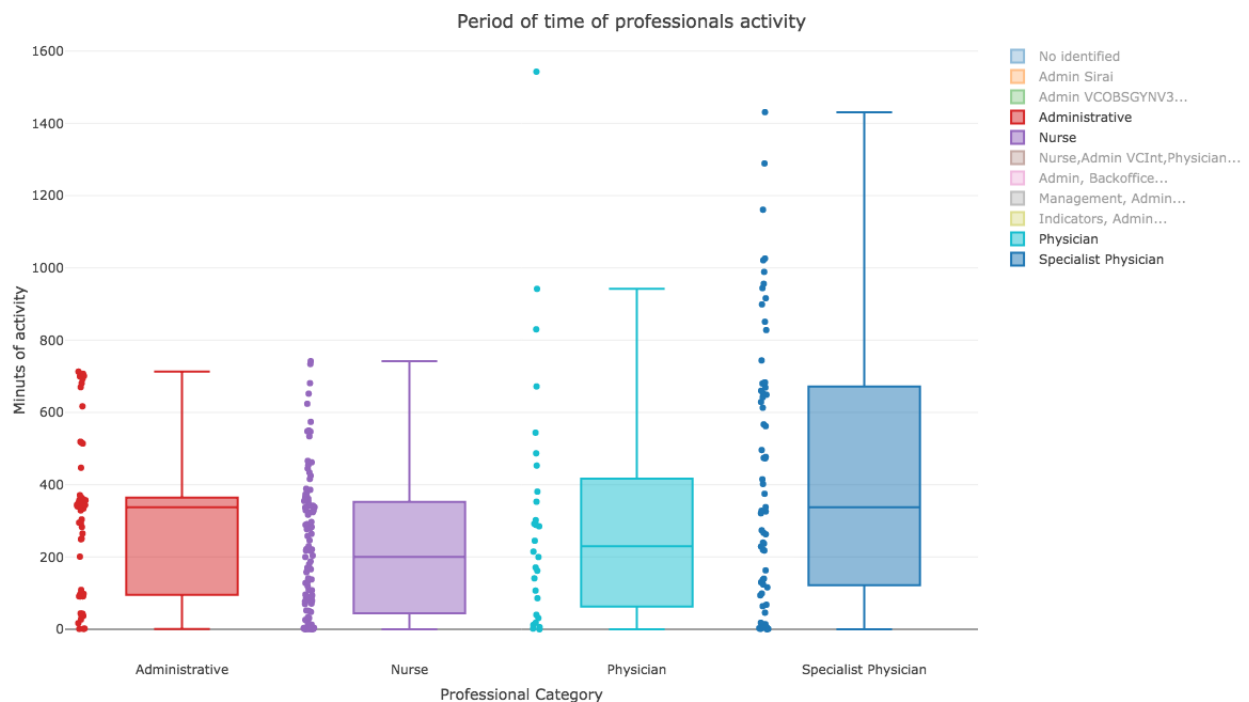


Figure 3.27: Consecutive time of activity in minutes on HIS by professional category related to healthcare delivery

Use case UC5: Days of activity

The second algorithm counts the consecutive days of professionals activity accessing EHR and classifies access as suspicious if a professional accesses a system for more than eight consecutive days. The code was tested over the same nine days period. For this period we obtained 795 results where 777 were classified as “normal access” and 18 were classified as “suspicious access”. Given that there are repeated results due to the accumulated daily results, we removed the duplicated ones. We got 318 results of which 17 were classified as “suspicious access” (figure 3.6).

Table 3.6: Results for UC5: Check days of activity

Professional Category	Nr results	nr of consecutive days						Standard deviation	Suspicious access
		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.		
No identified	9	11	11	11	11	11	11	0.00	9 (100%)
Admin Sirai	18	1.00	1.00	1.00	1.11	1.00	2.00	0.32	0
Admin VCOBSGYNV3,Create users	1	1.00	1.00	1.00	1.11	1.00	1.00	N/A	0
Administrative	61	1.00	1.00	2.00	2.18	3.00	7.00	1.46	0
Nurse	101	1.00	1.00	1.00	1.67	2.00	5.00	0.99	0
Nurse,Admin SIRAI	1	1.00	1.00	1.00	1.11	1.00	1.00	N/A	0
Nurse,Admin VCInt,Physician, Development team	2	1.00	1.00	1.00	1.11	1.00	1.00	0.00	0
Nurse,Administrative,Admin Physician,Admin Sirai, Admin_Backoffice	5	1.00	1.00	1.00	1.60	2.00	3.00	0.89	0
Management, Admin, VCOBSGYNV3,Creat Users,Physician	1	1.00	1.00	1.00	1.11	1.00	1.00	N/A	0
Indicators Sirai,Admin Sirai	7	1.00	1.00	1.00	1.43	1.50	3.00	0.78	0
Physician	30	1.00	1.00	2.00	2.03	2.00	5.00	1.19	0
Physician_Specialist,Physician	82	1.00	1.00	2.00	2.57	2.00	11.00	2.64	8 (9.7%)
	318	1.00	1.00	1.00	1.93	2.00	11.00	2.25	17(5,3%)

The boxplot graph (figure 3.28) gives a better look of the behaviour of the activity of consecutive days of HIS usage by professional category.

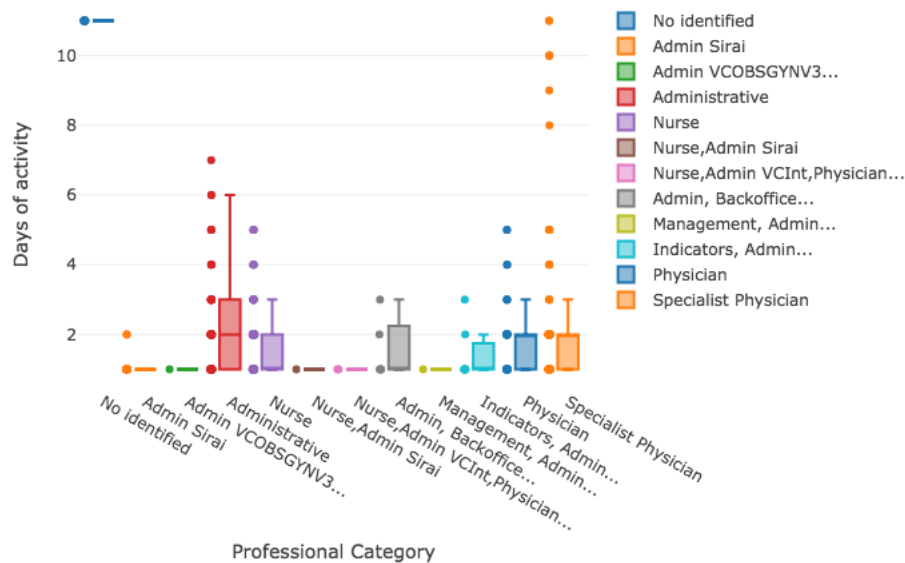


Figure 3.28: Consecutive days of work by professional category

We can also see that automatic processes run every day, and the management and research accesses occurs generally once or twice a week.

Observing the categories that are directly related to delivery of healthcare we can have a better look of what happens in terms of consecutive activity days (figure 3.29). Similarly to what happens in the results of UC4: Check time of activity, in this one the suspicious accesses are associated to "specialist physician" category.

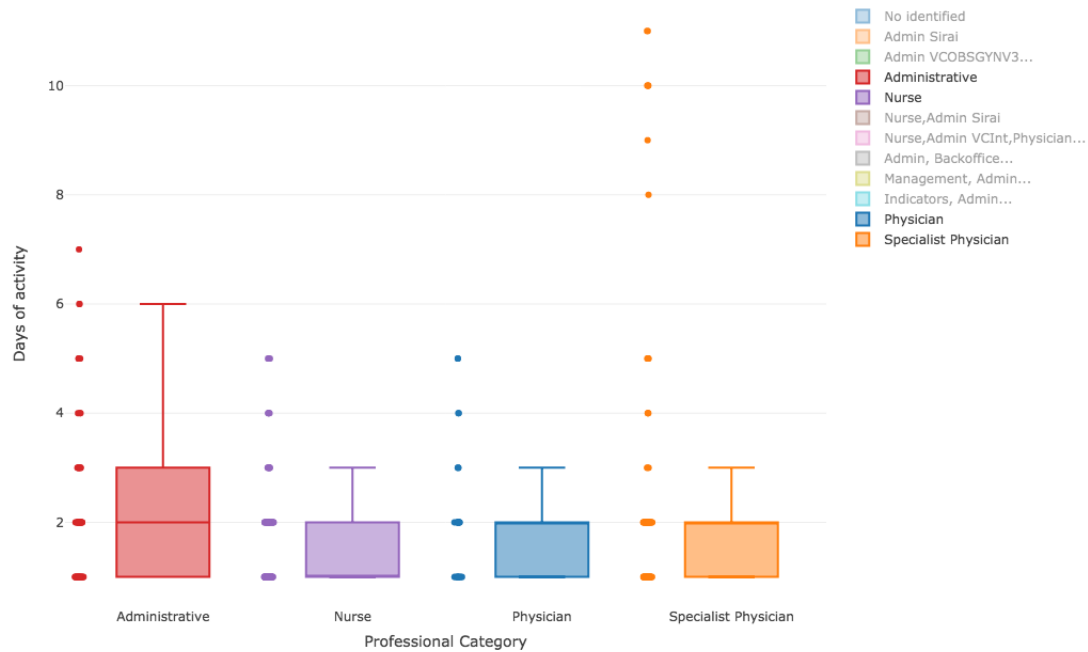


Figure 3.29: Consecutive days of work by healthcare delivery professional category

Use case UC10: Check EHR read access

The third algorithm was to identify accesses made by professionals to view patients' EHR on Obscare system. During the nine days it totalized 378 results, of which 300 (79%) were classified as "suspicious access" by our system. This means that, during this period, there were 300 sets of accesses of a professional to an EHR patient without any registry being made. Analysing the table 3.7 we can see that 32 accesses classified as "normal access" do not have professional ID and 28 others do not have patient ID.

Table 3.7: Results of EHR read access

Analysis date	Total of results	No		Suspicious access classification by system		
		Professional Id	Null Patient Id	Total	Null Professional Id	Null Patient Id
2019-07-22	49	10 (20%)	11 (22%)	39 (80%)	4 (10%)	9 (23%)
2019-07-23	38	6 (16%)	13 (34%)	30 (79%)	4 (13%)	8 (27%)
2019-07-24	49	7 (14%)	13 (27%)	44 (90%)	4 (9%)	12 (27%)
2019-07-25	59	9 (15%)	11 (19%)	43 (73%)	5 (12%)	7 (16%)
2019-07-26	31	11 (35%)	10 (32%)	24 (77%)	6 (25%)	8 (33%)
2019-07-27	31	10 (32%)	7 (23%)	24 (77%)	7 (29%)	4 (17%)
2019-07-28	29	6 (21%)	6 (21%)	22 (76%)	4 (18%)	2 (9%)
2019-07-29	44	10 (23%)	9 (20%)	34 (77%)	4 (12%)	7 (21%)
2019-07-30	48	8 (17%)	14 (29%)	40 (83%)	7 (18%)	9 (22%)
	378	77 (20%)	94 (25%)	300 (79%)	45 (15%)	66 (0,22%)

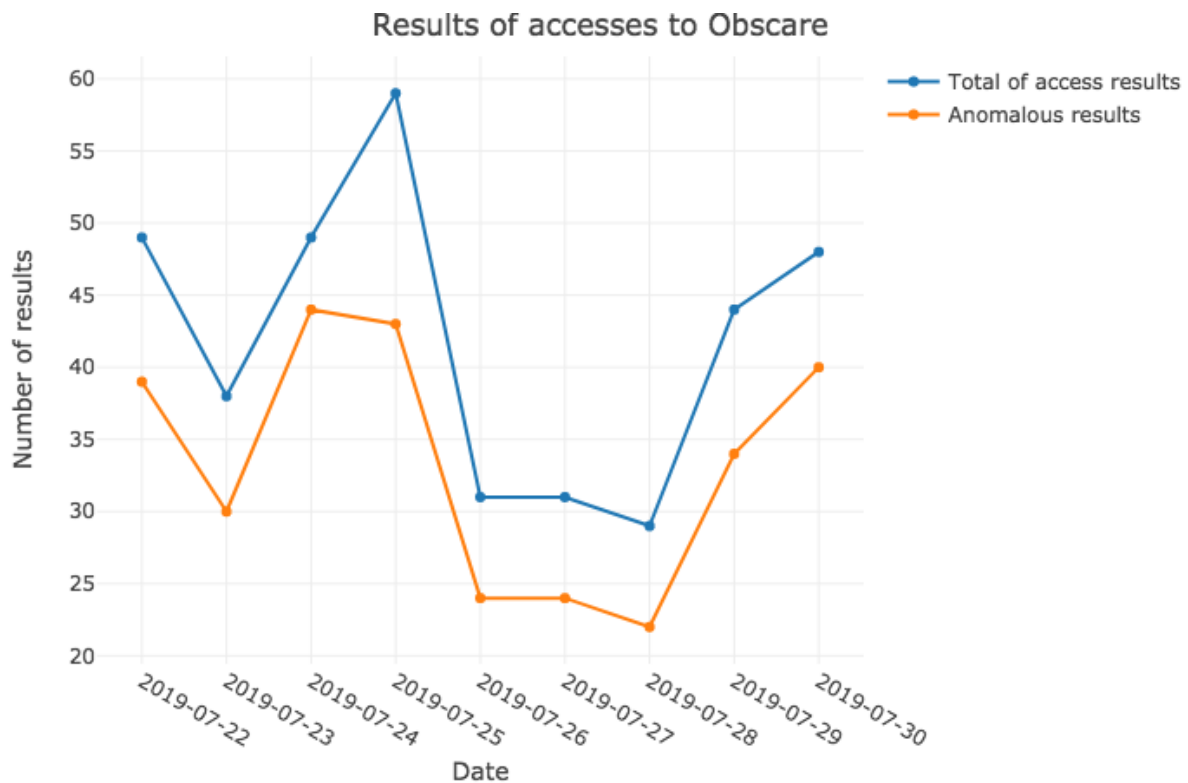


Figure 3.30: Results of EHR read access

The results show that 77 classifications (20%) do not have the professional Id. According the provider of Obscare system, such occurrences are related to automatic processes that run in parallel, to check, get and retrieve necessary data on EHR. The results with N/A patients id are 94 (25%), and they are users' processes that are not related with patients but to other type of reports instead.

