Development Engineering 5 (2020) 100054

Contents lists available at ScienceDirect

Development Engineering

journal homepage: http://www.elsevier.com/locate/deveng

Data Classification for Secure Mobile Health Data Collection Systems

Marriette Katarahweire [a,*], Engineer Bainomugisha [a], Khalid A. Mughal [b]

[a] Department of Computer Science, Makerere University, Uganda
[b] Department of Informatics, University of Bergen, Norway

ARTICLE INFO

Keywords:
Mobile health
Data collection systems
Security
Data classification
Data sensitivity
Confidentiality

ABSTRACT

Data collected in Mobile Health Data Collections Systems (MHDCS) are diverse, both in terms of type and value. This calls for different data protection measures to meet security goals of confidentiality, integrity, and availability. The majority of commonly used open-source MHDCS track and monitor individuals over a while. It is therefore important to have sensitive data defined and proper security measures identified. We propose a data classification model as a basis for secure design and implementation. Our method combines interviews with case studies. The case studies focused on three of the widely used MHDCS platforms in low-resource settings; that is Muzima, Open Data Kit (ODK), and District Health Information Software (DHIS) 2 Tracker Capture. Interviews with domain experts helped define the sensitivity of data in MHDCS. The proposed data classification model provides for three sensitivity levels: public, confidential, and critical. The model uses context information and multiple parameters as inputs to a classification scheme that maps data to sensitivity levels. The generated data classifications are intended to guide developers and users to build security into MHDCS starting from the early stages of the software development life cycle.

1. Introduction

The pervasive use of mobile devices in the health domain has led to functional needs beyond basic use as data collection tools. Current health-based data collection applications involve tracking of individual participants and events. Among the security requirements for Mobile Health Data Collections Systems (MHDCS) is the need for proper authentication and authorization, secure storage of data and credentials on the mobile device and the server, and secure communication between the mobile device and the server (Mancini et al., 2012; Cobb et al., 2018; Katarahweire et al., 2017; Katarahweire et al., 2019; Iwaya et al., 2019). Furthermore, mobile devices can be lost or stolen, and as the forms are being filled during data collection, there is a potential breach of confidentiality due to shoulder surfing and eavesdropping. MHDCS in low-resource settings are required to work even with no Internet connectivity (Muzima et al., 2016; DHIS, 2016; Open Data Kit, 2020) and this means that the application should be able to store data temporarily on the mobile device until connectivity is restored for transmission to the server for permanent storage.

According to the Open Web Application Security Project (OWASP) for mobile (OWASP, 2017), which ranks vulnerabilities in mobile applications, improper platform usage is one of the most common causes of security vulnerabilities in mobile applications. Improper platform usage is also ranked topmost closely followed by insecure storage in second place. This implies that data needs to be stored securely on the mobile device to minimize the risk of possible attacks. We contend that a good strategy is to support secure design in which security is considered early in the Software Development Life Cycle (SDLC) of the applications.

Since many MHDCS stakeholders including developers, managers, and users are not security experts (Green and Smith, 2016; Balebako et al., 2014; Viega et al., 2001), we seek to enable automation of security strategies as much as possible right from *form level design* which is usually the starting point for MHDCS. Current efforts (Gejibo, 2015; Simplicio et al., 2015; Gejibo et al., 1015) that attempt to enhance secure storage of data in MHDCS apply a blanket solution to all data irrespective of its type, value, and sensitivity. We propose the application of different security mechanisms to data depending on their sensitivity levels which will enhance the performance of the applications, utilize resources optimally while improving the security of the systems. Application of security controls than needed on specific data implies higher computational resources which are not readily available on budget mobile devices common in low-resource settings (Mancini et al., 2011).

In this paper, we specifically address the following Research

* Corresponding author.
E-mail addresses: kmarriette@cis.mak.ac.ug (M. Katarahweire), baino@cis.mak.ac.ug (E. Bainomugisha), Khalid.Mughal@ii.uib.no (K.A. Mughal).

https://doi.org/10.1016/j.deveng.2020.100054
Received 7 December 2018; Received in revised form 6 May 2020; Accepted 20 May 2020
Available online 12 June 2020

Questions (RQs):

RQ1. What kind of data is collected in MHDCS? In this research question, we are interested in finding out the nature and characteristics of data that is typically collected in MHDCS and the forms in which it is stored.

RQ2. What are the parameters for determining the sensitivity levels of data collected in MHDCS? Here we seek to understand the factors that determine and affect the sensitivity of data collected in MHDCS.

RQ3. How do existing MHDCS handle sensitive data? How can data sensitivity be incorporated into existing MHDCS? We aim to find out how existing MHDCS handle sensitive data and if not, how to incorporate data sensitivity into these systems. To address the above research questions, we interviewed domain experts and stakeholders involved in MHDCS. We then undertook a case study on three reference systems where we analyzed forms and data involved in some projects. Furthermore, we conducted reviews of forms used in the data collection process and security mechanisms applied in the reference systems.

The rest of the paper is organized as follows: In Section 2, we briefly describe mobile health data collection systems. Section 3 gives an overview of the related work and in Section 4, we describe the methods used in this study. In Section 5, we discuss the findings of our study concerning research questions RQ1 and RQ2. In Section 6, we describe the proposed data classification model in response to research question RQ3. Section 7 outlines changes to be made in existing MHDCS to enable data classification. Lastly, a conclusion and directions for future work are given in Section 8.

2. Background

In this section, we explain the data collection process and describe the need for data classification in mobile health data collection systems.

