2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW) | 978-1-6654-4488-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/EDOCW52865.2021.00052

An information security diagnostic of Electronic Data Capture Systems for the Personal Health Train

Zihan Wang,
Wallace Ugulino,
João Luiz Rebelo Moreira
University of Twente
the Netherlands
z.wang-9@student.utwente.nl,
w.corbougulino@utwente.nl
j.luizrebelomoreira@utwente.nl

ABSTRACT

A Case Report Form (CRF) is a questionnaire mostly used in clinical trial research to collect patients' information. Electronic Case Report Form (eCRF) is taking the place of the normal paper CRF due to its digital convenience. An eCRF is a feature of an Electronic Data Capture (EDC) system and it collects users' information, which might be sensitive. This research aims to find appropriate EDC systems for the Personal Health Train (PHT). This platform enables organizations and researchers to work with health data from diverse sources without sharing these data. Therefore, it is necessary to ensure that the system is prepared for privacy and security regulations. Due to the diversity of privacy regulations worldwide, this research investigates whether the EDC systems comply with the European Union General Data Protection Regulation (GDPR). This paper presents a short review of the PHT concept, the role of the Findable, Accessible, Interoperable and Reusable (FAIR) data principles, potential risks, protection measures for hospital information systems. In our evaluation, we provided an overview of five EDC systems: Castor, OpenClinica, REDCap, NowEDC, and ShareCRF. We evaluated their security and privacy features and techniques to store data. A comparison of their features is given, and a preliminary usability test is presented.

Keywords

Electronic Data Capture (EDC), Case Report Form (CRF), privacy, security, GDPR, Personal Health Train, Health Care

1. INTRODUCTION

In the health trial process and daily work of hospitals, eCRF plays an essential role. A comparison of eCRF and paper CRF carried in 2017 shows that *"eCRF increases time efficiency of data collection in clinical trials, irrespective of item quantity or patient age, and improves data quality"* [9]. The eCRF collects participants' information, including age, gender, medical records, etc. These are all sensitive information that is protected under the GDPR [1]. The GDPR was adopted on 14 April 2016 and became enforceable beginning 25 May 2018 [2]. However, the EDC systems had already been proposed in the late 1980s and became a trend since the US Food and Drug Administration suggested collecting information directly by EDC systems instead of first collecting data on paper and then transcribing it into electronic format on the cloud in 2013 [2]. The time gap between EDC dissemination and its corresponding regulations causes the problem that most existing EDC systems did not consider the GDPR when designing all functionalities. For further development based on these systems, it is crucial to consider the privacy rules and security measures to protect the participants. In this research, we analyze these EDC systems: Castor, OpenClinica, REDCap, NowEDC, and ShareCRF [2,3,4,5,6] about their privacy and security features to support the adoption of such systems in the context of Personal Health Train (PHT).

The current sanitary crisis brought by the Coronavirus pandemic makes it more necessary than ever for medical data to be able to be shared, found and operated on. Due to data sovereignty, the sensitivity and particularity of medical information, most organizations store these data on their premises and are unwilling to share it with others. However, data exchange is necessary for business evolution, e.g., for building improved algorithms. To solve this problem, the PHT approach is conceived to allow a decentralized way of data sharing, where the data follow the "Findable, Accessible, Interoperable and Reusable" (FAIR) principles, both for machines and humans [8]. While the relevance and utility of PHT are increasing during the pandemics, the concerns about privacy and information security become central. They must be assured, which makes necessary an investigation on the existing solutions. To guarantee user's data privacy, the PHT does not bring data to the requester. Instead, it brings the task to the data repository making the task executed in a secure environment. In the PHT approach, the organizations can retain their data in control as they can decide which part of data can be analyzed for the specific task. Data collected for one purpose cannot be used for another without the owner's permission [14]. However, the research by Peg Kerr [13] shows that many institutions share their users' information with their staff and other facilities, which deepens users' concern about the risk of their data leak. Therefore, an EDC system with support to privacy regulations is the basis for achieving the expectations mentioned above.

Personal information is always private, especially in the medical field, as it contains the patients' medical records. Besides the intrinsic harm of disclosing personal information, medical records may cause economic and psychological harm [12]. According to the *"Guidance on the Protection of Personal Identifiable Information"* published by the US Department of Labor [11], personal information can be divided into two categories. The first one is "*The information directly identifies an individual (e.g., name, address, social security number or other identifying number or code, telephone number, email address,*

978-1-6654-4488-0/21/$31.00 ©2021 IEEE
DOI 10.1109/EDOCW52865.2021.00052

etc.)," and the second one is the information by which an individual can be indirectly identified. In the medical field, the second category can be the combination of address and medical records. For example, if Jack is the only one who once had sex reassignment surgery at King's Street, it is not difficult to identify him.