So we adjusted the values excluding these results and obtained the values on table 3.8.

Table 3.8: Adjusted results obtained

Date	Total results nr = 216	Suspicious access nr= 198 (92%)
2019-07-22	29	27 (93%)
2019-07-23	20	19 (95%)
2019-07-24	30	29 (97%)
2019-07-25	40	32 (80%)
2019-07-26	11	11 (100%)
2019-07-27	15	14 (93%)
2019-07-28	18	17 (94%)
2019-07-29	26	24 (92%)
2019-07-30	27	25 (93%)

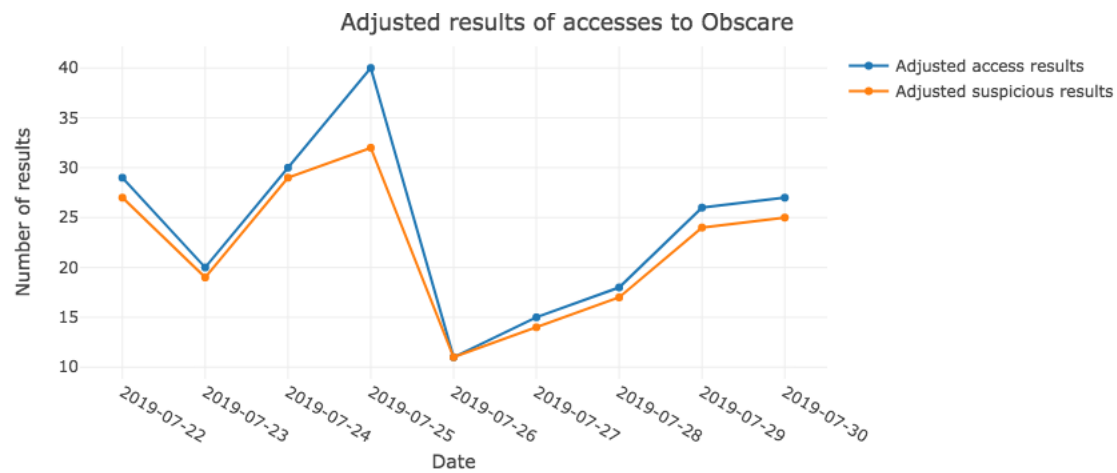


Figure 3.31: Adjusted results of access analysis

We can also see that there are categories that typically access data for management and research tasks. All these accesses were considered suspicious because they are query actions. Nonetheless all the other categories have a high percentage of access classified as “suspicious access”. This indicates that there are a several number of EHR accessed that did not had information updated, and were just consulted.

Table 3.9: Number of results obtained by professional category

Professional Category	Nr Accesses nr=378	Nr Suspicious nr=300 (79,3%)
No identified	50	30 (60)
Admin Sirai	7	7 (100)
Admin VCOBSGYNV3,Creat User	2	2 (100)
Administrative	52	38 (73)
Nurse	117	93 (79,5)
Nurse,Admin VCInt,Physician, Development team	1	1 (100)
Nurse,Administrative,Admin Physician,Admin Sirai, Admin_Backoffice	12	12 (100)
Management, Admin, VCOBSGYNV3,Creat Users,Physician	2	2 (100)
Indicators Sirai,Admin Sirai	2	2 (100)
Physician	25	18 (72)
Physician_Specialist,Physician	108	95 (88)

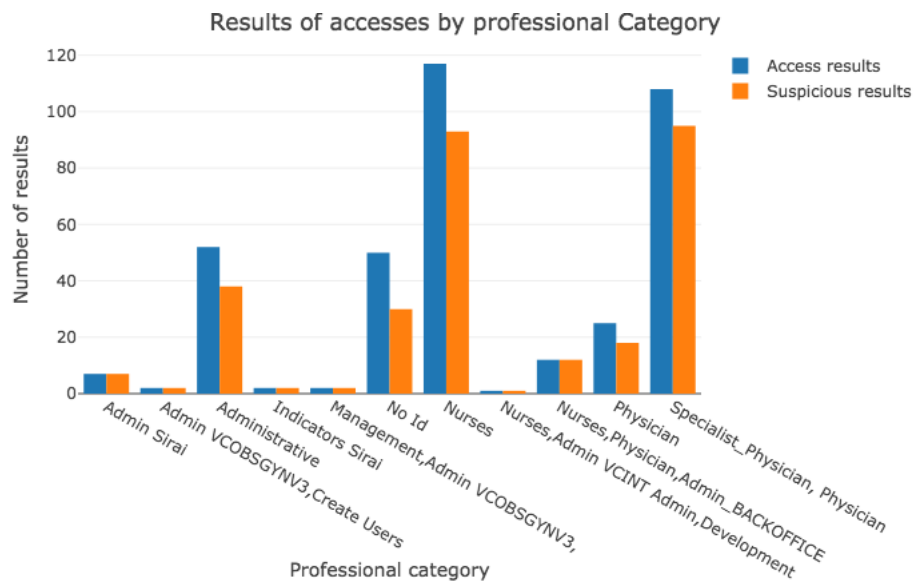


Figure 3.32: Results of access by professional category

Excluding the categories of management and research responsibilities we have four categories left, and all present a high percentage of suspicious accesses with a mean of 78%. We also can see that the “Physician” category has the lowest percentage and the “Specialist physician” has the highest.

Table 3.10: Number of results obtained for professional categories directly involved in care delivery

Category	Total results	Suspicious access
Administrative	52	38 (0.73)
Nurse	117	93 (0.79)
Physician	25	18(0.72)
Physician_Specialist,Physician	108	95 (0,88)

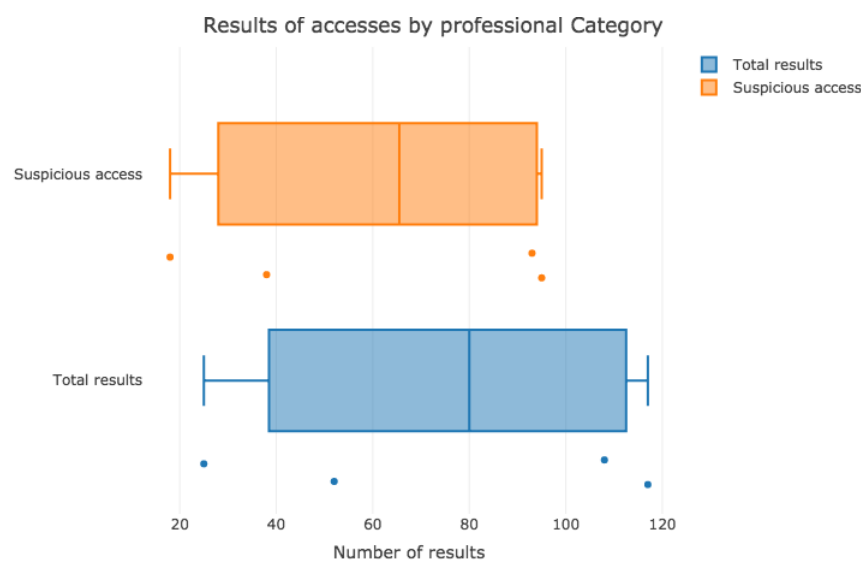


Figure 3.33: Results of access by professional category



Discussion

4. Discussion

There are many concerns about health data protection stated by different personalities with responsibilities in this field. The health data flow complexity is such that turns very hard to track all accesses to EHR and their purpose.

The survey based on discussion meetings with ISD, DPO and the jurist, and the professionals interviews allowed us to describe where, when, how and who access to HIS, motivations for accesses, constraints felt by professionals and how they get over them, and concerns reported by ISD and DPO on this subject. Although the complexity of health data management, it was possible to describe eleven scenarios based on the gathered information, which allowed us to model a use case for illegitimate access detection for each scenario. Sequence diagrams were also produced describing how users and systems interact between each other and what data flows from one to another. This task allowed us to identify the necessary logs' variables that we should use on algorithms to detect suspicious HIS accesses. We took advantage of audit trails, once they aggregate logs from different systems, which contain variables identified in sequence diagrams, and also cross information from various HIS, monitoring actions throughout all institution.

However, at this point, we just had available logs from one applicational system (Obscare) that were being collected to HS.Register in ULSM. So we decided to develop three algorithms for suspicious activity detection that could be tested.

The results immediately show that there are some aspects in common to the three analyses. There are actions that do not have the professional identification and are not null values as well. This actions are automatic processes that are running in background and all of them are classified as suspicious, in UC4: Check time of activity and UC5: Check days of activity use cases, because of their continuous behaviour, exceeding the time limits imposed by algorithm's rules. For use case UC10: EHR read access not all are suspicious because it depends on the automaic processes. Some are just to check information and others update the EHR. Being automatic processes they probably do not represent a threat, nonetheless they should be identified to make easier to spot and interpret them. We can also see that there are various professional categories of users and data management responsibilities, and they are not used regularly. So their accesses appear classified as suspicious accesses due to their pattern of very short usage, specially when comparing with the expected duration for a shift. Even seeming normal at a first glance, it would be recommendable to track the behaviour of these accesses in particular, once they provide confidential data. We think that a detailed analysis of the pattern of these accesses may give further indication of their legitimacy.

Another point of view of the analysis is related to professional category of the staff that does the access. The main categories that have access to HIS are administrative, nurses, physicians and specialist physicians. In general, they present normal activity in what concerns to activity longer then expected

(>780 minutes), 27 in 276 which represent 9,78% of the results. Physicians and Specialist physicians are the categories that have more cases of this type of activity, longer then expected. Some of these suspicious accesses may be explained by the fact that Obscare system is also used in emergency context, and not only for consulting or hospital stay context. In emergency context shifts may be longer then 12 hours, up to 24 hours. The suspicious access detected are in most cases for an activity shorter then expected (<300 minutes), 152 in 276 representing 55% of the results and it is common to all categories. We suppose that the consecutive time of activity on HIS by professionals in general probably may be shorter then six hours.

In the case of consecutive days of work, excluding the automatic processes which are 9 in 318 (2,83%), only the specialist physician category exceeds the expected number of consecutive days of activity, which are 7 in 309(2,27%). According to the information collected on interviews, when there are constraints to access necessary data to care delivery it is usually used the credentials of physicians because they normally do not have EHR access limitations. These occurrences might explain the values obtained.

In general, we see that accesses that show the highest percentage of being suspicious are the ones associated to physicians categories and are in line with our expectations. However a deeper analysis would be necessary to have further conclusions about these results.

Relatively to accesses made to read medical records, the mean of suspicious access is 79% of the total accesses, and when analysing by category we can see that this high percentage is transversal to all categories. Even when excluding the categories that normally access to get reports, every type of management and research categories that access to extrat data and do not update records, the percentage of suspicious access grows to 93%. This means that only 7% of the records are accessed and updated. For this use case we should evaluate again the data that it's being analysed, test it during a longer period of time and find out whether this numbers are correct, performing an analysis on the field.



Conclusion and Future Work

5. Conclusion and Future Work

The scope of the study conducted is very complex and requires a very thorough analysis. Although the difficulties we found along this project it was possible to create a proof of concept of a system to detect suspicious accesses by professionals from healthcare institutions.