2.1. The data collection process

MHDCS enable data collection in the form of electronic forms for surveys, clinical trials and interventions, immunization campaigns, community-extension health visits, maternal and child health, and other purposes in the health sector (Nankabirwa et al., 2017; Style et al., 2017; Mukunya et al., 2019; Macharia et al., 2015). Although data collection also includes sensor devices, our focus is on applications that use electronic forms with a person to fill in the data. These systems consist of a client-server architecture. Fig. 1 illustrates the different processes involved in data collection. There is a form designer who develops the forms and these form definitions are uploaded to a server. The form designer is not necessarily a tech-savvy person. The client is a mobile device on which the application is installed. During data collection, the form definitions are downloaded from the server by a data collector. Partially filled-in and filled-in forms are temporarily stored on the mobile device but with connectivity, the data is extracted and submitted to the server for further processing, analysis, and permanent storage.

The form definitions and data being exchanged between the mobile device and the server must be secured to maintain their integrity and confidentiality. Although it is best practice not to save sensitive data on the mobile device, MHDCS need to save the data temporarily on the mobile device due to intermittent Internet connectivity. For this reason, most MHDCS encrypt the data and the encryption keys are also stored on the mobile device. However, not all data in an MHDCS system are of the same value and kind. Different data have different sensitivity levels according to usage, creation, and purpose. Furthermore, the data temporarily stored on the mobile device and even on the server permanently need to be secured and kept tamper proof. Given that the form is the first executable artifact in the development of MHDCS, it is important to embed security requirements at the form design stage.

2.2. User groups in MHDCS

There are different groups of people in the mobile data collection process. These users may or may not have access to the system.

- Participants are the people who are interviewed or asked as the forms are being filled in. Generally, participants do not have access to the system. However, some systems are much more than simply data collection systems and may soon provide access to participants. One such system is Muzima which is more of a mobile health information system (Muzima et al., 2016).
- A Form Designer is responsible for developing the electronic forms used in the data collection. This person may or may not be a technical person. If the person is a technical person, they work hand-in-hand with the domain experts for the project at hand.
- Data Collector: this is the person with the mobile device who does the actual work of asking or interviewing the participants and filling in the forms. These have access to the application on the mobile device. These data collectors may or may not be technical people depending on their training and the specific project. These may include research assistants, Community Health Workers (CHWs), and medical personnel like nurses and doctors.
- Supervisor: this is a person who is at a level higher than the data collector. They may be more knowledgeable in the domain and may

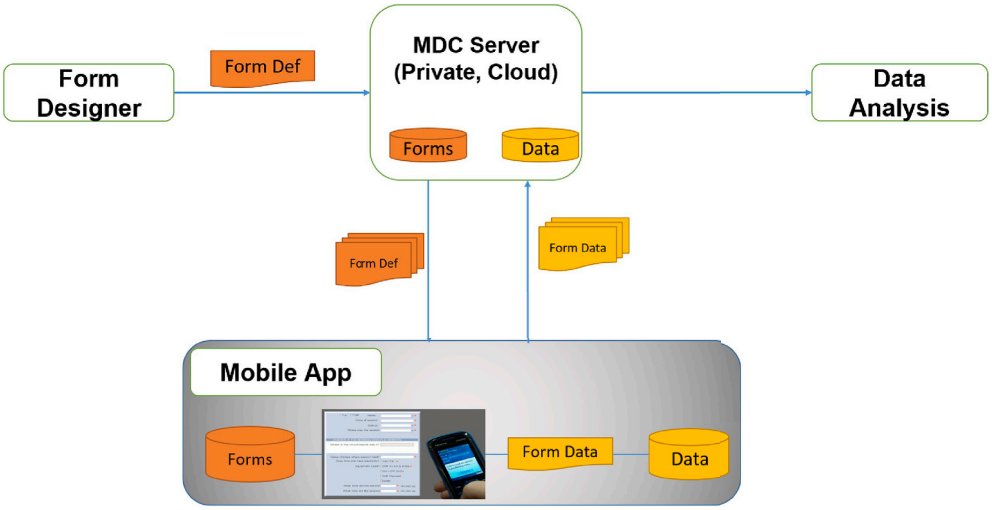

Fig. 1. The Data Collection Process, adopted from (Gejibo, 2015).

be consulted from time to time by the data collectors. Supervisors may have a different level of access to the system from the data collectors.

- Providers are medical personnel involved in the project such as nurses, doctors, laboratory attendants, consultants. These may not have access to the mobile application but may have access to the server for purposes of data processing and analysis.
- System administrators are responsible for the proper working of the entire system including user registration, and configuring the mobile devices and server. They have a different level of access from the data collectors, providers, and form designers.

2.3. Data classification in health

Well known levels of sensitivity include public, internal, confidential (highly confidential), restricted, regulatory, and top-secret (Health Level Seven Intern, 2017; University of Leicester, 2017; Union College Information Technology Services, 2017). They are used in several domains like the military, education, business, and health. The challenge however, lies in determining the sensitivity levels and the category to which the data belongs. Security policies and risk management can help in determining the classifications. Generally, data can be classified depending on its usage, the way and purpose for which it is created, the owner and user, the value and risk associated with its theft or disclosure to unauthorized persons, content of documents, location, and time of access (Shaikh and Sasikumar, 2015).

ISO/IEC 27001:2013 (ISO International Organisation for Standardisation, 2013) is a generic information security management standard. This standard has controls that guide how to manage information security and address associated security risks. ISO 27001 has controls in Annex A particularly A.8 under asset management which emphasizes information classification. The objective is to ensure that each piece of information receives the appropriate level of security. It provides guidelines on how to classify the data specifically in terms of its value, legal requirements, sensitivity, and criticality to the organization. Classification is done in four steps namely; asset identification, information classification, information labeling, and information handling. However, ISO 27001 does not provide the levels of classification though it advises against having too few or too many.

In this work, we classify data according to confidentiality. Confidentiality refers to protecting information from being accessed by unauthorized persons. Sensitive data should therefore not be disclosed to unauthorized persons. For instance, data collected about a participant should not be disclosed to any unauthorized person.