Our research aims to investigate the EDC systems, to check whether they have well-defined security and privacy features, which must be assured in the later development of PHT. In the PHT project, the EDC systems contribute to the process of eCRF generation. After the EDC systems generate the eCRF, they will be transmitted inside the PHT system for organizations to use. Therefore, the EDC systems' privacy and security regulations must be examined. Considering countries worldwide have their policies and regulations of data collection and usage, this project only examines whether these systems manage security issues in compliance with General Data Protection Regulation (GDPR) in Europe. Therefore, the principles listed in the GDPR Chapter 2 [1] were used as our inspection standard.

The remainder of this paper is structured as follows. Section 2 presents a review of the PHT, the FAIR principles, and the privacy protection of EDC systems used as part of hospital information systems. In Section 3, we summarize the results we get, including a comparison of the EDC systems. In Section 4, we proposed how we should protect privacy properly and conclude our findings in Section 5.

2. LITERATURE REVIEW

The review described in this section discusses the Personal Health Train (Section 2.1), the FAIR principles for modeling, sharing, and managing data and their models are explained in Section 2.2. Then, we proceeded with this review by defining Risks and Threats from the viewpoint of this research, as it is described in Section 2.3. Next, in Section 2.4, this review describes the main methods currently used to protect user's privacy. Finally, Section 2.5 concludes this review by presenting some key features of Hospital Information Systems that concern this research.

2.1 Personal Health Train

Data has been applied in many aspects in our daily lives, and most companies are adopting data-driven decision-making. However, at the same time, organizations must ensure data confidentiality. Nowadays, the data is mostly protected by anonymization and data masking. However, these methods cannot mitigate the risk of re-identification [15]. Furthermore, due to the sensitivity and particularity of medical information, most organizations are not willing to share the data with others [16].

Sometimes data exchange is necessary for business, and it is usually done in a centralized way. Centralized sharing means that different groups push their data to a shared database and retrieve others' data from it. The Personal Health Train's main contribution is to implement a move from centralized data sharing to a decentralized way of sharing that gives data owners more control over the use of their data. *"The Personal Health Train (PHT) is a novel approach, aiming at establishing a distributed data analytics infrastructure enabling the (re)use of distributed healthcare data, while data owners stay in control of their own data"* [7]. Although some decentralized storing methods based on Blockchain have been established [17,18,19], the major difference between the PHT and those methods is that the PHT brings models and algorithms to be executed in the data owner premises. In this way, the owner just needs to share the result generated by their data instead of the data itself. According to a survey of 603 participants asking whether they are willing to share their data for better service, 56% show a positive attitude. Still, the premise is that they can gain control of the use of their data [20]. Also, the biggest factor of people participating in a clinical trial is that they trust their information will be kept private and confidential [21].

To implement this, the PHT defines three core components: Station, Train, and Handler [7].

Station: Organizations of the medical field, such as hospitals and health service providers, are the "Stations". It is where data is retrieved to execute the query. In addition, the station provides a secure environment and computational resources to execute the analysis tasks.

Train: *"The set of all artifacts required to execute the distributed algorithm and return the results is called a "Train"*. Every train has its unique Digital Persistent Identifier (DPI). It transfers all required information between the relevant parties. In addition, it carries the Metadata that stores the DPI, algorithms, and models to execute the task, study description, queries to retrieve data.

Handler: It is also called Track, and it works as an intermediate station. It receives the trains and forwards them to the corresponding station by the Metadata. After the train is sent to the station, the result will be aggregated in the handler and sent back to the issuer who proposes the tasks.

During the research on PHT, we find that the terms used to describe the person who creates a train varies in the documentation [7] and the GitHub page [34]. For example, in the documentation it is called Train composer, while it becomes Train issuers on GitHub, which is confusing.

2.2 FAIR data principles

As mentioned in the Introduction, FAIR means "Findable, Accessible, Interoperable and Reusable" In PHT, the FAIR is not only applied to data but also to analytic skills. It pays specific attention to make the machine easy to read and use the data. The difference between humans and machines when searching data on the web is that humans have an intuitive sense of 'semantics'. To make the machine capable of reaching the same level as humans, it must manage tremendous data types, formats, access mechanisms/protocols in its autonomous exploration. At the same time, the data is also required to be formatted to be FAIR. Box 1 lists the FAIR guiding principles [25].