Some limitations we have are the short number of interviews. If we had a larger number of interviews we might have more detailed information to create more precise algorithms rules. We think that the analysis in the field, observing the professionals routines, would be also very useful to better understand the results and, probably, change the classification of the access. Another limitation is the availability and quality of HIS logs. Obscare system already has logs prepared for GDPR compliance, but many systems have not and healthcare institutions need to make a great effort on providers to have this information. Even the period of test should be longer than nine days so that we could detect some patterns in the results obtained. Nonetheless it was possible to model the scenarios of undue access and create algorithms to detect suspicious accesses. The results obtained give us a first picture of what is happening at the level of HIS access. A strength of using Obscare system was the fact that it is used on hospital stay, consulting and emergency context. It may explain some of the outliers detected, as the emergency shifts may have different durations. These results must be confirmed at an initial stage and, then, take advantage of this information to create a knowledge base that will allow to apply Artificial Intelligence (AI) models.

Even at this stage, which is still in a very embryonic stadium, the project reveals to be very useful to IS department. They are having the first picture of the access by the healthcare professionals, in the way that is being presented. Having the results of access classification based on the rules created according to staff routines, the identification number of the professional that made the access, the time the access was made and the patient accessed, it gives grounds for DPO and ISD investigate whether the access was in fact illegitimate.

Further work must be performed to completely accomplish the goal of this project. It will be necessary to perform a more detailed analysis to verify the correctness of access classification done by our system and determine its sensibility and specificity. It is important to create algorithms for the use cases left, however to implement them it is necessary to have logs with the necessary data and some of them require previous data treatment. Finally, the production of a knowledge base its recommended so that it will be possible to apply Artificial Intelligence models in the future.



References

6. References

(1995). Data Protection Act.

(1998). Lei n.º 67/98 LEI DA PROTECÇÃO DE DADOS PESSOAIS.

104th Congress, U. (1996). Health Insurance Portability and Accountability Act of 1996. Public Law 104-191.

AR (2016). Lei n.º 26/2016 - Aprova o regime de acesso à informação administrativa e ambiental e de reutilização dos documentos administrativos.

Arnold, K., Gosling, J., and Holmes, D. (2005). The Java programming language. Addison Wesley Professional.

Barrows, R. C. and Clayton, P. D. (1996). Privacy, Confidentiality, and Electronic Medical Records.

Booch, G., Rumbaugh, J., Jacobson, I., and Matter, F. (1996). The unified modeling language user guide.

Cimino, J. J. and Shortliffe, E. H. (2006). Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer-Verlag, Berlin, Heidelberg.

Coburn, A.W., D. J. S. A. B. J. L. E. S. S. H. T. (2018). Cyber Risk Outlook. Technical report, University of Cambridge.

Conselho de Ministros, P. (2018). Resolução do Conselho de Ministros n.º 41/2018.

Costa, G. d. B. P. d. (2014). PhD thesis, Universidade de São Paulo.

Council of the European Union, E. (2016). Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection.

Cruz-Correia, R., Boldt, I., Lapão, L., Santos-Pereira, C., Rodrigues, P. P., Ferreira, A. M., and Freitas, A. (2013). Analysis of the quality of hospital information systems audit trails. BMC Medical Informatics and Decision Making.

Cruz-Correia, R. J., Vieira-Marques, P. M., Ferreira, A. M., Almeida, F. C., Wyatt, J. C., and Costa-Pereira, A. M. (2007). Reviewing the integration of patient data: How systems are evolving in practice to meet patient needs. BMC Medical Informatics and Decision Making.

Delamaire, L., Abdou, H. A. H., Pointon, J., and Others (2009). Credit card fraud and detection techniques: a review. Banks and Bank systems, 4(2):57–68.

- Donovan, F. (2018). Healthcare data presents lucrative target for cyberattackers. Health IT Security.
- European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
- Fernández-Alemán, J. L., Señor, I. C., Lozoya, P. Á. O., and Toval, A. (2013). Security and privacy in electronic health records: a systematic literature review. J. Biomed. Inform., 46(3):541–562.
- Ferreira, A., Cruz-Correia, R., Antunes, L., and Chadwick, D. (2007). Access control: how can it improve patients' healthcare? Stud. Health Technol. Inform., 127:65–76.
- Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE, 11(4).
- Gonçalves-Ferreira, D., Leite, M., Santos-Pereira, C., Correia, M. E., Antunes, L., and Cruz-Correia, R. (2018). HS.Register - an Audit-Trail tool to respond to the general data protection regulation (GDPR). Stud. Health Technol. Inform., 247:81–85.
- Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11(1):1–21.
- Guimarães, R., Rodrigues, P., Santos, A., and Jorge Dinis-Oliveira, R. (2018). Reutilização de registos clínicos para investigação científica: Questões jurídicas relacionadas com a autorização dos titulares e a anonimização. Acta Médica Portuguesa, 31:299.
- Halamka, J. D., Szolovits, P., Rind, D., and Safran, C. (1997). A WWW Implementation of National Recommendations for Protecting Electronic Health Information. Journal of the American Medical Informatics Association, 4(6):458–464.
- Hauer, B. (2015). Data and information leakage prevention within the scope of information security. IEEE Access, 3:2554–2565.
- Issa, H. and Vasarhelyi, M. A. (2011). Application of anomaly detection techniques to identify fraudulent refunds.
- Joint Commission International (2017). JCI Accreditation Standards for Hospitals, 6th Edition.
- Jolliffe, I. T. (1986). Outlier detection, influential observations and robust estimation of principal components. In Jolliffe, I. T., editor, Principal Component Analysis, pages 173–198. Springer New York, New York, NY.
- Li, Z. and Oprea, A. (2016). Operational Security Log Analytics for Enterprise Breach Detection.
- Lobato, A., Andreoni Lopez, M., and Duarte, O. (2016). Um sistema acurado de detecção de ameaças em tempo real por processamento de fluxos. XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC'2016), Salvador, Bahia.
- Morrow, S. (2016). Insider threats at hospitals. Accessed: 2018-12-22.

- Ng, J., Joshi, D., and Banik, S. M. (2015). Applying data mining techniques to intrusion detection. In 2015 12th International Conference on Information Technology - New Generations, pages 800–801.
- Obikee, A. C., Ebuh, G. U., and Obiora-Ilouno, H. O. (2014). Comparison of outlier techniques based on simulated data. Open Journal of Statistics, 4(07):536.
- Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. SIGMOD Rec., 29(2):427–438.
- Séneca, H. (2018). CNPD: Hospital do Barreiro multado em 400 mil euros por permitir acessos indevidos a processos clínicos.
- SPMS (2017). Privacidade da informação no setor da saúde.
- Sun, J., Xie, Y., Zhang, H., and Faloutsos, C. (2008). Less is more: Sparse graph mining with compact matrix decomposition. Stat. Anal. Data Min., 1(1):6–22.
- Valli, C. (2006). The insider threat to medical records: has the network age changed anything?
- Wikina, S. B. (2014). What caused the breach? an examination of use of information technology and health data breaches. Perspect Health Inf Manag., 11.



Annexes

7. Annexes

7.1 Published article in HCIST2019 conference proceedings.



Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies

Illegitimate HIS access by healthcare professionals: scenarios, use cases and audit trail-based detection model

Liliana Sá Correia^{a,b,*}, Ricardo Cruz Correia^c, Pedro Pereira Rodrigues^c

^a*Faculdade de Ciências da Universidade do Porto, Porto, Portugal*

^b*Faculdade de Medicina da Universidade do Porto, Porto, Portugal*

^c*CINTESIS - Centro de Investigação em Tecnologias e Serviços de Saúde, Porto, Portugal*

Abstract

Healthcare institutions face serious security challenges, namely confidentiality, integrity and availability of patient's data due to the amounts of sensitive data collected on Health Information Systems (HIS) and the complex data management processes in health care. This paper describes scenarios of undue HIS access by staff in healthcare institutions, use cases (UC) that model the activities on HIS and identify the variables on audit trails (AT) logs that can be used to detect illegitimate actions on patients' data.

Firstly, a survey was conducted through discussion meetings with Information Systems Director (ISD), Data Protection Officer (DPO) and a jurist to discuss their concerns about patient data access, followed by interviews to professionals from healthcare institutions to gather information about their routines and HIS access practices. Then, undue access scenarios were described and UC of activities on HIS which allow their detection were modelled. Lastly, necessary log variables were identified in order to produce algorithms for illegitimate accesses detection. UC and variables selected were matched with the specific requirements of Ministers Council Resolution (MCR) nr.41/2018 which provides guidelines for technology to be compliant with General Data Protection Regulations (GDPR). Discussions with ISD, DPO and the jurist, and professionals' interviews allowed us to describe nine scenarios of undue access. For each scenario we modelled one UC. 32 variables from different type of logs were identified for illegitimate access detection, of which 14 are mandatory according to MCR nr. 41/2018. Despite we might have some limitations related to poor HIS log quality, the mandatory data that logs must comply will be very useful for the development of UC presented. In addition, it is possible to request systems' vendors the improvement of logs' data to meet the detail we propose for this model, which may be very useful to comply not only with GDPR requirements but also with the Standard "Management of Information" (MOI.11) of Joint Commission International Standards for Hospitals (JCI) certification.

As future work, we intend to develop the algorithms for the UC modelled, that will detect suspicious activities and produce alarmistic in their presence, testing them in real environment of a hospital to help Information Systems department and DPO on investigation and prevention of data breaches.

* Corresponding author. E-mail address: lilianamsacorreia@gmail.com

© 2019 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANAgement / HCist - International Conference on Health and Social Care Information Systems and Technologies

Keywords: HIS access; patient data access; undue access; data breach;

1. Introduction

Since computers were introduced in the routine of healthcare institutions, notable advances were made in providing care. Health Information Systems (HIS) have proliferated in healthcare facilities and the amount of electronic health data has since increased dramatically. A new paradigm was imposed to the practice of medicine and health research thanks to the quantity and quality of information available in this domain [1]. Electronic Health Records (EHR) have been developed to integrate patient data and help healthcare professionals to make the best decisions [2]. Despite the benefits achieved with all the data collected, there are several challenges that are faced by healthcare institutions, namely, ensuring patient privacy, and the data integrity, confidentiality and availability [3], [4]. The concern for personal data protection exists since long time ago. Laws such Data Protection Act [5] and Health Insurance Portability and Accountability Act (HIPAA) [6] date from 1995/96. Nonetheless in May 2018 the GDPR [7] came to reinforce the processes of personal data protection.

Healthcare institutions process great amounts of sensitive data and have serious issues to demonstrate that are compliant with the GDPR, due to the large number of HIS in the institutions. The “Management of Information” Standard (MOI.11) certification by Joint Commission International Standards for Hospitals (JCI) also requires compliance with security requirements related to HER access[8]. Audit Trails (AT) can be very useful on both topics. They comply important requirements, such as audit and traceability of actions on patients' data [9]. However, these repositories are used in most cases as a forensic investigation tool, once a complaint occurs [10] and do not prevent or alert for possible undue accesses. Such occurrences may be data breaches, and if they are, they should be communicated to authorities and even to the affected person, within 72 hours after the breach is known [7]. Moreover, if the healthcare institutions have this knowledge, they can preventively act to avoid future incidents. Healthcare typically implies complex data management processes by many different users, with multiple roles (e.g. physician, head of department, nurse, researcher, auditor) of many patients with whom they have momentarily a relation in a context of healthcare delivery, clinical research, education, among others. Distinguishing non-legitimate access from the legitimate is a very hard task in this environment.

Focused on this problem, we described scenarios and produce a set of UC that model different data access activities and a list of variables that must be monitored to detect suspicious accesses.

2. Methods

First, we discussed the theme of undue health data access on meetings with the Information System Director (ISD) and Data Protection Officer (DPO) from healthcare institutions and an expert jurist on data access laws that is also Data Access Responsible (DAR) of a hospital. From this discussion we described some scenarios of undue access to HIS data.

Secondly, in order to better describe the scenarios of undue access and model UC representing HIS access routines by professionals, a survey based on interviews with eleven questions of qualitative and open answers, to professionals from healthcare institutions was conducted. The information collected was about work routines and HIS access practices: (1) professional category of the interviewed, (2) consecutive hours and days of work in order to define limits for existing activity on HIS by users, including exceptional shifts; (3) HIS used by each professional's category and (4) how they access them, to identify misused profiles or misconfigured permissions; (5) computers used for HIS access; (6) permission for remote access, in order to understand who can access from outside the healthcare institution and for what propose; (7) type of information accessed, to match with profile permission; (8) access limitations felt during their activities and how they get over them, to identify abnormal accesses motivated by task accomplishment like credential sharing; (9) motivations for EHR access that can explain the use of another professional' credentials,

(10) practices adopted by professionals in general and (11) other experiences that the interviewed wish to report so that we could detect different scenarios from the ones we initially previewed. We choose a convenience sample of volunteers to interview from three public hospitals from North Portugal, based on the volunteer's availability, so that it included at least two representatives per profession (physicians, nurses, administrative staff, others) and type of care (primary, secondary or tertiary). The interviews were applied between November 2018 and February 2019. Thirdly, based on gathered information we described scenarios of undue accesses and UC were produced, using Unified Model Language (UML), to represent activities on HIS that could spot those accesses. Lastly, for each UC we identified the necessary log variables to detect suspicious activity.