According to the Health Insurance Portability and Accountability Act (HIPAA) (Department of Health, 2009), data in health systems particularly individually identifiable health information need extra protection and is termed Protected Health Information (PHI). PHI falls into two categories: demographics data and medical data. HIPAA provides eighteen (18) identifiers that need to be protected and are individually identifiable data. These identifiers include person's name, telephone number, email addresses, dates such as date of birth or death, discharge date, social security number, account number, address, finger and voice prints, photographic images, and any unique identification numbers like license or vehicle registration numbers.

The HIPAA privacy rule also categorizes a person's physical and mental condition, data that relates to the provision of health care and payment of provision of health care as protected health information. Examples of such data include a person's mental condition, Human Immunodeficiency Virus (HIV) status, drugs administered, and genetic disorders.

Health Level Seven (HL7) International is an ANSI-accredited standards organization that is responsible for developing standards and frameworks for transfer, exchange, integration, and sharing of data between different software and systems in health (Health Level Seven Intern, 2017). HL7 has six confidentiality classifications codes namely; unrestricted, low, moderate, normal, restricted, and very restricted. These confidentiality classes describe the sensitivity of the information concerning whether it should be made available or disclosed to unauthorized individuals, entities, or processes. HL7 describes the kind of data that should be under what classification. We have used these standards in the proposed data classification scheme.

3. Related work

MHDCS either send data to local servers or into the cloud. The context of the related work, therefore, covers mobile health data in the cloud as well as on local servers. Attempts have been made to enhance the security of data in the cloud by applying different security controls depending on the sensitivity level of the data. However, not much work has been extended to handle sensitive health data on mobile devices.

A lot of research has been done on data classification in cloud computing based on machine learning algorithms (Zardari et al., 2014; Kaur and Zandu, 2016; Rubal, 2016; Kiran, 2017; Tamanna, 2017). All the authors contend that it is best to first know the security needs of the data before applying any security mechanisms to it. They, therefore, use different machine learning algorithms to help classify the data into different sensitivity levels. Thereafter, different security mechanisms can be applied accordingly such as encryption algorithms of varying strength and storing sensitive data in inner clouds. Whereas this technique helps classify the data, it needs an initial training set and most of the authors do not specify how to come up with one or simply give examples of sensitive data.

Other authors (Ding and Klein, 2010; Tawalbeh et al., 2015; Zardari et al., 2013) agree that data should first be classified according to security needs and afterwards the right encryption algorithms applied with varying key size and strengths. Zardari et al. (2013) suggest that sensitive data are stored in different clusters of cloud storage. However, there is no clear methodology used by the authors to classify the data apart from (Ding and Klein, 2010) who uses domain experts in health. The other authors give examples of data in each sensitivity class.

Shaikh and Sasikumar (2015) propose a classification of data based on certain parameters namely access control, content, and storage. The authors are of the view that once data is classified, then varying degrees of protection can be applied to the data accordingly. The diverse protection is applied to data during storage and communication in the cloud. However, it is not feasible to use the parameters provided for data classification in MHDCS since the data is stored temporarily on the mobile device, and in some cases, the data may not be edited or accessed again once the data collection form has been saved.

Harel et al. (2012) define a misuseability weight measure, M-score, which can be used to estimate the risk of data leakage to insiders. The M-score value obtained is dependent on the sensitivity level of the data in the tabular database. The sensitivity level of the data is determined by domain experts using a sensitivity score function which gives a value ranging between 0 and 1. They divide attributes into quasi-identifier attributes and sensitive attributes. The sensitivity score function is run on the sensitive attributes for each value of the record in the table in a certain context. Contextual attributes include time, location, and the user's role. Factors that the data owner needs to put into consideration while determining the sensitivity level of a data attribute include privacy and legislation.

Vavilis et al. (2014) extend the Harel et al.'s sensitivity function by developing a data model with hierarchies and inference relations to determine the sensitivity value of data attributes and their instance values. The authors, too, like Harel et al. use the help of domain experts to give the initial set of sensitive attributes and their corresponding sensitivity scores. We build on both models to determine the sensitivity of data with more than one context variable.

4. Methodology

To answer research questions RQ1, RQ2 and RQ3, we combine interviews and case studies. The case studies were used to study current technical implementations for security mechanisms while the stakeholder interviews addressed the operational challenges and concerns during project implementation and data usage.

4.1. Case studies

We analyzed data from three systems widely deployed in low-resource settings namely Muzima (Muzima et al., 2016), DHIS 2 Tracker Capture (DHIS, 2016), and Open Data Kit (ODK) (Open Data Kit, 2020) to understand the kind of data collected. Muzima is an Android-based platform developed by the Institute of Biomedical Informatics, Moi University, Kenya. It is widely deployed in Kenya to monitor patients with chronic illnesses such as hypertension, diabetes, HIV/AIDS, and tuberculosis. DHIS 2 Tracker Capture is an add-on application to DHIS 2 which was first released in 1998. Since then, DHIS 2 has been deployed in many low-resource settings in over sixty countries including Rwanda, Uganda, Ghana, and Kenya (Sahay et al., 2013). ODK is an open-source suite of tools (ODK Build, Collect and Aggregate) designed to empower users to build information services for low-resource settings. Whereas ODK Collect and corresponding ODK Aggregate server tool are general purpose, they have found more use for mobile health systems (Tom-Aba et al., 2015; Labrique et al., 2013; Hartung et al., Borriello) and are commonly deployed as MHDCS.

For each platform, we studied one project implementation. Specifically, we analyzed the electronic forms for data collection, metadata, settings, and configurations. We also carried out reviews of code segments related to security.

4.2. Stakeholder interviews

We engaged various stakeholders involved in the use and implementation of the three reference systems. The participants were drawn from ten institutions, that included two health centers, two application development organizations that focus on mobile health systems, Ministry of Health (MoH) Monitoring and Evaluation department, two Local Governments and three Non-Governmental Organizationss (NGOs) that focus on health monitoring. The common requirement for the selected institutions were those with a deployed MHDCS. The domain experts included five medical officers, five monitoring and evaluation experts, six software developers, six form designers, and ten data collectors.