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Box 1. FAIR Guiding principles [25]

For PHT, it ensures **Findability** by the DPI for each train, **Accessibility** by making it possible for every organization to obtain their desired result once they obtain successful authentication and authorization, **Interoperability** by restricting the Metadata transported by the train, and **Reusability** by designing the trains to be reused in multiple locations [7].

2.3 Risks and Threats

This part reviews chapter 3 of the book "For the Record: Protecting Electronic Health Information" [28]. According to statistical analysis, including 400 organizations, 42% have experienced intrusion and 20% did not even realize they had been invaded. In this book, concerns over the privacy and security of Information systems are divided into two categories: *"(1) concerns about inappropriate releases of information from individual organizations and (2) concerns about the systemic flows of information throughout the health care and related industries"* The biggest difference between these two is the first one is about risks of data storage/usage inside one organization, while the other is about the concerns during the data transmission process.

For the first category, the forms of risk differ mainly due to the following factors: **Motive, Resources, Initial Access,** and **Technical Capability. Motives** include economic and non-economic motives such as curiosity. **Resource**, in this research, refers to the resources needed by attackers to execute the crime, and it determines how much damage they can achieve. A precise judgment of the extent of a potential attack helps organizations to determine the input on defense. To date, most attacks against health care organizations are from individuals and small groups. **Initial Access** means the permission the attackers have at the target organization. It ranges from the employee who has full permission to the data to intruders who physically break into the organization and rob the data. Finally, technical Capability refers to techniques that attackers use, and it is constantly evolving, which means the defense needs to be evaluated constantly.

The second concern arises due to the circumstances of developing information technology. For example, a patient's data recorded in a hospital, may be sent to all kinds of institutions like insurance companies, health care organizations, retail pharmacies, or even government programs. This information exchange is necessary as organizations can provide more targeted services based on it, but the exchange process is highly complex and dynamic. Moreover, although GDPR has had a huge influence, the regulations in different organizations vary in detail due to the difference between their stakeholders. This leads to the problem that data cannot be guaranteed to be treated the same as people comprehend.

2.4 Methods to protect privacy

Access control is widely accepted as the most powerful tool for authentication. It has many specific categories, among which Role-Based Access Control (RBAC) is one of the most utilized. It does provide not only authentication but also manages the groups of users that have the same permissions. RBAC divides users into different groups by their needs or positions, and the user's access is restricted in his role. This access granting function is especially attractive to clinical research due to the variety of participants. Its ability to group people diminishes the data leak risk as only the people with needs can access the data. When they are granted a role, they need to sign a consent that they must be responsible for the use or misuse of the information they change and view [23]. However, RBAC brings trouble when the number of users becomes huge as everyone needs to be assigned to a role manually. Another access control is called Attributed-based access control (ABAC). In this model, access to a specific database will be granted when his attributes match the requirements. This includes any kind of attribute (user attribute, environment attribute, etc.). For example, a project checks where you access the database (environmental attribute) and your title (user attribute). This model is regarded as the next-generation model due to its dynamic, context-aware, and risk-intelligent access control. ABAC can dynamically evaluate different attributes compared to the RBAC that checks only the defined set of values [29]. However, both have limitations because they cannot prevent the threat from Internal Agents who abuse their privileges by accessing information for inappropriate reasons or uses [24].

The audit trail is another popular way and broadly applied to check the data flow. It reduces the risk from internal agents as it keeps monitoring the whole system. *"An audit trail (also called audit log) is a security-relevant chronological record, set of records, and/or destination and source of records that provide documentary evidence of the sequence of activities that have affected at any time a specific operation, procedure, event, or device."* [26] With this method, who accesses /views what data is recorded. It can even reconstruct the system from a catastrophic attack [27].

The last part is about the Extensible Access Control Markup Language (XACML) and Web Service privacy (WS-Privacy). They both belong to policy enforcement, which refers to *"The creation, categorization, management, monitoring, and automated execution of a specific set of requirements for the use of a computer or communications network"* [31]. XACML is a standard for evaluating requests according to the policies. It is an implementation of ABAC and is based on the XML language. Its biggest advantage is that it enables segregation of the authorization logic from the application logic, which enables separate administration, auditing, etc., for the authorization [30]. WS-Privacy is a protocol to apply privacy to Web services. It describes how to encrypt/sign SOAP messages such that confidentiality and integrity can be assured.

2.5 Hospital Information system

The main user of PHT is hospitals, so it is necessary to investigate how hospitals benefit from the information systems and what information they collect. By reviewing papers about the performances of Hospital information systems (HIS), we find the main components involved in it. HIS is applied in all hospital departments for various purposes, and it can be divided into seven categories in table1.