3. Results

From the information provided by ISD, DPO and the jurist we could identify some major concerns: (1) failed attempts to login Information Systems (IS); (2) the correspondence between the professional category and the access profile; (3) type of information that professionals access; (4) the legitimacy of an access by a professional; (5) remote access; (6) access by inactive professionals.

Table 1. Motivations for access HIS with another professionals' credentials. Each interviewee could order up to 4 reasons based on the frequency.

Motivation	Order of frequency for access motivation				Not answered
	1st	2nd	3rd	4th	
Professional need to access information that own credentials do not allow	24	0	0	0	4
Give patient information to relatives, friends or acquaintances	0	15	3	1	9
Support other professionals	2	4	1	0	21
Curiosity	0	1	2	3	22
Other	--	--	--	--	28

28 professionals from six categories that access HIS (6 physicians, 12 nurses, 1 diagnostic and therapeutic technician (DTT), 2 nutritionists, 1 psychologist, 1 pharmaceuticals, 1 social assistance technician (SAT) and 4 administrative) were interviewed. All interviewees confirmed the schedules usually vary between 6 and 8 hours daily, 35 and 40 hours weekly, and 5 consecutive days per week. Exceptionally, they work 12 hours one day a week, and the head physician may work 24 hours in emergency. Physicians, nurses, DTT, nutritionists, psychologists, pharmaceuticals and SAT have access to many IS with health data, although they have different level of patient data access permissions. Administrative usually do not have access to patient health data. According to all interviewees and ISD, access to HIS that are managed internally by healthcare institution itself, is done by username and password credentials, unique for each user. Username includes the staff id or professional association number and the password is defined by the user. 23 professionals interviewed share computers during the shifts, usually just the head of departments have their own computer to work, as well as they do not always access from the same computer. All professionals can access from any computer on the hospital using their credentials. About remote access, just 1 professional confirm to access remotely by VPN. According to ISD, VPN access is granted by a special authorization by the institution Administration Council, and is necessary for tasks accomplishment, such as hospital encounter coding by coder physicians. Nonetheless remote accesses are always vulnerabilities because they are open doors to get into the HIS. Nurses and DTT stated that they have constraints related to patient data access that makes their job harder. This problem is sometimes solved by asking to head of department or to patient's physician the necessary information. When that is not possible, they, simply, do not access to information needed or, sometimes, another professional with more permission provides his or her credential so that the work can be done without restrictions.

Professionals were asked about motivations for access HIS with another professional credential. The answers obtained were ordered by the occurrence frequency (see Table 1).

Table 2 presents the answers obtained when professionals were asked about the knowledge that they had related to HIS access practices by other professionals in general. It also includes other situations of HIS accesses that could be suspicious of being undue accesses that were not in the predefined list.

Table 2 - HIS access practices by professionals in general

Situation	Frequency
Predefined list	28
Sharing of credentials	10
Use of credentials that belonged to a professional that is not on service on the healthcare institution	14
Patient's data access of a patient without encounter context	18
To help other professional on his or her tasks	18
Other added by the interviewee	4
There are professionals that share credential to get help from other professionals or to help other professionals on their tasks. Medical students are an example of this practice, or the administrative that uses physicians' credentials to help them with some tasks	2
Professionals that use information available in systems to confront other professionals about their personal life	1
Patients that have been blackmailed because someone unduly access to his or her EHR	1

3.1. Scenarios and Use Cases

From the gathered information we describe nine scenarios that may identify undue or, at least, suspicious accesses and match them with the specific requirements from Ministers Council Resolution n° 41/2018 which provides guidelines for technology to be compliant with GDPR [11].

Scenario 1 - Failed access attempts: *A user tries to access a HIS and is not succeeded. The user can try a second chance and succeed or try over again without success.*

Username and passwords are frequently used and so it is not easy to forget them. Consecutive failed login may indicate that something is wrong with the access and must be tracked [11] (see Fig.1).

Scenario 2 - User with wrong profile: *IT department set a less restricted profile to a user because he needs to access more information than usual, to accomplish his tasks for that month. The administrative logs into HIS with a physician profile. The user profile does not match to the user's professional category.*

In this case, a user may need information that usually cannot access to accomplish tasks. This need may legitimate the access [12] and ISD confirmed that it can be done assigning user another profile or customizing profile access permissions. However, when the job is completed if the access permissions are not restored, the access becomes illegitimate since then (see Fig.2). The permissions or profile must be restored.

Scenario 3 - Access of professional out of service: *A user logs into HIS inside the healthcare institution and access to patients EHR but his presence is not registered on the attendance system.*

A user must access HIS when in service performing his or her duties and, thus, their HIS activity must match the time period of the user presence registry. If they do not match, this may indicate that the user is accessing the systems outside the professional activity, and in this case access is undue (see Fig.3). Although it is possible that some exceptions occur, this should not happen and must be monitored [11].

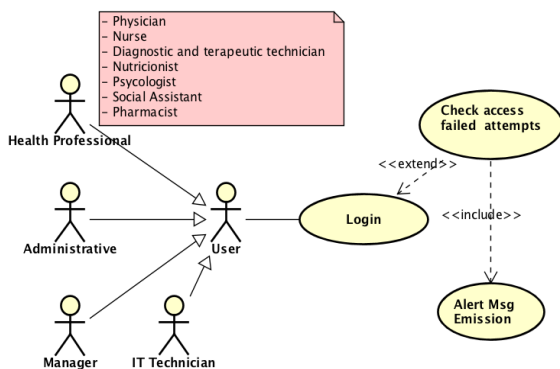


Fig. 1 - UC1: Failed access attempts

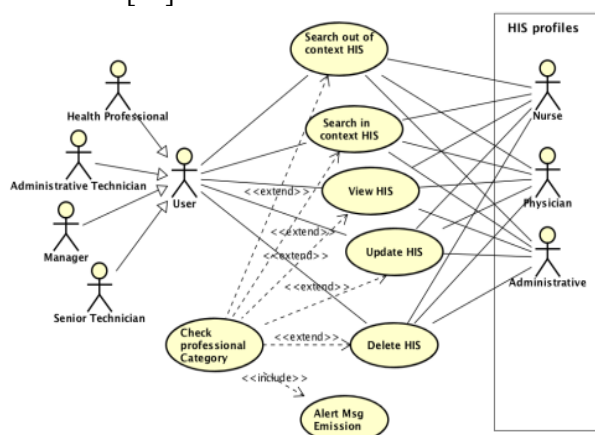


Fig. 2 - UC2: User with wrong profile

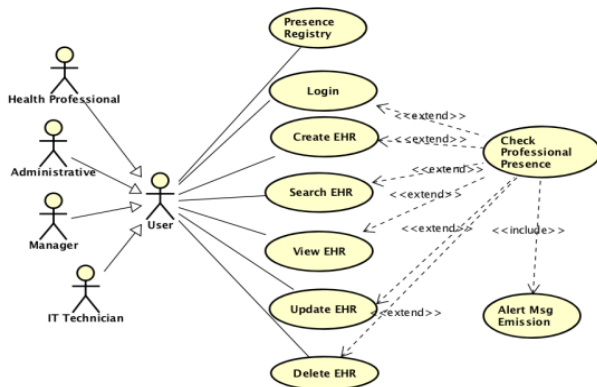


Fig. 3 - UC 3: Access of professional out of service

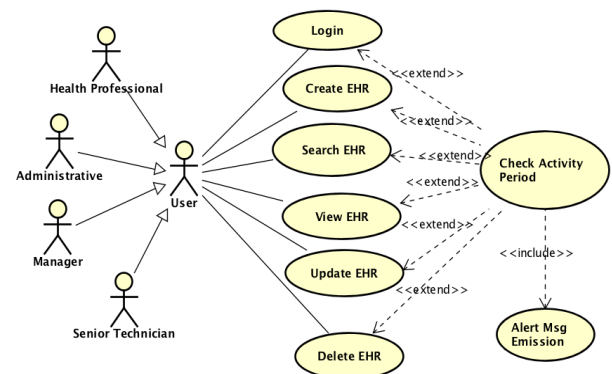


Fig. 4 - UC4: Activity period longer or shorter than the shift time

Scenario 4 - Activity period longer or shorter than the shift time: *A professional uses his credentials during his shift to accomplish his tasks. In the end of his shift another user uses his credential to access a HIS, to take a look at a patient EHR.*

For this scenario we propose counting the period of time of users' actions on HIS, analysing all consecutive period of activity, in terms of daily hours and consecutive works days of credentials. If the added time it is longer or shorter than the shift duration should be, it gives information of suspicious users' activity (see Fig.4). This scenario has the advantage of not depending on the presence system. If AT does not have the necessary information from presence system or if this system is down, it is still possible to detect such accesses this way.

Scenario 5 - Access of an inactive professional: *A user access to HIS with someone's credential, but the credential's owner is no longer working on the healthcare institution. Despite the credential has constant and recent activity, there is no registry on the presence system neither on the payroll system of that professional (the owner of the credential used).*

Active credentials of users that are no longer healthcare institution employees, is a critical issue reported by ISD, and which motivated the first fee applied to a Portuguese Hospital for GDPR noncompliance [13]. It was also reported by 3 professionals interviewed that credentials of professionals that are no longer employees at the institution are still in use by other users. We propose to detect actions of a user that does not have any registry on presence or payroll systems during three months period (see Fig.5) [11].

Scenario 6 - Simultaneous access of same user in different IP address: *A user accesses to a HIS and another user accesses to HIS at the same time with same credentials from another computer. The computer may be from the same department or not.*

This scenario shows the possibility of two different people using the same credential, which results in bad conduct (see Fig.6). However, according to all interviewees it is possible for same professional to access a system from one computer and for a valid reason access again from another computer while the other session is still active, accounting that the physical proximity of computers used allowed it.

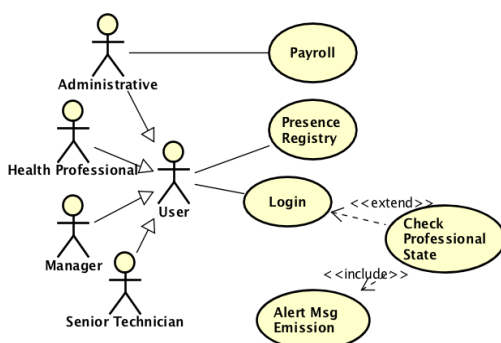


Fig. 5 - UC 5: Access of an inactive professional

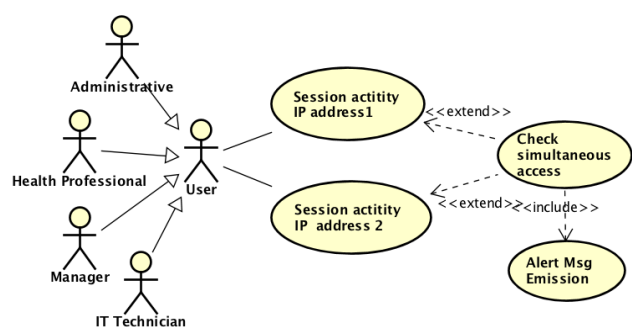


Fig. 6 – UC6: Simultaneous access of a user in different IP address

Scenario 7 - Access to HER of healthcare institution employee: *A patient has an appointment in the same hospital where she works. The physician does not fill the EHR fields because he wants to be sure that no one sees this patient's information. The physician is aware that it happened in the past and can bring harm to the patient, who is also that hospital's employee.*

This scenario shows that there are critical issues concerning patient data privacy, confidentiality and research data quality related to patients which are simultaneously employees on same healthcare facilities. Thus, we propose to check the accesses to employees EHR and produce a report to send the affected person with the information of who accessed to his or her health data (see Fig.7).