We carried out customized interviews for the different stakeholders. The focus of the interviews was on the variety of data collected, parameters for sensitivity of data, and how data is currently secured. The results of the interviews were coded into common themes and categories. We coded the interviews using an integrated deductive and inductive approach (Cruzes and Dyba, 2011). The initial set of codes for data categories, concerns and sensitivity levels were defined by a small set of leading domain experts after the first set of interviews.

5. Findings

This section describes the findings for research questions RQ1 and RQ2. With research question RQ1, we are interested in finding out the kind of data collected in MHDCS while RQ2 seeks to understand the factors that influence the sensitivity of the data.

5.1. Findings from domain experts

Fig. 2 gives a breakdown of the views of the domain experts to data classification, sensitivity, and security of MHDCS. Whereas some respondents (78%) agreed that data classification is important for MHDCS, 28% disagreed, and think that all data collected should be treated as sensitive.

When asked about security threats experience with MHDCS, the respondents mentioned the following: data loss due to malicious programs on the mobile device such as viruses, mobile device theft and loss which would lead to loss of the data collected and breach of privacy when the data falls in wrong hands. There were fears too, that data was being collected and accessed by unauthorized persons. It was further observed that confidentiality could be breached as data is transmitted from the mobile device to the server for permanent storage. Data on the mobile device should be stored securely in order to gain trust and confidence in the data collection systems.

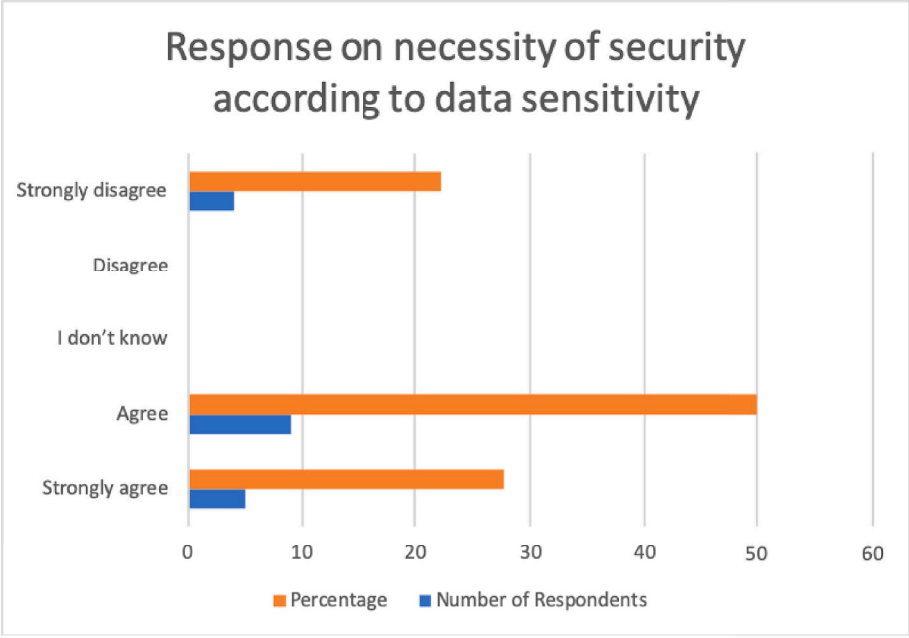

Fig. 2. Number of responses on the need for data classification.

5.2. Categories of data in MHDCS

In this subsection, we provide findings for research question RQ1, in which we present the kind of data collected in MHDCS.

Categories of data in MHDCS were derived from the coding process during case studies and interviews. Most respondents referred to the OpenMRS concept dictionary (Ball, 2018) which we also studied to align the data categories. The concept dictionary defines the medical concepts including questions and answers that are the building blocks for forms in MHDCS. Given that MHDCS are run under different projects with varying objectives and purposes, the data collection forms hold different data. We identified the data categories according to similarity, usage, and purpose. It was noted that some of the data may fall into several categories.

The summary of the data categories is presented in Table 1 and a detailed description is provided in the paragraphs that follow.

Participant characteristics (demographics) This is the kind of data that is used to identify the participant or patient. It includes personal data like names, date of birth, unique identification numbers, gender, telephone number, physical address, biometrics such as fingerprints and facial photographs, vehicle registration number, and email address. Participant characteristics also include socio-economic data about the participant such as their occupation or place of work, level of education, property owned, salary and wages earned, household and family composition, among others.

Personal data can be further categorized into direct and indirect identifiers (Hrynaszkiewicz et al., 2010). Direct identifiers are the kind that can be linked to an individual whereas indirect identifiers need to be matched with other pieces of data to identify an individual. Participant characteristics are therefore divided into direct and indirect identifiers. Direct identifiers include participant's names, telephone number, date of birth, unique identification number, driver's license, physical address, email address, vehicle registration number, fingerprints and facial photographs among others. Indirect identifiers include weight, height, race, tribe, nationality, birthplace, gender, place of work, occupation, family size, and income among others.

Medical Data includes details such as laboratory test results like HIV and malaria blood test results, diagnosis made, treatment and drugs administered, facility and provider information. Medical data also includes radiographic images like ultrasound, X-rays, and Computed Tomography (CT) scans. Some projects may have video and audio recordings of say a child coughing or photo showing a skin condition or even a video showing how an activity like breastfeeding a baby is done.

Meta-data: This category contains data such as facility and data collector identity, mobile device identity including telephone number, Subscriber Identification Module (SIM) card and International Mobile Equipment Identity (IMEI) numbers, Global Positioning System (GPS) coordinates for location, date and time of access, audit trials, server Uniform Resource Locator (URL).

Internal Data: This category covers project related information with details about the project, purpose of the project, duration of the data collection process, consent forms if needed, employee contracts, actual data collected such as results of a trial conducted, effects of a drug administered, survey data and many others depending on the project.