Table 1. Roles involved in Hospital Information system [32]

Component	Department	User
Financial Information System (FIS)	Financial	Accountants
Clinical Information System (CIS)	Clinical	Doctors and nurses
Nursing Information System (NIS)	Ward	Nurses and Doctors
Laboratory Information Systems (LIS)	Laboratory	Lab officers
Radiology Information System (RIS)	imaging	Radiologist
Picture Archiving Communication System (PACS)	Imaging	Imaging Officer
Pharmacy Information System (PIS)	Pharmacy	Pharmacist

To have a deeper understanding of the sensitivity of the information in the medical field, we interviewed 2 medical staff. One is a doctor and the other one is a hospital receptionist. They both need HIS but with different aims. This means the information they enter varies. After the interview, we find the information collected by hospitals covers the history of patients' data, personal information, main symptoms, consequences of the research center test, judgments, charging, allergy history. Also, they all agreed that with HIS, the duplication of data passages and time for archiving patients' information is minimized [32].

3. EDC SYSTEMS COMPARISON

This section discusses our evaluation regarding some EDC systems' privacy and security features. From the most popular EDC systems, we selected 5 EDC systems and evaluated them. However, we paid more attention to Castor as it has already been a partner of PHT[40]. They are evaluated by case study and their manuals. The evaluation criteria were determined based on the GDPR principles, as follows:

1. Data is collected for a specific aim and will not be further processed for incompatible purposes.
2. The collection tries to minimize data, which means the collection should be restricted to what is necessary for that purpose.
3. Data must be ensured to be accurate and up to date
4. The identification data cannot be stored longer than needed unless it is needed solely for public interest, historical or scientific research purposes.
5. Data is protected against unauthorized visits and accidental loss, destruction, or damage.
6. The data subject must sign consent and have the right to withdraw his/her consent at any time.
7. The controller should have the ability to restore availability and access control
8. Processing of special categories of personal data, including racial or ethnic origin, political opinions, religious or philosophical beliefs, genetic/biometric data, data concerning sex life or sexual orientation and trade union membership, should be prohibited unless for situations given in GDPR Art.9 [1].

The criteria to compare systems are based on the principles we summarize above. Further explanations of evaluation criteria are described in Section 3.3

3.1 Castor EDC

The structure for Castor consists of three main parts: **Study**, **Phase** and **Step**. A Study can consist of multiple Phases and the same for Step, and Step is where the questions are defined. In this way, questions with different purposes can be categorized. As a case study, we set up an eCRF that WHO published for COVID-19 [35] with Castor[41]. In this example, the **Study's** goal is to document the patient's post-COVID-19 condition. The three modules of the study are the **Phases**. Module 1 includes background demographic and clinical information. Module 2 is to help identify patients who require further evaluation. Module 3 includes the past test results and medical assessments. Under these modules, there are sections with a different focus, and these sections are called **Steps**. During the case study, we find its manual describes every functionality for new users. It includes all functionalities from the most basic ones, like creating a study, to the most advanced techniques, like encrypting and synchronizing data.

What is worth mentioning is that Castor aims at incorporating the FAIR principles into their EDC. They implemented a Resource Description Framework (RDF) endpoint. In addition, they added semantic data to the Castor Study to allow the export of data in RDF format. These methods provide support to make the data FAIR.

To ensure **Confidentiality**, Castor applies a combination of RBAC and Discretionary Access Control (DAC). With DAC, the project owner can decide the list of members who can visit a specific location. DAC is the most flexible access control; However, one security vulnerability is that the profile "admin" can control the whole system. Having one profile that can control the whole system makes it especially vulnerable if the attackers illegally possess an admin account. In Castor, the project owner invites other members by email address and assigns them to an institute. Then, the predefined roles should be set for members. There are three roles: Admin, Data-entry and Monitor. You can also create your role and assign the permissions according to your needs. Figure 1 shows all the permissions included in Castor and how they define the roles.

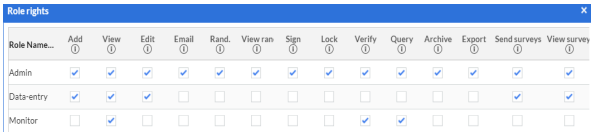


Figure 1. Castor user roles and permissions

Considering that a member may hold different positions in multiple institutions, Castor provides the function to assign a user to different roles in different institutions. In Castor, even if a member has been assigned to a role, the admin can still grant him new permissions or remove one of his permissions that his role should have owned, and that is how DAC works. This feature improves flexibility, as the admin does not have to create a new role for one user or just a small group of members. Besides these, Castor also makes it possible to hide a specific part for certain users. For example, admins can choose to hide a phase, report, or step for a specific role. This means that, for instance, even if a physician has standard access to all Studies, the system may hide a specific Study to this user, which increases the privacy even further (one particular study may be restricted to only the patient and one physician, which is one important right – guaranteed by law in most countries – those patients have in their doctor-patient relationship). Besides access control, Castor provides many other functions to ensure participants' privacy is protected to the greatest extent.