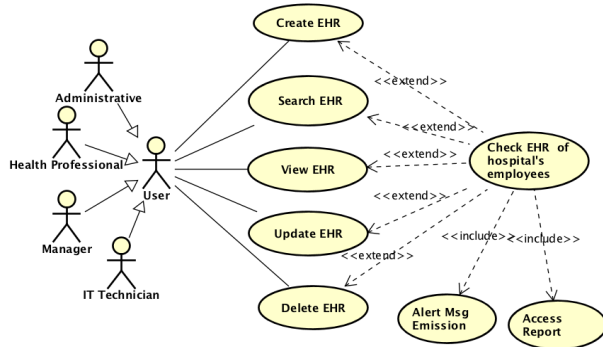


Fig. 7 – UC7: Access to HER of healthcare institution employee

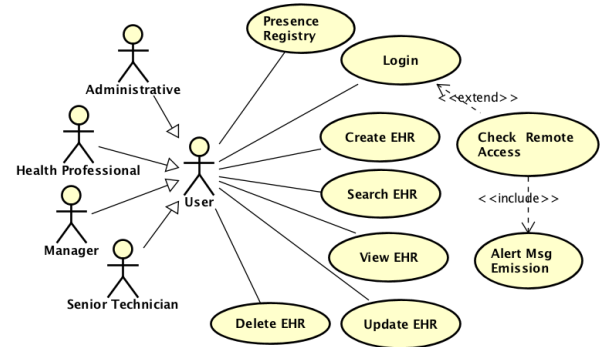


Fig. 8 – UC8: Remote access

Scenario 8 - Remote access: *A employee of healthcare institution, that has a VPN access authorization, accesses late at night and sees some EHR during a short period of time.*

This scenario describes potential risks of VPN accesses. According the ISD they are granted through a special authorization, and they are necessary for task accomplishment such as encounter coding by coder physicians, but in fact, are potential vulnerabilities. It is an open door to get into the HIS and must be monitored. This will allow to identify the VPN accesses on use and the ones that are active without use, and also to track the activity with this type of access (see Fig.8).

Scenario 9 - Access to EHR analysis based on a set of variables: *A professional access to a patient EHR. Why does he access? What is his role? Is he responsible for that patient? Is he giving a second opinion? Has he been following the patient? Did he get Complementary Diagnosis Test and Therapeutic (CDTT) results? Is he prescribing? Does the patient have an appointment? Has he been admitted at the hospital? What was the motivation for the access to occur?*

This describes what should be analyzed when an action occurs on the EHR. How can we be sure that is not an undue access? This is another concern for ISD and DPO, to find out if an access apparently normal is in fact legitimate or not. We propose to perform a context analysis, using different variables' information that can influence decisions about the access legitimacy and present the probability of being undue (see Fig.9).

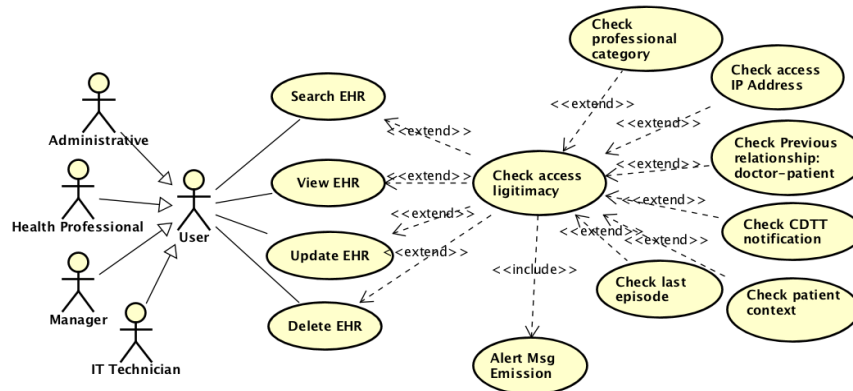


Fig. 9 – UC9: Access to HER analysis based on a set of variables

3.2. Logs Variables

Based on the scenarios described we identified 34 relevant variables needed by the algorithms of undue access suspicions, of which 14 are required to comply with the MRC nr. 41/2018. It is necessary logs from three different sources: System logs, Firewall logs and Application logs (from HIS, presence registry system and payroll system).

Table 3. Relation between log sources and variables, and the UCs. Last column calculates relevance of each variable to all nine UCs

Log Source	Variables	UC1	UC2	UC3	UC4	UC5	UC6	UC7	UC8	UC9	Variable Usage
Systems logs (Logon Session Events)	security id	✓		✓	✓	✓	✓				56%
	account name*	✓		✓	✓	✓	✓				56%
	timestamp*	✓		✓	✓	✓	✓				56%
	logon type code*	✓		✓	✓	✓	✓				56%
	source IP*	✓					✓				22%
Application logs	professional id	✓	✓	✓	✓	✓	✓	✓		✓	88%
	professional category		✓		✓			✓		✓	44%
	account name*	✓	✓	✓	✓	✓	✓	✓		✓	88%
	account profile		✓					✓		✓	33%
	timestamp*	✓	✓	✓	✓	✓	✓	✓		✓	88%
	patient id		✓					✓		✓	33%
	context		✓					✓		✓	22%
	action type			✓		✓		✓		✓	44%
	action id*	✓	✓	✓	✓	✓	✓	✓		✓	88%
	source IP*	✓	✓				✓	✓		✓	56%
	last CDDT notification date							✓		✓	22%
	last patient episode date							✓		✓	22%
	previous patient-physician interaction							✓		✓	22%
Presence system registry	professional id			✓		✓					22%
	timestamp entry registry			✓		✓					22%
	timestamp exit registry			✓		✓					22%
Payroll system	professional id					✓					11%
	month processed					✓					11%
Firewall logs	timestamp*	✓		✓		✓	✓		✓		55%
	account name*	✓		✓		✓	✓		✓		55%
	action type*	✓		✓		✓	✓		✓		55%
	protocol					✓	✓		✓		33%
	source IP*	✓		✓		✓	✓		✓		55%
	destination IP*								✓		11%
	source port*								✓		11%
	destination port								✓		11%
	path								✓		11%
Total of variables per UC		14	9	16	9	19	15	13	9	13	

* Mandatory variables required on the Ministers Council Resolution n° 41/2018

4. Discussion and future work

The survey based on discussion meetings with ISD, DPO and the jurist, and the professionals interviews allowed us to describe where, when, how and who access to HIS, motivations for accesses, constraints felt by professionals and how they get over them, and concerns reported by ISD and the DPO on this subject. Nine scenarios were described using the gathered information and UC for illegitimate access detection were modeled. From the UC we could identify the necessary variables to implement algorithms that will detect suspicious HIS accesses. AT aggregate logs from different systems, which contain variables identified in UC, and have the advantage of crossing information from various HIS and monitor actions throughout all institution. Nonetheless, obtained results will depend on quality of collected logs. If logs comply legal requirements [11] they possess the necessary data to implement algorithms for all UC. However, to obtain a very detailed analysis is necessary more than just the mandatory data. Some AT collect HL7 messages [9], which is very useful to implement the UC 2, 7 and 9 because they include information and workflow from delivery care. Implementation of UC 3 and 5 depends on information from human resources systems as they need the registry of staff activity and, so, it will be necessary to prepare information from presence, payroll or another similar system.

A limitation we might find is the poor logs quality of some HIS [14] and the need to ask systems vendors to produce logs with the necessary data for this propose. Another difficulty is that health data workflow is very complex and it is not trivial to assure that an access is undue when the evidence indicates it is. However, it is possible to identify issues, detect suspicious actions on HIS and associate a probability of being an illegitimate access. Once they are noticed these accesses must be analyzed and classified by humans. There are several reasons for a suspicious access do not be undue, it depends on the task that it is being accomplished at a given moment [8]. If these activities are identified, it will allow conducting an investigation focused on the suspicious access detected.

Despite limitations related to logs quality and access legitimacy uncertainty, the model we want to propose, based on AT logs and probabilities can help to identify abnormal activities on HIS by healthcare institution staff. It will benefit from data collected in AT and act as tool for early breach detection. We also expect that the identifications of such occurrences will allow avoiding illegitimate accesses on the future, providing information for acting preventively. The proposed model may be very useful to comply with GDPR requirements and also with the MOI Standard of Joint Commission International Standards for Hospitals (JCI).

As future work we intend to develop and implement algorithms for the UC modeled, that will detect suspicious activities and produce alarmistic in their presence and test them in real environment of a hospital, helping IS department and DPO on investigation and prevention of data breaches.

Acknowledgements

This work is supported by the Master Degree in Medical Informatics of Faculty of Sciences and Faculty of Medicine of the University of Porto (<http://mim.med.up.pt>).

References

- [1] J. J. Cimino And E. H. Shortliffe, *Biomedical Informatics: Computer Applications In Health Care And Biomedicine (Health Informatics)*. Berlin, Heidelberg: Springer-Verlag, 2006.
- [2] R. J. Cruz-Correia, P. M. Vieira-Marques, A. M. Ferreira, F. C. Almeida, J. C. Wyatt, And A. M. Costa-Pereira, “Reviewing The Integration Of Patient Data: How Systems Are Evolving In Practice To Meet Patient Needs,” *Bmc Med. Inform. Decis. Mak.*, 2007.
- [3] A. Ferreira, R. Cruz-Correia, L. Antunes, And D. Chadwick, “Access Control: How Can It Improve Patients’ Healthcare?,” *Stud. Heal. Technol. Inform.*, Vol. 127, Pp. 65–76, 2007.
- [4] J. L. Fernández-Alemán, I. C. Señor, P. Á. O. Lozoya, And A. Toval, “Security And Privacy In Electronic Health Records: A Systematic Literature Review,” *J. Biomed. Inform.*, Vol. 46, No. 3, Pp. 541–562, 2013.
- [5] “Data Protection Act,” *Official Journal L 281*. P. P. 0031-0050, 1995.
- [6] U. States, “Health Insurance Portability And Accountability Act Of 1996. Public Law 104-191,” *Us Statut.*, Vol. 110, Pp. 1936–2103, 1996.
- [7] E. P. Council Of The European Union, “Directive (Eu) 2016/680 Of The European Parliament And Of The Council Of 27 April 2016 On The Protection Of Natural Persons With Regard To The Processing Of Personal Data.” 2016.
- [8] “Joint Commission International Accreditation Standards For Hospitals,” 6th Edition, 2017.
- [9] D. Gonçalves-Ferreira, M. Leite, C. Santos-Pereira, M. E. Correia, L. Antunes, And R. Cruz-Correia, “Hs.Register - An Audit-Trail Tool To Respond To The General Data Protection Regulation (Gdpr),” *Stud. Heal. Technol. Inform.*, Vol. 247, Pp. 81–85, 2018.
- [10] Z. Li And A. Oprea, “Operational Security Log Analytics For Enterprise Breach Detection,” *Ieee Cybersecurity Development (Secdev)*. 2016.
- [11] P. Do Conselhode Ministros, “Resolução Do Conselho De Ministros N.º 41/2018.” 05-Apr-2018.
- [12] A. Da República, “Lei N.º 26/2016 - Aprova O Regime De Acesso À Informação Administrativa E Ambiental E De Reutilização Dos Documentos Administrativos,” *Diário Da República*, 2016.
- [13] H. Séneca, “Cnpd: Hospital Do Barreiro Multado Em 400 Mil Euros Por Permitir Acessos Indevidos A Processos Clínicos,” *Exame Informática*, 2018. [Online]. Available: [Http://Exameinformatica.Sapo.Pt/Noticias/Mercados/2018-10-19-Cnpd-Hospital-Do-Barreiro-Multado-Em-400-Mil-Euros-Por-Permitir-Acessos-Indevidos-A-Processos-Clinicos](http://Exameinformatica.Sapo.Pt/Noticias/Mercados/2018-10-19-Cnpd-Hospital-Do-Barreiro-Multado-Em-400-Mil-Euros-Por-Permitir-Acessos-Indevidos-A-Processos-Clinicos). [Accessed: 17-Apr-2019].
- [14] I. Boldt, L. Lapao, P. P. Rodrigues, A. Freitas, And R. Cruz-Correia, “Poor Quality Of Hospital Information Systems Audit Trails,” In *Sistemas Y Tecnologias De Informacion, VOLS 1 AND 2*, 2012

7.2 Interview Script

Acesso aos Sistemas de Informação da Saúde

Esta entrevista, composta por 11 questões, tem como objetivo recolher informação sobre a forma como os profissionais das instituições de saúde utilizam os sistemas informáticos no seu dia-a-dia permitindo obter um contexto de base para a realização de estudo sobre acessos indevidos aos registos clínicos dos utentes.

Identificação da categoria profissional do entrevistado

1) Categoria profissional: _____

Rotinas de trabalho: tempo normal de trabalho e exceções (em horas e dias)

2.1) Diário (em horas)

Nº de horas normal do turno: _____ Nº máximo de horas em casos de exceção: _____

2.2) Semanal (em dias)

Nº de dias consecutivos normais de trabalho: _____ Nº máximo de dias em casos de exceção: _____

Caracterização do acesso a Sistemas de Informação/Informáticos ou Plataformas informáticas

3) A que sistemas informáticos /plataformas informáticas têm acesso na sua atividade profissional?

4) Como é feito o acesso aos sistemas informáticos/plataformas informáticas?

4.1) Por senhas de acesso (login e password)? Sim _____ Não: _____

4.1.1) Se respondeu sim: Cada profissional utiliza sempre a mesmo login e password para aceder aos diferentes sistemas informáticos/plataformas informáticas?

4.1.2) Se respondeu não: de que forma é feito o acesso aos sistemas informáticos/ plataformas informáticas?