User Credentials category includes all data that is used to grant access to the system such as user names, emails and passwords, fingerprints, location coordinates, bar codes, and encryption and decryption keys.

We observe that it is possible to establish vertical or horizontal relations between some categories. Vertical relations could be in the form of sub-categories, while horizontal relations imply a form of dependence. For instance, disease and prescriptions have a horizontal dependency because the prescriptions apply to specific diseases.

Table 1
Categories of data in MHDCS.

Data Category	Example Data
User credentials	username, password
Participant characteristics	names, identification numbers, telephone number, physical address, birth dates, gender, tribe
Medical Data	Diagnosis, Prescription, Drugs/Medication, Laboratory Tests, Procedures, Findings, Anatomy, Symptoms
Project/Internal Data	data collector details, coverage and location, mobile devices allowed, consent forms, project goals, actual data related to the survey, participant selection criteria
Meta Data	user and device identity such as telephone number, SIM card, time, GPS coordinates

5.3. Sensitivity in MHDCS

In this section, we describe findings from the case studies performed on the three platforms regarding how sensitive data is handled.

On the whole, the three mobile data collection platforms studied handle all data as sensitive and only store images on the external storage card. Furthermore, no encryption is made on the mobile device unless the data is being transmitted to the server from the mobile device for permanent storage.

ODK Collect stores data in readable XML files on internal storage on the mobile device. Some of the projects analyzed coded the data. However, a determined malicious user can be able to decode the data. All images collected are stored on the external storage of the mobile device and can be accessible to any user of the mobile device.

Muzima and DHIS 2 Tracker on the other hand store data in SQLite databases on internal storage of the mobile device which is not easily accessible by a novice user. Muzima, however, uses one encryption key for all mobile devices and as such, it is possible to have access to the data once a user can decrypt this key.

In conclusion, the three platforms studied do not classify data into sensitivity levels apart from the user login and encryption keys which are given security priority. All the other data from electronic forms and even the meta-data are treated the same with no segmentation into classes for security purposes. This work aims to define data classes for sensitivity and appropriate security mechanisms.

5.4. Parameters for data security sensitivity levels in MHDCS

In this subsection, we provide findings for research question RQ2. RQ2 investigates different parameters that influence sensitivity levels of data collected in MHDCS. Through the coding process described in Section 4, we identified the following key parameters that influence the sensitivity of data.

The Participant: the participant in the data collection process may decide to or not disclose the value of a data attribute. For instance, in HIV/AIDS scenarios, some patients choose to disclose their HIV status especially if found to be HIV positive, despite the stigma still associated with it, while others may prefer not to have their status disclosed to even their partners. In other scenarios the social standing of the participant in society is put into consideration and may influence the sensitivity of the data. The participant may be a celebrity or community leader yet it is a sensitive issue like a history of mental illness, fertility issues, and others that may be used against the person if disclosed to the public.

Lifetime of the Data: This includes the changes in societal attitudes over time and also when the data becomes obsolete. With sensitization of the public and as society changes norms, values, and attitudes, what is sensitive now may not be sensitive a few months down the road.

Content: data collected in MHDCS includes multi-media data such as images, videos, and audios. Depending on the content of this data, some may or may not be sensitive data. As some of this data may be medical data, the rest may be informative or educational. For some projects, the social-economic status of participants is gathered, and depending on whether it is stigmatizing or not, the data may or may not be sensitive. This may include property owned, level of education, and salary.

Legal and Regulatory Obligations: Health standards like HIPAA, the Health Information Technology for Economic and Clinical Health (HiTECH) act, and others may impose regulations on how health data should be protected. Other laws and regulations may be from countries such as the General Data Protection Regulation (GDPR) for European Union area (The European Commission, 2018) and the Data Protection and Privacy bill in Uganda (The Goverment of Uganda, 2019).

Importance of the Data to the Organization: The health organization or unit itself may have data that is of more value or importance to them compared to other data. For instance if it is a tuberculosis (TB) clinic, all data related to TB will be more important to them compared to other illnesses. This is defined in their security policies. Data that is most important will be classified as highly sensitive whereas data that is of low importance may be regarded as less sensitive.

Impact of Disclosure: If the impact of disclosure of the data to unauthorized persons is high, then the sensitivity level of that particular data will be high. The effect may lead to legal and psychological issues, reputation, and brand image destruction to both the individual and health organizations collecting the data.

Severity of the Disease: diseases like HIV/AIDS have a different mortality rate than others like flu and fever. Such severe diseases will tend to have a higher sensitivity score compared to the less severe ones.

Context: The context in which the data is collected and handled may affect its sensitivity level altogether. Context includes the location and time when the data was obtained, who is viewing the data, the usage and purpose of the data among others.

It was observed that in some cases a combination of parameters may affect the resultant sensitivity level of data. For instance, the participant may decide to disclose his HIV status and the laws of the land may dictate that HIV status should be private and protected as such. We further noted that the parameters can be grouped into two major categories: those that directly relate to the type of data and those that relate to the context. We thus refer to the first category as *attribute parameters* and the later as *context variables*.

6. Proposed data classification model

To respond to research question RQ3, on how to incorporate data sensitivity into MHDCS, we developed a data classification model. Our proposed model comprises a definition for sensitivity levels in MHDCS, appropriate security controls for each security level, and a data model with a mapping scheme and algorithm. Overall, our data classification model is depicted in Fig. 3.