Data Integrity is also one of the requirements in GDPR, and it should be protected in multiple ways. To prevent the mistakes by negligence in data entry, the fields in Castor can be defined to check the input validation to ensure the data collection process complies with GDPR. For example, when collecting personally identifiable information, a polar question (yes/no) should be set for the data entry operator to check whether the participant has already signed the consent. If not, a warning shows and the data entry cannot proceed. Also, data collected from participants can only be put into use with verification. Members with 'verify' permission can mark data as verified after the examination. In case the verified data is modified after verification, the verification will be removed. This prevents the trouble brought by incorrect data modification/entry.

Castor has a monitoring system, including the Audit Trail and a monitoring tab for all queries, verifications and validations. In the monitoring tab, the admin and monitors can check the records to prevent data misuse. Castor's Audit trail is expected to reduce the risk from internal agents as members' every action is recorded there. The tremendous events can be categorized by the view, change, creation, lock, unlock, copy, etc., and the monitor can filter records by type of events, users, or other secondary information.

To improve data safety, an encryption module makes it possible to store the data in an encrypted manner. By this, the data is shown in the form of a random code; only people with an encryption key can view the value behind these codes. For login, Castor provides Two-factor authentication. This makes it harder for attackers to steal data as the login does require not only account information but also the verification code. Besides this, a session in Castor expires in 20 minutes of inactivity.

Except for its advantages, Castor has limitations. As we all know, the first thing to do before collecting information is to ask participants to sign the consent. In our evaluation, we find Castor does not put eConsent as a function into its EDC system (Castor EDC). Instead, they build an individual system called Castor eConsent. In this way, the consent cannot be generated in Castor EDC and users have to switch to the eConsent to build and sign the consent. This causes inconvenience for both using and connecting with PHT.

3.2 Other Systems

For OpenClinica, REDCap, NowEDC and ShareCRF, are all featured with RBAC, Source Data Verification (SDV) and measures to check the validity of fields. The query system is adopted in them to ensure data integrity. The monitor uses it as an inquiry or a notification of incorrect data. Also, the system can automatically create queries to notify the errors made during data entry. SDV is used to verify and review the data to ensure accuracy. It provides an SDV table that displays SDV Status, Open Queries, SDV Requirements, CRF Status, etc., provided for Monitors to operate.

Most of them have a similar structure as Castor, and it also consists of three parts: **Study**, **Event** and **Form.** An Event is a group of Forms that will be used in the Study. Forms are the eCRF we use to collect information. Events are divided into two types: Visit-Based and Common. Visit-Based events refer to the activities scheduled to occur, and usually, it is defined with a date. Common events refer to the events that occur anytime during the daily routine. Informed consent is a typical example of common events.

In the following, we list the features that one system uniquely has.

OpenClinica's forms have the Form Logic, including calculations, cross-form logic, skip logic and constraints. This feature provides support for data entry and prevents mistakes during data collection.

To comply with GDPR, OpenClinica designs diverse functions to ensure data Confidentiality, Integrity and Availability. RBAC is adopted and there are 5 roles: Data manager, Investigator, Monitor, Clinical Research Coordinator and Participant. It individually sets a role for participants and endows the right to view, enter, edit, lock, remove/restore, and re-assign. This measure helps to ensure both confidentiality and integrity at the same time by allowing data subjects to remain their data in control. To ensure data integrity,

OpenClinica built a tool for secure data transmission via an SSH tunnel. It enables a direct read-only connection to the database. The set-up phase consists of three steps: securing the connection, authenticating the connecting user and defining the user's authorization. Then, a 'jump host' ensures that the internal server cannot be directly reached from the internet. In this way, data confidentiality is protected during the transmission process. Besides the protection during transmission, protection for storage is also established.

REDCap provides a function to implement data de-identification when exporting. This method automatically removes dates, performs data shifting, etc.