5) Quando está a realizar as suas funções, acede sempre do computador no seu posto de trabalho habitual? Sim _____ Não _____

6) É normal outros profissionais de categorias diferentes da sua acederem aos sistemas a partir do posto de trabalho onde trabalha habitualmente? Sim _____ Não _____

Acesso Remoto

7) Consegue aceder aos sistemas informáticos/plataformas informáticas a partir de computadores que estão fora do hospital? Sim _____ Não: _____

7.1) Se sim, que motivos podem levar a fazer este tipo de acesso? _____

Tipo de informação no acesso

8) Em algum dos sistemas informáticos/plataformas informáticas que acede com as suas credenciais tem permissão para aceder a informação de saúde dos utentes? Sim _____ Não _____

8.1) Se sim qual/quais? _____

- 9) Existe algum tipo de informação que tenha necessidade de aceder no âmbito das suas funções e à qual não tenha acesso através das permissões da sua credencial (por exemplo algum tipo de informação de saúde do utente ou outra)? Sim _____ Não _____

9.1) Se sim, de que forma é resolvida esta dificuldade? _____

Motivos para acesso com credenciais de outros utilizadores

- 10) Na sua opinião quais são os motivos que dão origem à utilização de credenciais (login e password) de outros profissionais para aceder aos sistemas de informação do hospital? Selecione as opções que na sua opinião se aplicam (indicar quais e por ordem de frequência de ocorrência):

- a. ____ Necessidade profissional de acesso a informação que as suas credenciais de acesso não permitem.
- b. ____ Dar informação a pessoas da família, amigos ou pessoas próximas.
- c. ____ Curiosidade pontual
- d. ____ Dar apoio a outros profissionais na execução das suas funções
- e. ____ Outro (pode especificar?): _____

Ações que podem dar origem a (ou serem suspeitas de) acesso indevidos

- 11) Das seguintes ações, que envolvem os sistemas informáticos/plataformas informáticas, quais é que tem conhecimento que são praticadas por profissionais de instituições de saúde:

- f. ____ Partilha de credenciais
- g. ____ Utilização de credenciais de profissionais que já não estão ao serviço da instituição
- h. ____ Acesso a registos de saúde de pacientes sem contexto de consulta ou seguimento do estado de saúde
- i. ____ Executar tarefas a pedido de outros profissionais
- j. ____ Usurpação de credenciais

- 11.1) Tem conhecimento de outras ações de legitimidade duvidosa que sejam feitos no dia a dia? Sim _____ Não _____

- a. Se sim, quais são os motivos? _____

Obrigada pela colaboração,

Liliana Correia

Estudante de Mestrado de Informática Médica

Faculdade de Ciências e Faculdade de Medicina da Universidade do Porto

7.3 ULSM application for study authorization

Liliana Maria Barroso de Sá Cachada Correia
Estudante da Faculdade de Ciências
e Faculdade de Medicina da Universidade do Porto
email: up201700866@med.up.pt
Telemóvel: 968690747

Ex.mo Sr. Presidente do Conselho de Administração da Unidade Local de Saúde de Matosinhos,

Assunto: Pedido de autorização para realização de investigação

Porto, 24 de setembro 2018

Eu, Liliana Maria Barroso de Sá Cachada Correia, estudante do 2º ano do mestrado de informática médica da Faculdade de Ciências (FCUP) e Faculdade de Medicina da Universidade do Porto (FMUP), venho por este meio solicitar a colaboração da V. prestigiada instituição, no sentido de realizar a recolha de dados do sistema HS.REGISTER para fins de investigação no âmbito da Dissertação de mestrado intitulado "Detection of anomalous activities on personal data of healthcare organizations based on HIS's logs applying a Bayesian Network model", sobre orientação do Doutor Pedro Pereira Rodrigues e coorientação do Doutor Ricardo João Cruz Correia. Os dados recolhidos são confidenciais, acrescentado ainda sob compromisso de honra que o funcionamento da instituição não será posto em causa.

O objetivo deste estudo é desenvolver uma plataforma que consiga detetar ações indevidas sobre os dados através da análise dos *logs* dos Sistemas de Informação na Saúde (SIS), e que possa apoiar o Encarregado de Proteção de Dados (EPD) a investigar se, de alguma forma, se tratam de violação de dados. Pretende-se que seja uma ferramenta que monitoriza a informação contida nos *logs* e chame a atenção do EPD através da emissão de alertas.

No âmbito do projeto Demonstrador HS.REGISTER que a ULSM vai iniciar, em copromoção, com as entidades HealthySystems (HLTSYS) e com a FCUP, e que conta com a instalação de um sistema (HS.REGISTER®) que agrega e guarda os *logs* de todos os SIS num único repositório permitindo cumprir com determinadas exigências do RGPD, tais como a rastreabilidade e atribuição de responsabilidade das ações executadas, considereei uma mais valia para o meu estudo, realizá-lo na vossa instituição.

Anexo I: Apresentação do estudo, descrição dos métodos e cronograma

Anexo II: Curriculum dos investigadores

Cordialmente,



Liliana Correia

Apresentação do Estudo

Investigadores:

Liliana Maria Sá Correia
Estudante da Faculdade de
Ciências e Faculdade de Medicina
da Universidade do Porto
up201700866@med.up.pt

Pedro Pereira Rodrigues
Docente da Faculdade de Medicina
da Universidade do Porto
pprodrigues@med.up.pt

Ricardo João Cruz Correia
Docente da Faculdade de Medicina
da Universidade do Porto
rcorreia@med.up.pt

Introdução

O projeto surge na sequência dos novos desafios estão a emergir para as organizações de saúde face à aplicação do novo Regulamento Geral de Proteção de Dados (RGPD). Apesar da anterior legislação endereçar grande partes das questões tratadas no RGPD, este trás consigo um novo paradigma [1] [2]. As instituições de Saúde passam a assumir maior responsabilidade na forma como atuam em matéria de proteção de dados pessoais.

Os dados da saúde são dados muito apetecíveis por diversas razões [3]. É sabido que existe uma grande preocupação das organizações no que concerne a ataques externos, mas pouca consciencialização sobre as ações indevidas sobre os dados por pessoas de dentro. [4]; [5]; [6].

O controlo das atividades realizadas nos Sistemas de Informação na Saúde (SIS) são uma demanda para estas instituições. Os *logs* dos SIS desempenham um papel essencial no processo de deteção de atividades suspeitas de serem irregulares. De acordo com a SPMS, as instituições de saúde devem munir-se de ferramentas de rastreabilidade que permitam rastrear a acesso, inserção, alteração ou apagamento de dados através dos *logs* dos SIS [2]. Isto poderá ser muito útil ao Encarregado de Proteção de Dados (EPD) no cumprimento das suas atribuições, uma das quais é notificar as autoridades da ocorrência de violação de dados no período de 72 horas após a tomada de conhecimento de tal ocorrência, e caso possa constituir um perigo para os indivíduos afetados, também deverá ser-lhes feita essa comunicação.

O objetivo deste estudo é desenvolver uma plataforma que consiga detetar atividades anormais sobre os dados através da análise dos *logs* dos SIS, e que possa apoiar o EPD a investigar se, de alguma forma, se tratam de violação de dados. Pretende-se que seja uma ferramenta que monitoriza a informação contida os *logs* e chame a atenção do EPD através da emissão de alertas.

No âmbito do projeto Demonstrador HS.REGISTER que a ULSM a iniciar, em promoção, com as entidades HealthySystems (HLTSYS) e com a FCUP considere de grande interesse a introdução do meu estudo neste contexto. O projeto em questão contará com instalação e evolução de um sistema (HS.REGISTER®) que agrega e guarda os *logs* de todos os SIS num único repositório, e que permite cumprir com algumas exigências do RGPD, tais como a rastreabilidade e atribuição de responsabilidade das ações realizadas.

Métodos:

Numa primeira fase, serão analisados os requisitos necessários ao cumprimento do RGPD no âmbito da violação de dados. Será realizado um estudo dos diferentes cenários que constituem ações ilegítimas. Seguir-se-á, um levantamento das variáveis necessárias à identificação dessas ações. Nesta fase será feita leitura de documentos técnicos, legais e científicos, entrevistas a indivíduos que desempenham diferentes funções no processo de gestão de dados, discussão do tema numa atividade de *focus group* e análise de dados, em concreto os dados dos *logs* dos SIS.

Anexo 1

Na segunda fase será desenhado um modelo de rede Bayesiana baseada na informação recolhida. As probabilidades na que serão usadas na rede Bayesiana serão determinadas pela análise dos *logs* e pela informação obtida no *focus group*.

A terceira fase será a implementação de um protótipo da plataforma com o desenvolvimento de, pelo menos, um caso de uso descrito aplicando rede Bayesiana. Nesta fase pretende-se desenvolver um protótipo da plataforma, fazer testes em ambiente real e promover uma avaliação do sistema.

Método de recolha de dados:

Os dados a utilizar no estudo serão dados do sistema HS.REGISTER.

A amostra dos dados deve conter todos os *logs* compreendidos num período de pelo menos 2 meses, sendo utilizados apenas as variáveis que são estritamente necessários ao modelo.

Algumas dessas variáveis já estão identificadas:

- Relacionadas com as ações no sistema: ação executada; identificador da ação; justificação; base legal; deadline de preservação; data e hora;
- Relacionadas com os dados acedidos: Campos acedidos; identificação do documento; tipo do documento; tempo de execução; motivo; dados retornados;
- Relacionadas com o proprietário da ação: categoria; identificador;
- Relacionadas com o proprietário dos dados: categoria; identificador;
- Relacionadas com o sistema dos dados acedidos: Tipo de sistema de dados; departamento; ip; localização; tipo de profissional;
- Relacionadas com o sistema de pedido dos dados: tipo de sistema; departamento; ip; localização; categoria; identificador do utilizador; tipo de utilizador; tipo de profissional; identificador da sessão.

Outras variáveis poderão ser identificadas como necessárias apenas quando os casos de uso estiverem concluídos.

Gestão e análise de dados

Os dados serão analisados na ULSM. Serão exportadas as variáveis necessárias e tratadas com o software R. Os dados serão analisados aplicando as regras para detetar acessos e ações sobre dados não justificados pelo propósito da prestação de cuidados de saúde, de acordo com os casos de uso produzidos.

Será utilizada estatística descritiva e inferencial para a análise e apresentação dos resultados.

Modelo de consentimento Livre e esclarecido:

Não se aplica

Informação de notificação à comissão Nacional de Proteção de Dados (CNPd):

Não se aplica

Anexo 1

Cronograma:

Nr.	Activity Identification	2018						2019					
		08	09	10	11	12	13	14	15	16	17	18	18
		A	S	O	N	D	J	F	M	A	M	J	J
1	Research Protocol	x	x	x	x								
1.1	Literature Review	x											
1.2	Production of protocol research	x	x										
1.3	Authorizations from ethical Committee of ULSM		x	x	x								
2	Stage 1: Requirements Analysis		x	x	x	x							
2.1	Literature Review		x										
2.2	Participants Interviews		x	x									
2.3	Focus Group			x	x								
2.4	Use cases design				x	x							
2.5	HIS's logs analysis					x							
3	Stage 2: Bayesian network model design					x	x						
3.1	Bayesian network model design					x	x						
4	Stage 3: Prototype implementation						x	x	x	x	x		
4.1	Implementation of 1 use case						x	x	x				
4.2	Test of implemented use case									x			
4.3	Evaluation of implemented use case										x		
5	Stage 4: Article submission										x	x	x
	Article submission										x	x	x

Referências

- [1] Council of the European Union, "REGULAMENTO (UE) 2016/679," 2016.
- [2] SPMS, "Privacidade da Informação no setor da Saúde," SPMS, 2017.
- [3] F. Donovan, "Healthcare Data Presents Lucrative Target for Cyberattackers," *Health IT Security*, 09 2018.
- [4] S. B. Wikina, "What Caused the Breach? An Examination of Use of Information Technology and Health Data Breaches," *PMC*, 2014.
- [5] C. Valli, *The Insider Threat to Medical Records: Has the network Age Changed anything?*, ECU Publications, 2006.
- [6] S. Morrow, "resources.infosecinstitute.com," 09 2016. [Online]. Available: <https://resources.infosecinstitute.com/insider-threats-at-hospitals/>. [Acedido em 31 08 2018].

7.4 ULSM study authorization

ULSM 02/11/18 26272 ENT

Teresa Fernandes

De: Paula Teixeira
Enviado: 29 de outubro de 2018 17:24
Para: Teresa Fernandes
Cc: José Castanheira; Tiago Morais
Assunto: FW: Pedido de autorização para realização de investigação
Anexos: Requerimento_Investigacao_15102018085832.pdf; RicardoCorreia_CV- pt-uk.pdf; CV_LilianaCorreia2PT.pdf; Resultado-SAMA2020.pdf

Importância: Alta

Boa tarde Drª Teresa,

Tendo em consideração a apreciação do Dr. José Castanheira em relação ao projeto "HS.REGISTER", em nome da Comissão Local de Proteção de Dados Pessoais da ULSM, consideramos que o mesmo vem apoiar a ULSM na área da monitorização do controlo de acessos às aplicações informáticas disponíveis na ULSM.