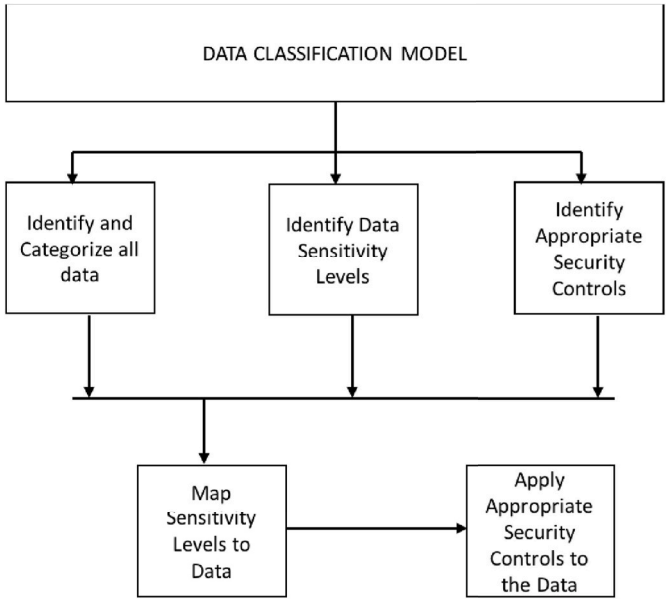

Fig. 3. Proposed data classification model.

6.1. Data sensitivity levels

Well known levels of data sensitivity include public, internal, confidential (highly confidential), restricted, regulatory, and top-secret. These levels have been used in several domains like the military, education, business, and health (Health Level Seven Intern, 2017; University of Leicester, 2017; Union College Information Technology Services, 2017). The challenge, however, lies in determining the sensitivity levels and the assignment of data accordingly. Each domain exhibits unique characteristics that require different sensitivity levels and criteria. Generally, data can be classified depending on its usage, the way and purpose for which it is created, the owner and user, the value and risk associated with its theft, destruction or disclosure to unauthorized persons, content of documents, location and time of access (Shaikh and Sasikumar, 2015; ISO International Organisation for Standardisation, 2016) among others. However, there are times when a data attribute changes from one sensitivity level to another due to certain factors. We sought to find out the determinants of data sensitivity in MHDCS, appropriate security levels and corresponding security mechanisms, and the criteria for mapping data to sensitivity levels.

We propose a list of sensitivity levels for MHDCS based on confidentiality. Confidentiality refers to the protection of information from being accessed by unauthorized persons. As discussed by several authors (Baig et al., 2015; Mancini et al., 2012; Cobb et al., 2018), confidentiality in healthcare is critical and any breach of confidentiality can lead to legal issues, mental and psychological issues for the participants, reputation and brand image damage, blackmail, patient-doctor mistrust, participant discrimination and stigmatization among others.

Several concerns that relate to confidentiality were identified. The concerns affect both the participants and the health organization. Participants concerns include harm, reputation damage, psychological and mental anguish, blackmail, and discrimination. Concerns for the health organization are legal issues that may arise from a breach of confidentiality, damage to their brand image, and patient-doctor mistrust among others.

From the interviews, case studies, and analysis of the concerns, three levels of sensitivity were identified namely: public, confidential, and critical as summarized in Table 2. Furthermore, from the interviews with the domain experts, it emerged that data can be highly sensitive, moderately sensitive, and with little or no sensitivity at all. It is from these, together with established sensitivity levels from the existing health standards that we derived the three sensitivity levels for MHDCS. Table 2 further highlights the potential impact due to a breach of confidentiality. *Public* sensitivity level has the lowest level of sensitivity whereas *Critical* is at the highest level. The security controls applied to critical data have to be more stringent than those applied to confidential data, while public data may have no security controls applied at all.

Public data is the kind of information that is freely available for everyone to see, have, and know with no restrictions. Disclosure of this data causes negligible harm to anyone or the project. The availability of public data without restrictions can be valuable to a wider audience in health implementation and research. Such data may include information about the survey and the project, disease prevention and control, immunization schedules and dates, geographical coverage of the different programs, and others.

Table 2
Data classes in MHDCS.

Data Class	Sensitivity Level	Impact due to Confidentiality Breach
Public	Low	None
Confidential	Moderate	Low to Moderate
Critical	High	High

Data marked *Confidential* is not public but needs some protection. Exposure or leakage of this data to unauthorized persons may cause, to some extent, danger, damage or harm either psychologically, monetary, or in other ways to the participants, the data collectors, the health facility, or the project itself.

Datum that is highly sensitive and potentially stigmatizing is marked as *Critical*. The impact of disclosure of critical data to unauthorized persons is very high. Critical data includes sensitive conditions that the participants may have such as mental cases, genetic disease, HIV status, drug abuse, child abuse, domestic violence. It also includes demographic sensitive information like the participant's social standing in society as a leader, politician, or celebrity. Proprietary or classified information arising from activities such as clinical trials and interventions may also fall under critical data.

7. Discussion and insights

Applying the proposed Data Classification Model to existing MHDCS would require some modifications to the platforms. Generally, all platforms have to be modified such that form fields can have their data classified into appropriate sensitivity levels and tagged accordingly. As the forms are loaded on the mobile devices and data collected, appropriate security mechanisms have to be activated and applied depending on the sensitivity levels of the data.

ODK platform: The ODK platform is a suite of data collection tools including ODK Build, Collect, Aggregate, and Briefcase amongst others. Our focus is mainly on ODK Build and ODK Collect. ODK Build is a tool used for form design by any person who is not technically savvy. ODK Collect on the other hand, runs on the mobile device rendering forms for data collection. The forms designed from ODK Build are loaded into ODK Collect for use during data collection in the field.

To have the data classification model running in ODK tools, ODK Build has to be modified such that data classification can be done for the form fields during form design. ODK forms are in XML format. Furthermore, ODK Collect will be modified to be able to read the sensitivity tags attached to the form fields and have the security mechanisms applied when the form is submitted, saved or loaded.

DHIS 2 Tracker Capture and Muzima do not have a dedicated form designer. Muzima uses HTML5 for form design, and thus, form fields have to be tagged using CSS-style headers. The security mechanisms and corresponding sensitivity levels are defined in the headers. The corresponding handlers for each form, in JavaScript, have to be modified such that the data is validated and classified into sensitivity levels as per the security policies defined in the headers. DHIS 2 on the other hand will have to be modified from the server-side where the form fields are determined and the corresponding data handlers redefined as well.