3.3 COMPARISON

In this section, we compare the five EDC systems that we mentioned. In the evaluation, we found that most of the principles mentioned at the beginning of section 3 by GDPR are administration regulations. It depends on the users to follow the rules, so in this part, we cannot draw a conclusion that a system complies with GDPR. Instead, we focus on whether the systems provide mechanisms to support the execution of these regulations. Finally, we list out the features the systems should include. The core principle to protect security is Confidentiality, Integrity and Availability [37]. By this, we categorized our criteria into the following, and the corresponding GDPR principles, which are listed at the beginning of Section 3 are included:

Confidentiality:

1. Discretionary Access Control (DAC): Whether this system provides DAC. -- 5
2. Role-Based Access Control (RBAC): Whether this system provides RBAC-- 5
3. Multi-Factor Login: This checks whether logging into this system requires more than a password. One example is the Two-Factor Authentication. --5
4. Data Anonymization: This is to evaluate whether the system can encrypt the data such that only people with access can view it. -- 1,5
5. SQL injection/XSS attack defense: This checks whether the system can defend this basic attack. -- 5,7
6. Data Transmission Protection: This checks whether the system provides measures to protect the safety of data transmission. -- 5

Integrity:

7. Monitoring system: This checks whether the system provides monitoring mechanisms to check the validity, queries and verifications. -- 3
8. Data Expiration Check: This is set to check whether the system provides mechanisms to erase the expired data. This is to comply with the GDPR that the data cannot be stored longer than needed. -- 3
9. Source Data Verification: This is a mechanism to check whether the data on CRF is the same as the data source.-- 3
10. Audit Trail: Whether audit trails/logs can be viewed by the administrator. –1,3,5

Other advanced features:

11. RESTful API: This checks whether the system provides a RESTful API to integrate.
12. Open Source: Whether the system is open source.

From Table 2 (next page), we can see that Castor and OpenClinica provide better protections than the others. This is because castor provides more measures to ensure the safety of data storage, while OpenClinica provides a guarantee in data transmission. Also, OpenClinica is open source, which may help in the development of the PHT.

Besides the privacy and security features mentioned above, they both have a RESTful API, making it easy for us to integrate it with PHT. The RESTful service can make PHT more efficient due to its greater portability and scalability [36]. The huge amount of data involved in PHT also makes RESTful service an essential factor for us to consider. The REST architecture also requires security and privacy configurations. There is a checklist to secure REST APIs: Use HTTPS, Password Hash, Never expose information on URLs, Timestamp in request, and Input parameter validation. This helps to achieve stateless APIs [42].

In addition, these principles can be used to check whether the systems implement the FAIR principles. Findability is ensured in all systems we evaluated by a data management system. Criteria 1, 2, 3, 4 ensure Accessibility. The data transmission and exportation provided by Castor and OpenClinica ensure the data

to be Interoperable. To achieve Reusability, descriptions can be added to the data and connections between data can be set up.

Table 2. Comparison table of 5 popular EDC systems

	Castor	Open Clinica	REDCap	NowEDC	ShareCRF
DAC	√	√	√		
RBAC	√	√	√	√	√
Multi-Factor Login	√	√			
Data anonymization	√		√		
SQL injection/XSS defense	√	√	√		
Data Transmission Protection		√			
Monitoring System	√	√	√	√	√
Data expiration check					
Source Data Verification	√	√	√	√	√
Audit Trail	√	√	√		√
Restful API	√	√			
open Source		√ (Only the community edition)			

Although the five systems provide a relatively privacy-friendly environment, we can see that they still lack some features to support organizations to comply with the GDPR. In Table 3, we list out the limitations and give our suggestions for improvements. Most of these suggested improvements function by notifying people who solve the problems. These administration problems can never be solved without people's participation.

Except for the limitations we mentioned in the table, the last one is that both systems do not support multiple languages. This may cause problems for international studies.

Table 3. Limitations and possible improvements

GDPR principles that Castor does not provide support	**Possible improvements**
Data collected for one purpose cannot be used for others without the data subject's permission.	Assign unique Intention Identifiers (Iid) to the studies with the same intentions. When data transmission between studies with different Iid and data exportation is executed, an email asking for data subjects' permission should be automatically sent if no relevant consent was signed in advance.
The collection should try to achieve data minimization	When setting up questions of eCRF, the reasons for proposing each question should be required to be filled. And, the eCRF can only be published in a production environment for clinical trial environments with more than one admins' approval.
The identification data cannot be stored longer than it is needed	A validation period should be defined when setting up the eCRF. When any participant's data expires, the system should notify admins and trigger the data removal process.
Collection of special categories of personal data *(Mentioned in the last GDPR principle at the beginning of Section 3)*should be prohibited.	The special categories of information given in GDPR should be defined beforehand in the system. When a data entry clerk tries collecting this information, the system should notify the particularity of this information. The collection cannot proceed if no consent has been signed

4. DISCUSSION

When talking about data, privacy is always the first thing we care about. There has always been an ethical dilemma: Is it morally right to collect private information for better lives? Nowadays, technology has been adopted to be the key to fighting against many difficult situations. One can argue that we should not resist technology; instead, we need to set up regulations to comply with digital ethics. Digital ethics refers to the moral regulations to counteract the negative effects brought by information technologies. By browsing related research, we summarized the potential data leakage. These threats arise from both internal and external sources, so different measures must be applied to resist.