Nesse sentido, a CLPDP é de parecer que o projeto vem apoiar as ações a dinamizar no âmbito da implementação do Regulamento Geral da Proteção de Dados Pessoais, não se opondo à realização do mesmo.

Ana Paula Teixeira
 Data Protection Officer ULSM
 Telefone Geral: +351 22 939 10 00 | Directo: +351 22 939 13 69
 Extensão móvel: 6022

De: José Castanheira
Enviada: 26 de outubro de 2018 17:28
Para: Paula Teixeira
Cc: Teresa Fernandes; Tiago Morais
Assunto: FW: Pedido de autorização para realização de investigação

Bom dia Engª Paula Teixeira,

Sobre este pedido de recolha de dados para fins de investigação no âmbito de um projeto designado HS.REGISTER, informamos o seguinte:

1. A ULSM de Matosinhos em 2017 foi convidada a integrar um consórcio para uma candidatura ao SAMA2020, constituído por uma empresa e pela Faculdade de Ciências da Universidade do Porto;
2. O CA da ULSM aceitou o convite para integrar este consórcio. A candidatura obteve financiamento em 2018 conforme de pode verificar no documento em anexo – "Resultado-SAMA2020";
3. O Projeto submetido designa-se por HS.REGISTER ou seja, esta investigação tem a ver com um projeto do qual a ULSM faz parte e, curiosamente, o enquadramento desta candidatura tem a ver com a criação de algumas condições no sentido da conformidade dos sistemas de informação com o RGPD;
4. Como de pode comprovar no documento "Resultado-SAMA2020", este projeto pretende demonstrar num contexto clínico real uma solução para a integração e auditoria de logs produzidos por sistemas clínicos previamente desenvolvida. Pretende-se adicionalmente adaptar a solução por forma a torná-la compatível com

Autógrafa

CONSELHO DE ADMINISTRAÇÃO ULSM, EPE
 Dr. VITOR HERDEIRO 20/10/31
 Presidente
 Dr.ª BEATRIZ DUARTE Prof. Doutor FÁBIO GOMES
 Vogal Diretor Clínico
 Doutora MARGARIDA FILIPE Mestre CARLOS MOUTA Dr.ª TERESA FERNANDES
 Enfermeira Diretora Vogal Vogal

as exigências do Regulamento Geral sobre a Proteção de Dados Pessoais - RGPD. Com a realização do projeto pretende-se atingir os seguintes objetivos:

- a. Implementar em contexto real de utilização numa instituição de saúde o repositório centralizado de logs HS.REGISTER para centralização dos logs (daí a integração da ULSM neste consórcio);
 - b. Realizar um estudo de análise de impacto da privacidade do HS.REGISTER ao nível da informação recolhida, tendo em vista demonstrar a conformidade com o RGPD - Validação de viabilidade legal ao nível da privacidade;
 - c. Desenvolver dashboards com base na informação recolhida pelo integrador HS.REGISTER, específicos para apoiar a monitorização da rastreabilidade da informação, alinhados com o cumprimento das obrigações legais decorrentes do RGPD e com a realidade da instituição de saúde.
5. Assim, com este estudo, pretende-se efetuar uma análise a alguns logs gerados por sistemas de informação (ou passíveis de serem gerados) de modo a validar alguns algoritmos que se pretendem implementar para atingir os objetivos do projeto;
6. Como a ULSM é a instituição do consórcio onde se vai testar o projeto submetido ao SAMA2020, faz todo o sentido que seja a ULSM a fornecer a informação necessária ao desenvolvimento do estudo, não invalidando que a mesma, à luz do RGPD, deva ter a necessária autorização que é o que se está a fazer com este pedido de autorização;
7. A informação a recolher para este efeito corresponde aos logs gerados durante o período de 2 meses contudo, devido ao gap que existe entre as aplicações informáticas e os requisitos do RGPD, teremos ainda de ver quais são os aplicativos que estão em melhores condições para fornecer os logs que se pretende analisar (a informação dos logs é genericamente identificada no documento em anexo "requerimento..." pois os dados em concreto vão depender dos sistemas de informação e a identificação concreta dos dados/logs é um dos resultados que se pretende deste projeto) – ou seja, no âmbito deste projeto, está previsto um trabalho conjunto com a ULSM para identificar estes logs.

Este projeto constitui uma forte evidência da preocupação da ULSM em assegurar que os seus sistemas de informação estão em conformidade com o RGPD e a investigação inerente a este pedido é vital para atingir os objetivos a que o projeto de propõem.

Face ao exposto, o nosso parecer sobre a recolha de dados com vista à realização desta investigação é claramente favorável.

Ao dispor para qualquer esclarecimento adicional.

Com os melhores cumprimentos,
José Castanheira.
Serviço de Tecnologias de Informação e Comunicação
Ext: 5401 | Tlm: 911 058 204

-----Mensagem original-----

De: up201700866 [mailto:up201700866@med.up.pt]
Enviada: 15 de outubro de 2018 09:10
Para: José Castanheira
Cc: Teresa Fernandes
Assunto: Pedido de autorização para realização de investigação

Ex.mo Senhor Diretor do Departamento de Tecnologias da Informação, Dr. José Castanheira,

Eu, Liliana Maria Barroso de Sá Cachada Correia, estudante do 2º ano do mestrado de informática médica da Faculdade de Ciências (FCUP) e Faculdade de Medicina da Universidade do Porto (FMUP), venho por este meio solicitar a colaboração da V. prestigiada instituição, no sentido de realizar a recolha de dados do sistema HS.REGISTER para fins de investigação no âmbito da Dissertação de mestrado intitulado "Detection of anomalous activities on personal data of healthcare organizations based on HIS's logs applying a Bayesian Network model", sobre orientação do Doutor Pedro Pereira Rodrigues e coorientação do Doutor Ricardo João Cruz Correia.

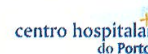
Junto envio o requerimento e respetivos anexos para vossa análise, estando disponível para informação e esclarecimentos adicionais.

Agradeço desde já a vossa melhor atenção para o assunto.

Cordialmente,

Liliana Correia
Estudante do Mestrado de Informática Médica
Faculdade de Ciências e Faculdade Medicina da Universidade do Porto

7.5 CHUP application for study authorization



Departamento de Ensino, Formação e Investigação

Hospital de Santo António – CMIN – CICA – CGMJM

FOLHA DE ROSTO DO ESTUDO DE INVESTIGAÇÃO

TÍTULO

Detection of suspicious activities on EHR of healthcare organizations based on logs of HIS applying a Bayesian Network model.

CLASSIFICAÇÃO

Trabalho Académico de Investigação ☒

Não conferidor de grau ☐ Conferidor de grau ☒ (Licenciatura ☐ Mestrado ☒ Doutoramento ☐)

Projeto de Investigação ☐

Ensaio Clínico ☐ (Medicamentos ☐ Dispositivos médicos ☐)

Outro ☒ Qual? Estudo não Clínico.

VERSÃO

Novo ☒ Modificação / Adenda ☐ Prolongamento ☐

CALENDARIZAÇÃO

Data início: Setembro 2018 Data conclusão: Junho 2019 Prazo a cumprir: Abril 2019

INVESTIGADORES (Aplicável aos Projetos de Investigação e Ensaio Clínicos)

É FUNDAMENTAL O E-MAIL, JÁ QUE ESTE CONSTITUIRÁ A FORMA PREFERENCIAL DE CONTACTO COM OS INVESTIGADORES.

Investigador Principal (Nome, Instituição, Serviço, Grupo profissional) Contactos (e-mail, telefone e telemóvel).

Investigador Responsável no CHUP (Nome, Serviço, Grupo profissional) (Contactos: e-mail, telefone e telemóvel).

Assinale aqui se for idêntico ao Investigador Principal ☐

Indique o Investigador Responsável no CHUP quando o Investigador Principal não pertencer à Instituição.

Outros Investigadores (Nome, Instituição, Serviço, Grupo profissional.)

Indique os restantes elementos da Equipa de Investigação.

ALUNOS E ORIENTADORES (Aplicável aos Trabalhos Académicos de Investigação)

É FUNDAMENTAL O E-MAIL, JÁ QUE ESTE CONSTITUIRÁ A FORMA PREFERENCIAL DE CONTACTO COM OS INVESTIGADORES.

Aluno Nome; Instituição (Universidade / Faculdade ou Escola); Curso e Ano; (Contactos: e-mail, telefone e telemóvel).

Liliana Maria Barroso de Sá Cachada Correia
Faculdade de Ciência e Faculdade de Medicina da Universidade do Porto
Estudante do 2º Ano de Informática Médica
E-mail: up201700866@med.up.pt
Telemóvel: 968690747

Orientador / Supervisor da Instituição de Ensino

Nome; Instituição (Universidade / Faculdade ou Escola); Grupo e Categoria Profissional; Grau Académico; (Contactos: e-mail, telefone e telemóvel).

EQUIPA DE ORIENTAÇÃO

Nome: Pedro Pereira Rodrigues
Instituição: Faculdade de Medicina da Universidade do Porto
Grupo e Categoria Profissional: Docente, Professor Auxiliar
Grau Académico: Doutoramento
Contactos: pprodrigues@med.up.pt

E

CENTRO HOSPITALAR UNIVERSITÁRIO DO PORTO

Morada: Largo Professor Abel Salazar 4099-001 PORTO | E-mail: geral.investigacao.defi@chporto.min-saude.pt | Telefone: +351 222077508

www.chporto.pt

Departamento de Ensino, Formação e Investigação

Hospital de Santo António – CMIN – CICA – CGMJM

Nome: Ricardo João Cruz Correia
Instituição: Faculdade de Medicina da Universidade do Porto
Grupo E Categoria Profissional: Docente, Professor Auxiliar
Grau Académico: Doutoramento
Contactos: rcorreia@med.up.pt

Orientador / Supervisor no CHUP (Nome, Serviço, Grupo profissional) (Contactos: e-mail, telefone e telemóvel).

Assinale aqui se for idêntico ao Orientador / Supervisor da Instituição de Ensino ☐

Indique o Orientador / Supervisor no CHUP quando o Orientador / Supervisor da Instituição de Ensino não pertencer à Instituição.

Nome: Maria Manuel Salazar
Serviço: Sistemas de Informação
Contactos: msalazar@chporto.min-saude.pt

PROMOTOR O próprio ☐ Outro ☐ (especifique no espaço disponível)

INSTITUIÇÕES E SERVIÇOS

Unidades, Departamentos e Serviço do CHUP (de entre as indicadas, mencione qual é a proponente)

Outras Instituições intervenientes (Indique outras Instituições, Unidades, Departamentos e Serviços)

CARACTERÍSTICAS do estudo (Assinale as opções corretas)

Alvo do estudo

Animais ☐ Humanos ☒

Países / Instituições envolvidos

Multinacional ☐ Nacional ☐
Multicêntrico ☐ Institucional ☐

Natureza do estudo

Clínico ☐ Terapêutico ☐
Epidemiológico ☐ Laboratorial ☐
Revisão literatura ☐ Revisão casuística ☐
Outro ☐ Levantamento de processos de acesso aos RCE

Características do estudo (desenho)

Descritivo ☐ Analítico ☐
Observacional ☐ Experimental ☐
Transversal ☐ Longitudinal ☐
Retrospectivo ☐ Prospetivo ☐

Participantes

Existência de grupo controlo: Não ☒ Sim ☐ Seleção dos Participantes: Aleatória ☐ Não aleatória ☐

Estudos observacionais:

Tipo: Caso-controlo ☐ Coorte ☐ Outro ☐

Estudos experimentais:

Conhecimento: Aberto ☐ Cego ☐ (Duplamente cego ☐)
Ensaio Clínicos: Fase I ☐ Fase II ☐ Fase III ☐ Fase IV ☐

Outros aspetos relevantes para a apreciação do estudo:

Participação de grupos vulneráveis Não ☒ Sim ☐ (Crianças ☐ Grávidas ☐ Outros:)
Convocação de doentes / participantes Não ☒ Sim ☐ (especificamente para participar no Estudo de Investigação)
Consentimento informado Não ☒ Sim ☐ (Carta a solicitar dispensa: Não ☐ Sim ☐)
Realização de inquéritos / questionários Não ☒ Sim ☐ (Contacto entre Investigadores e Participantes: Não ☐ Sim ☐)
Realização de entrevistas Não ☒ Sim ☒