8. Conclusion and future work

Classifying data into sensitivity levels enables identification of data classes that need protection and appropriate security mechanisms to protect the data. In this paper, we identified key data categories in MHDCS and associated confidentiality concerns. We have defined three sensitivity levels and what needs to be done in three data collection platforms to handle data sensitivity. It is important to note, however, that a piece of data may have various sensitivity levels throughout its lifetime in the system depending on the context. For this, we have proposed a data model that considers multiple context variables and attribute parameters.

As future work, a plug-in implementation of the data classification model will be developed to enable automated annotation of form attributes with sensitivity levels. We shall also explore the management of confidentiality throughout the lifetime of datum as it moves across system boundaries in the MHDCS ecosystem.

Declaration of competing interest

None.

Acknowledgements

We thank the survey and interview respondents for taking their valuable time to give us feedback. This work was supported by the Health Informatics Training and Research in East Africa for Improved Health Care (HI-TRAIN) project under the NORHED program within the Norwegian Agency for Development Cooperation (NORAD).

References

Baig, M.M., GholamHosseini, H., Connolly, M.J., 2015. Mobile healthcare applications: system design review, critical issues and challenges. Australas. Phys. Eng. Sci. Med. 38, 23–38.

Balebako, R., Marsh, A., Lin, J., Hong, J.I., Cranor, L.F., 2014. The privacy and security behaviors of smartphone app developers. In: Workshop on Useable Security (USEC). Internet Society, p. 10.

Ball, E., 2018. Concept Dictionary Basics. https://wiki.openmrs.org/display/docs/Concept+Dictionary+Basics. (Accessed 5 March 2019).

Cobb, C., Sudar, S., Reiter, N., Anderson, R., Roesner, F., Kohno, T., 2018. Computer security for data collection technologies. Develop. Eng. 3, 1–11.

Cruzes, D.S., Dyba, T., 2011. Recommended steps for thematic synthesis in software engineering. In: International Symposium on Empirical Software Engineering and Measurement. IEEE, pp. 275–284.

U.S. Department of Health and Human Services, 2009. Health Information Privacy. https://www.hhs.gov/hipaa/index.html. (Accessed 19 December 2019).

Ding, Y., Klein, K., 2010. Model-driven application-level encryption for the privacy of E-health data. In: ARES 2010, Fifth International Conference on Availability, Reliability and Security, 15-18 February, pp. 341–346. Krakow, Poland.

DHIS 2, Android Tracker Capture App, 2016. https://docs.dhis2.org/2.25/en/android/html/android_tracker_capture.html. (Accessed 1 February 2018).

Gejibo, S., 2015. Towards a Secure Framework for mHealth. Ph.D. thesis, University of Bergen, Norway.

S. Gejibo, F. Mancini, K. A. Mughal, Mobile data collection: a security perspective, in: S. Adibi (Ed.), Mobile Health: A Technology Road Map, Volume 5 of Springer Series in Bio-/Neuroinformatics, Springer, Cham, pp. 1015–1042. ISBN 978-3-319-12817-7.

Green, M., Smith, M., 2016. Developers are not the enemy!: the need for useable security APIs. IEEE Secur. Priv. 14, 40–46.

Harel, A., Shabtai, A., Rokach, L., Elovici, Y., 2012. M-Score: a misuseability weight measure. IEEE Trans. Dependable Secure Comput. 9, 414–428.

C. Hartung, A. Lerer, Y. Anokwa, C. Tseng, W. Brunette, G. Borriello, Open data Kit: tools to build information services for developing regions, in: ICTD '10: Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development, ACM, p. 18.

Health Level Seven International, Security Levels, 2017. http://www.hl7.org/fhir/documentation.html. (Accessed 22 January 2018).

Hrynaszkiewicz, I., Norton, M.L., Vickers, A.J., Altman, D.G., 2010. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ 340.

Iso International Organisation for Standardisation, 2013. ISO/IEC 27000 Family - Information Security Management Systems. https://www.iso.org/isoiec-27001-information-security.html. (Accessed 19 January 2018).

Iso International Organisation for Standardisation, 2016. ISO 27799:2016 Health Informatics – Information Security Management in Health Using ISO/IEC 27002. https://www.iso.org/standard/62777.html. (Accessed 19 January 2018).

Iwaya, L., Fischer-Hübner, S., Åhlfeldt, R.-M., Martucci, L., 2019. Mobile health systems for community-based primary care: identifying controls and mitigating privacy threats. JMIR Mhealth Uhealth 7, e11642.

Katarahweire, M., Bainomugisha, E., Mughal, K.A., 2017. Authentication in selected mobile data collection systems: current state, challenges, solutions and gaps. In: Proceedings of the 4th International Conference on Mobile Software Engineering and Systems, MOBILESoft '17. IEEE Press, Piscataway, NJ, USA, pp. 177–178.

Katarahweire, M., Bainomugisha, E., Mughal, K.A., 2019. In: Rocha, Á., Adeli, H., Reis, L., Costanzo, S. (Eds.), New Knowledge in Information Systems and Technologies. WorldCIST'19 2019. Advances in Intelligent Systems and Computing, 932. Springer, Cham, ISBN 978-3-030-16186-6, pp. 547–556.

Kaur, K., Zandu, V., 2016. A secure data classification model in cloud computing using machine learning approach. Int. J. Grid Distrib. Comput. 9, 13–22.

Kiran, S. Sharma, 2017. Enhance data security in cloud computing using machine learning and hybrid cryptography techniques. Int. J. Adv. Res. Comput. Sci. 8, 393–397.

Labrique, A.B., Vasudevan, L., Kochi, E., Fabricant, R., Mehl, G., 2013. mHealth innovations as health system strengthening tools: 12 common applications and a visual framework. Glob. Health: Sci. Pract. 1, 160–171.