In our research, we find several kinds of techniques to protect data privacy/security. It includes access control, data anonymization, audit trail, etc. However, we must understand that techniques can only function as support, and we cannot fully depend on them to solve the problem. It is the administration that solves the problems. For example, GDPR requires that the data collection should only include the necessary information and achieve data minimization. This is something that can never be assured only by an EDC system. The system can only provide users with guidance to support such a policy. Therefore, only a combination of strict administration and supportive techniques can draw a comprehensive solution.

After our evaluation, we found that all systems have privacy and security measures. This provides support to the administration of users' data. These features reduce the efforts needed in administration to ensure data safety. The difference between them can be discovered with the comparison table we give. For their limitations, we propose possible improvements. Of course, these measures limit functionality and usability to some extent, but they are necessary. Requirements can be divided into functional and non-functional requirements, while Security and Privacy requirements are categorized as non-functional requirements. A paper by Goertzel states that *"The people involved are not likely to know or care about non-functional requirements. Stakeholders tend to take for granted non-functional security needs."* [38] For an EDC system, the data entry clerks do not pay as much attention to the data they enter as data objects do because that is not their data. However, the participants usually do not have enough knowledge of related regulations. That's why the improvements we mentioned are supposed to be considered.

Besides security/privacy features, we conducted usability testing to assess how people with different backgrounds feel about using the Castor. It is attached in Appendix A. Due to time reasons, the usability testing only included three participants, which – although useful to show some key differences in user's point of view – makes it necessary to perform more comprehensive usability tests for the sake of the completeness and robustness of the usability results. Future works include more usability tests, with more participants and systems, to allow comparison of results.

5. CONCLUSION

The eCRF is being rapidly adopted by health institutions to collect patients' information, but privacy and security issues in data exchange have not yet been solved properly. The PHT is an approach that aims to solve these issues. EDC systems with well-designed privacy and security features are the premise. This research is conducted to answer the following question: *How do the existing EDC systems support companies to comply with GDPR and which is the best choice for the PHT approach?*

We concluded that OpenClinica, Castor, NowEDC, RedCap and ShareCRF as the most appropriate EDC systems for the PHT. As a preparation, we analyzed the potential attacks that a clinical trial system may suffer. A summary of popular methods to ensure privacy and security is given, along with PHT requirements and how they relate to the FAIR data principles. This literature review helped to define the criteria for the evaluation phase.

We evaluate the systems by assessing how they comply with the GDPR principles. For principles about administration problems, we check whether the systems provide support to address them. After comparing the five systems, we concluded that Castor and OpenClinica are the most appropriate EDC systems for the development of the PHT. Their access control, audit trails and other security measures, such as Multi-Factor login, provide support for organizations to comply with the GDPR. However, Castor provides relatively diverse protection measures compared to OpenClinica. Castor can anonymize data, is designed to impede regular cybersecurity attacks and provide data access controls for specific groups of people. These are all features that OpenClinica does not provide. Thus, we recommend Castor as an appropriate candidate for the PHT development. Moreover, to understand Castor's usability, we also conducted usability testing, which showed that Castor is adequate for experienced medical staff but needs improvements for beginners.

We found that the EDC features can only support the GDPR for the PHT if proper administrative measures are taken. For example, GDPR requires that the data collection should only include the necessary information and achieve data minimization. The EDC system itself can only provide users with guidance to support such a policy. Therefore, only a combination of strict administration and supportive techniques can draw a comprehensive solution.

This is a preliminary research, and future work should be carried to confirm the findings here. First, the assessment of the systems is limited to binary by the comparison table. A more fine-grained scale of analysis and comparison of the features will be more appropriate. For example, the multi-factor login is more advanced than the SQL injection defense, so the systems with multi-factor login should be granted more emphasis while choosing the systems. Second, this research does not cover detailed instructions on how to develop the PHT approach with Castor. Future in-depth research should focus on Castor's Interoperability to figure out the feasibility of building PHT with Castor. This includes discovering the potential problems and estimating workload. A thorough understanding of PHT's development details and communication with Castor is necessary.