CENTRO HOSPITALAR UNIVERSITÁRIO DO PORTO

Morada: Largo Professor Abel Salazar 4099-001 PORTO | E-mail: geral.investigacao.defi@chporto.min-saude.pt | Telefone: +351 222077508

www.chporto.pt

Departamento de Ensino, Formação e Investigação

Hospital de Santo António – CMIN – CICA – CGMJM

Colheita de produtos biológicos Não ☒ Sim ☐ (No CHUP ☐ Noutro local ☐
(Não anonimizados ☐ Anonimizados ☐
(Anonimização reversível ☐ irreversível ☐
Armazenamento de produtos biológicos Não ☒ Sim ☐ (No CHUP ☐ Noutro local ☐
Criação de bancos de produtos biológicos Não ☒ Sim ☐ (No CHUP ☐ Noutro local ☐ (ADN ☐ Outros ☐
(Não anonimizados ☐ Anonimizados ☐
Realização de exames / análises Não ☒ Sim ☐ (No CHUP ☐ Noutro local ☐
Realização de estudos genéticos Não ☒ Sim ☐ (No CHUP ☐ Noutro local ☐
Recolha de dados Não ☒ Sim ☐ (Dados clínicos ☐ Dados laboratoriais: analíticos ☐ /
imagem ☐
Criação de bases de dados Não ☒ Sim ☐ (Não anonimizadas ☐ Anonimizadas ☐

Saída para outras instituições* Não ☒ Sim ☐
(Dados Clínicos ☐ , especificar
Produtos biológicos ☐ , especificar
Imagens ☐ , especificar
Públicas ☐ Privadas ☐ Nacionais ☐ Estrangeiras ☐
Anonimização dos produtos biológicos / dados saídos ☐

*Em cumprimento do Despacho n.º 913-A/2017 de 19 de janeiro de 2017

ORÇAMENTO E FINANCIAMENTO

Orçamento total: _____ Euros Contrato financeiro em anexo: Não ☐ Sim ☐
Financiamento: Interno (CHUP) _____ Euros Externo (Outros) _____ Euros
Entidade(s) financiadora(s): _____
Conflitos de interesse Não ☐ Sim ☐ (Quais? _____)

INDICADORES

Relatórios de progresso ☐ (periodicidade: _____) Relatório final ☐ Outros ☐ Quais? _____

Data: 07/01/2019

Assinatura do proponente (Investigador Responsável / Aluno):



Departamento de Ensino, Formação e Investigação

Hospital de Santo António – CMIN – CICA – CGMJM

PEDIDO DE AUTORIZAÇÃO INSTITUCIONAL

Trabalho académico de investigação (título)

Detection of suspicious activities on EHR of healthcare organizations based on
logs of HIS applying a Bayesian Network model
Presidente do Conselho de Administração do CHUP

Exmo(a). Senhor(a) Presidente do Conselho de Administração do CHUP

Liliana Maria Barroso de Sá Cachada Correia (nome completo), na
qualidade de Investigador Principal ☐ Investigador Responsável ☐ Aluno ☒, vem por este meio, solicitar a Vossa
Exa. autorização para realizar no Centro Hospitalar Universitário do Porto o Estudo de Investigação acima
mencionado, de acordo com o programa de trabalhos e os meios apresentados.

Data

07/01/2019

Assinatura

Liliana Correia

Presidente da Comissão de Ética para a Saúde do CHUP

Exmo(a). Senhor(a) Presidente da Comissão de Ética para a Saúde do CHUP

Liliana Maria Barroso de Sá Cachada Correia (nome completo), na
qualidade de Investigador Principal ☐ Investigador Responsável ☐ Aluno ☒, vem por este meio, solicitar a Vossa
Exa. autorização para realizar no Centro Hospitalar Universitário do Porto o Estudo de Investigação acima
mencionado, de acordo com o programa de trabalhos e os meios apresentados.

Data

07/01/2019

Assinatura

Liliana Correia

Diretor(a) do Departamento de Ensino, Formação e Investigação do CHUP

Exmo(a). Senhor(a) Diretor(a) do Departamento de Ensino, Formação e Investigação do CHUP

Liliana Maria Barroso de Sá Cachada Correia (nome completo), na
qualidade de Investigador Principal ☐ Investigador Responsável ☐ Aluno ☒, vem por este meio, solicitar a Vossa
Exa. autorização para realizar no Centro Hospitalar Universitário do Porto o Estudo de Investigação acima
mencionado, de acordo com o programa de trabalhos e os meios apresentados.

Data

07/01/2019

Assinatura

Liliana Correia

Departamento de Ensino, Formação e Investigação

Hospital de Santo António – CMIN – CICA – CGMJM

TERMO DE RESPONSABILIDADE - ALUNO

Trabalho académico de investigação (título)

Detection of suspicious activities on EHR of healthcare organizations based on logs of HIS applying a Bayesian Network Model.

Aluno

Na qualidade de Aluno, comprometo-me a executar o Trabalho Académico de Investigação acima mencionado, de acordo com o programa de trabalhos e os meios apresentados, respeitando os princípios éticos e deontológicos e as normas internas da instituição.

Aluno

Liliana Maria Barroso de Sá Pacheco Correia

Data

07/01/2019

Assinatura



Instituição de Ensino

Faculdade de Medicina e Faculdade de
Ciências da Universidade do Porto

Curso

Mestrado em Informática
Médica

Ano

2.º

Departamento de Ensino, Formação e Investigação

Hospital de Santo António – CMIN – CICA – CGMJM

TERMOS DE RESPONSABILIDADE - ORIENTADORES

Trabalho académico de investigação (título)

Detection of anomalous activities on EHR's of healthcare organisations based on HIS logs applying a Bayesian Network model."

Aluno

Liliana Maura Barroso de Sá Machado Correia

Orientador da Instituição de Ensino

Na qualidade de Professor Orientador, solicito autorização do Conselho de Administração para que o Aluno acima referido possa desenvolver no CHUP o seu Trabalho de Investigação. Informo que me comprometo a prestar a orientação necessária para uma boa execução do mesmo e a acompanhar o Aluno nas diferentes fases da sua realização, de acordo com o programa de trabalhos e meios apresentados, bem como por zelar pelo respeito dos princípios éticos e deontológicos e pelo cumprimento das normas internas da instituição.

Nome

Ricardo João Cruz Correia

Data

3/01/2019

Assinatura



Instituição de Ensino

Faculdade de Medicina da Universidade do Porto

Cargo / função

Professor Auxiliar

Orientador / Supervisor do CHUP

Na qualidade de Orientador / Supervisor do Aluno no CHUP, comprometo-me a prestar a orientação necessária para uma boa execução do Trabalho de Investigação, de acordo com o programa de trabalhos e meios apresentados. Mais declaro que acompanharei o Aluno, responsabilizando-me por supervisionar a execução do trabalho no CHUP, bem como por zelar pelo respeito dos princípios éticos e deontológicos e pelo cumprimento das normas internas da instituição.

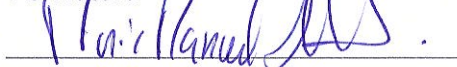
Nome

Mané Manuel Salazar

Data

07/01/2019

Assinatura



Departamento

Serviço

Sistemas de Informação

Grupo profissional

Departamento de Ensino, Formação e Investigação

Hospital de Santo António – CMIN – CICA – CGMJM

TERMOS DE AUTORIZAÇÃO LOCAL

Estudo de investigação (título)

Detection of suspicious activities on EHR of healthcare organization based on logs of HIS applying a Bayesian Network model.

Diretores de Serviço

Na qualidade de Diretor de Serviço, declaro que autorizo a execução do Estudo de Investigação acima mencionado e comprometo-me a prestar as condições necessárias para a boa execução do mesmo, de acordo com o programa de trabalhos e os meios apresentados.

Serviço

Sistemas de Informação

Nome do Diretor

Marta Manuel Salazar

Data

07/01/2019

Assinatura

MARTA MANUEL SALAZAR
DIRECTORA. S.S.I.

Diretores de Departamento

Na qualidade de Diretor do Departamento, declaro que autorizo a execução do Estudo de Investigação acima mencionado e comprometo-me a prestar as condições necessárias para a boa execução do mesmo, de acordo com o programa de trabalhos e os meios apresentados.

Departamento

Nome do Diretor

Data

Assinatura

_____	_____	____/____/____	_____
_____	_____	____/____/____	_____
_____	_____	____/____/____	_____

7.6 CHUP study authorization



SNS SERVIÇO NACIONAL
DE SAÚDE

centro hospitalar
do Porto

Dra. Liliana Correia

Rua da Barranha, 1055, Bloco B R/C Drt Frt

4460-256 Senhora da Hora

ASSUNTO: Trabalho Académico - Doutoramento - **“Detection of suspicious activities on EHR of healthcare organizations based on logs on HIS applying a Bayesian Network model”** – N/ REF.^a 2019.027(024-DEFI/025-CE)

O Conselho de Administração do CHUP autoriza a realização do estudo acima mencionado, a realizar no Serviço de Sistemas de Informação desta Instituição e tendo como Investigador Principal a Dra. Liliana Correia, Estudante do Mestrado em Informática Médica da FCUP/FMUP.

O estudo foi previamente analisado pela Comissão de Ética CHUP|ICBAS, pelo Serviço de Investigação Clínica, pela Direção do Departamento de Ensino, Formação e Investigação do CHUP e pelo Presidente do Conselho de Administração, tendo obtido parecer favorável.

Cumprimentos,

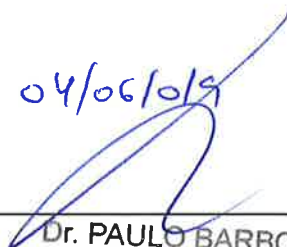
CONSELHO DE ADMINISTRAÇÃO
5/6/2019

Dr. PAULO BARBOSA	Dr.ª ÉLIA GOMES
Presidente	Vogal Executiva
Prof. Doutor JOSÉ BARROS	Dr.ª RITA MOREIRA
Director Clínico	Vogal Executiva
Enf.ª EDUARDO ALVES	
Enfermeiro Diretor	

* Em todas as eventuais comunicações posteriores sobre este estudo é indispensável indicar a nossa ref.^a.

APRECIÇÃO E PARECER PARA A REALIZAÇÃO DE TRABALHO ACADÉMICO – DOUTORAMENTO

Título: "Detection of suspicious activities on EHR of healthcare organizations based on logs on HIS applying a Bayesian Network model"		Ref.º: 2019.027(024-DEFI/025-CE)
		Investigador: Dra. Liliana Correia Estudante da FCUP/FMUP

DIREÇÃO DE ENFERMAGEM: ☒ NÃO SE APLICA ☐ PARECER FAVORÁVEL ☐ PARECER NÃO FAVORÁVEL Data: _____ _____	PRESIDENTE DO CONSELHO DE ADMINISTRAÇÃO: ☒ PARECER FAVORÁVEL ☐ PARECER NÃO FAVORÁVEL Data: 04/06/19  _____ Dr. PAULO BARBOSA Presidente do Conselho de Administração do CHUP
---	---

Em conformidade. Pode ser autorizado


Luísa Lobato
Diretora do DEFI


Prof.ª Doutora Luísa Lobato
Diretora do DEFI

03/06/2019

APRECIÇÃO E VOTAÇÃO DO PARECER

Deliberação	Data: 29.5.2019	Órgão: Reunião Plenária
Título: "Detection of suspicious activities on EHR of healthcare organizations based on logs on HIS applying a Bayesian Network model"		Ref.ª: 2019.027(024-DEFI/025-CE)
Protocolo/Versão: TA-DT	Promotor: o(a) próprio(a)	Investigador / Local: Liliana Maria Barroso de Sá Cachada Correia – Licenciada em Informática e Matemáticas Aplicadas, Estudante do Mestrado em Informática Médica da FCUP/FMUP.

A Comissão de Ética CHUP / ICBAS, ao abrigo do disposto no Decreto-Lei n.º 80/2018, de 15 de Outubro, em reunião realizada nesta data, apreciou a fundamentação do relator sobre o pedido de parecer para a realização do **TA-DT** acima referenciado:

Ouvido o Relator, o processo foi votado pelos Membros da Comissão de Ética CHUP / ICBAS presentes:

Presidente: Dr.ª Luisa Bernardo

Vice-Presidente: Dr.ª Paulina Aguiar

Prof.ª Doutora ~~Carla Teixeira~~, Dr.ª Cármen de Carvalho, Dr.ª Fernanda Manuela Costa, Dr. Gonçalo Senra, Prof. Dr. José António Pinho, Dr.ª Helena Ramos, Prof.ª Doutora Maria Manuel Jorge, Prof.ª Doutora Maria Strecht, Prof. Doutor ~~Paulo Costa~~.

Resultado da votação:

PARECER FAVORÁVEL

A deliberação foi aprovada por unanimidade.

Pelo que se submete à consideração superior.

PARECER FAVORÁVEL

04/06/019
Dr. SEVERO TORRES
Assessor do Presidente do Conselho de Administração

Data 29.5.2019

A Presidente da Comissão de Ética CHUP / ICBAS



Dr.ª Luisa Bernardo