Macharia, P., Abuna, F., Bukusi, D., Dunbar, M.D., Betz, B., Cherutich, P., Sambai, B., Njoroge, A., Farquhar, C., 2015. Enhancing Data Security in Open Data Kit as an mHealth Application, 2015 International Conference on Computing, Communication and Security. ICCCS, pp. 1–4.

Mancini, F., Mughal, K.A., Gejibo, S.H., Klungsøyr, J., 2011. Adding security to mobile data collection. In: IEEE 13th International Conference on E-Health Networking, Applications and Services, pp. 86–89.
Mancini, F., Gejibo, S., Mughal, K.A., Valvik, R.A.B., Klungsøyr, J., 2012. Secure mobile data collection systems for low-budget settings. In: Seventh International Conference on Availability, Reliability and Security, pp. 196–205.
Mukunya, D., Nankabirwa, V., Ndeezi, G., Tumuhamye, J., Bruno Tongun, J., Kizito, S., Napyo, A., Achora, V., Odongkara Mpora, B., Anna Arach, A., Tylleskär, T., Tumwine, J., 2019. Key decision makers and actors in selected newborn care practices: a community-based survey in Northern Uganda. Int. J. Environ. Res. Publ. Health 16, 1723.
Muzima, mUzima, 2016. http://muzima.org. (Accessed 1 February 2018).
Nankabirwa, V., Tylleskär, T., Tumuhamye, J., Tumwine, J., Ndeezi, G., Martines, J.C., Sommerfelt, H., 2017. Efficacy of umbilical cord cleansing with a single application of 4% chlorhexidine for the prevention of newborn infections in Uganda: study protocol for a randomized controlled trial. Trials 18, 322–322.
Open Data Kit, 2020. The Standard for Mobile Data Collection. https://getodk.org//. (Accessed 10 April 2020).
OWASP, 2017. Mobile Top 10 2016 - Top 10. https://www.owasp.org/index.php/Mobile_Top_10_2016-Top_10. (Accessed 20 September 2017).
Rubal, S. Kalra, 2016. Securing data confidentiality in cloud computing using improved boosting technique. Int. J. Eng. Appl. Sci. Technol. 1, 113–119.
Sahay, S., Sæbø, J., Braa, J., 2013. Scaling of HIS in a global context: same, same, but different,. Inf. Organ. 23, 294–323.
Shaikh, R., Sasikumar, M., 2015. Data classification for achieving security in cloud computing. Procedia Comput. Sci. 45, 493–498. International Conference on Advanced Computing Technologies and Applications (ICACTA).
Simplicio, M.A., Iwaya, L.H., Barros, B.M., Carvalho, T.C.M.B., Näslund, M., 2015. SecourHealth: a delay-tolerant security framework for mobile health data collection. IEEE J. Biomed. Health Inform. 19, 761–772.
Style, S., Beard, B.J., Harris-Fry, H., Sengupta, A., Jha, S., Shrestha, B.P., Rai, A., Paudel, V., Thondoo, M., Pulkki-Brannstrom, A.-M., Skordis-Worrall, J., Manandhar, D.S., Costello, A., Saville, N.M., 2017. Experiences in running a complex electronic data capture system using mobile phones in a large-scale population trial in Southern Nepal. Glob. Health Action 10, 1330858. PMID: 28613121.
Tamanna, R. Kumar, 2017. Secure cloud model using classification and cryptography. Int. J. Comput. Appl. 159, 8–13.
Tawalbeh, L., Darwazeh, N.S., Al-Qassas, R.S., AlDosari, F., 2015. A secure cloud computing model based on data classification. Procedia Comput. Sci. 52, 1153–1158. The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015).
The European Commission, 2018. Data Protection: Rules for the Protection of Personal Data inside and outside the EU. https://ec.europa.eu/info/law/law-topic/data-protection_en. (Accessed 10 June 2018).
The Goverment of Uganda, 2019. The Data Protection and Privacy Act, 2019. https://ict.go.ug/wp-content/uploads/2019/03/Data-Protection-and-Privacy-Act-2019.pdf. (Accessed 10 July 2019).
Tom-Aba, D., Olaleye, A., Olayinka, A.T., Nguku, P., Waziri, N., Adewuyi, P., Adeoye, O., Oladele, S., Adeseye, A., Oguntimehin, O., et al., 2015. Innovative technological approach to ebola Virus disease outbreak response in Nigeria using the open data Kit and form hub Technology. PloS One 10, e0131000.
Union College Information Technology Services, 2017. Procedure: Data Classification and Handling. In: https://its.union.edu/sites/default/files/Procedure%20-%20New%20York%20Six%20-%20Data%20Classification%20and%20Handling%201.14.pdf. (Accessed 28 January 2018).
University of Leicester, 2017. University Data Classification Decision Tree and Model. https://www2.le.ac.uk/offices/ias/university-data-classification/decision-tree-model. (Accessed 28 January 2018).
Vavilis, S., Petković, M., Zannone, N., 2014. Data leakage quantification. In: Proceedings of the 28th Annual IFIP WG 11.3 Working Conference on Data and Applications Security and Privacy XXVIII - Volume 8566, DBSec. Springer-Verlag, Berlin, Heidelberg, pp. 98–113.
Viega, J., Kohno, T., Potter, B., 2001. Trust (and mistrust) in secure applications. Commun. ACM 44, 31–36.
Zardari, M.A., Jung, L.T., Zakaria, M.N.B., 2013. Hybrid multi-cloud data security (HMCDS) model and data classification. In: International Conference on Advanced Computer Science Applications and Technologies, pp. 166–171.
Zardari, M.A., Jung, L.T., Zakaria, N., 2014. K-NN classifier for data confidentiality in cloud computing. In: International Conference on Computer and Information Sciences. ICCOINS, pp. 1–6.