Moreover, this research provides the background of the principle to make data FAIR, and we found limited information on how EDC systems deal with FAIR. Therefore, a future in-depth research is required to investigate how the EDC systems cover each of the FAIR data principles. Because the way of making data FAIR shows in diverse ways, and some of them may be unexpected for researchers, future research should be conducted by a combination of literature review on possible methods to make data FAIR and evaluation with clinical use cases.

REFERENCES

[1] European Union General Data Protection Regulation, 2016,https://gdpr-info.eu/

[2] Wikipedia, GDPR https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

[3] REDcap, https://www.project-redcap.org/

[4] Castor, https://www.castoredc.com/

[5] ShareCRF, https://www.sharecrf.com/

[6] nowEDC, https://www.datatrial.com/edc/

[7] O. Beyan et al., "Distributed Analytics on Sensitive Medical Data: The Personal Health Train," Data Intell., vol. 2, no. 1–2, pp. 96–107, 2020, doi: 10.1162/dint_a_00032.

[8] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Sci. Data, vol. 3, pp. 1–9, 2016, doi: 10.1038/sdata.2016.18.

[9] R. Fleischmann, A. M. Decker, A. Kraft, K. Mai, and S. Schmidt, "Mobile electronic versus paper case report forms in

clinical trials: A randomized controlled trial," BMC Med. Res. Methodol., vol. 17, no. 1, pp. 1–10, 2017, doi: 10.1186/s12874-017-0429-y.

[10] L. O. Gostin and J. G. Hodge, "Personal privacy and common goods: a framework for balancing under the national health information privacy rule.," Minn. Law Rev., vol. 86, no. 6, 2002, doi: 10.2139/ssrn.346506.

[11] US DEPARTMENT OF LABOR, Guidance on the Protection of Personal Identifiable Information https://www.dol.gov/general/ppii

[12] Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule; Nass SJ, Levit LA, Gostin LO, editors. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington (DC): National Academies Press (US); 2009. 2, The Value and Importance of Health Information Privacy. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9579/

[13] P. Kerr, "Protecting patient information in an electronic age: a sacred trust.," Urol. Nurs. Off. J. Am. Urol. Assoc. Allied, vol. 29, no. 5, pp. 315–318, 2009.

[14] S. Landau, "Control use of data to protect privacy," Science (80-.)., vol. 347, no. 6221, pp. 504–506, 2015, doi: 10.1126/science.aaa4961.

[15] K. El Emam, S. Rodgers, and B. Malin, "Anonymising and sharing individual patient data," BMJ, vol. 350, 2015, doi: 10.1136/bmj.h1139.

[16] A. J. Rohm and G. R. Milne, "Just what the doctor ordered. The role of information sensitivity and trust in reducing medical information privacy concern," J. Bus. Res., vol. 57, no. 9, 2004, doi: 10.1016/S0148-2963(02)00345-4.

[17] Y. Chen, J. Guo, C. Li, and W. Ren, "FaDe: A blockchain-based fair data exchange scheme for big data sharing," Futur. Internet, vol. 11, no. 11, pp. 1–13, 2019, doi: 10.3390/fi11110225.

[18] M. Shabani, "Blockchain-based platforms for genomic data sharing: a decentralized approach in response to the governance problems?," J. Am. Med. Informatics Assoc., vol. 26, no. 1, pp. 76–80, 2019, doi: 10.1093/jamia/ocy149.

[19] A. K. Shrestha, J. Vassileva, and R. Deters, "A Blockchain Platform for User Data Sharing Ensuring User Control and Incentives," Front. Blockchain, vol. 3, no. October, pp. 1–22, 2020, doi: 10.3389/fbloc.2020.497985.

[20] B. Fecher, S. Friesike, and M. Hebing, "What drives academic data sharing?," PLoS One, vol. 10, no. 2, 2015, doi: 10.1371/journal.pone.0118053.

[21] Damschroder LJ, Pritts JL, Neblo MA, Kalarickal RJ, Creswell JW, Hayward RA. Patients, privacy and trust: Patients' willingness to allow researchers to access their medical records. Social Science & Medicine. 2007;64(1):223–235.

[22] M. A. De Carvalho Junior and P. Bandiera-Paiva, "Health Information System Role-Based Access Control Current Security Trends and Challenges," J. Healthc. Eng., vol. 2018, 2018, doi: 10.1155/2018/6510249.

[23] Electronic HealthRecords: Privacy,Confidentiality,and Security https://journalofethics.ama-assn.org/article/electronic-health-records-privacy-confidentiality-and-security/2012-09

[24] C. C. Ma, K. M. Kuo, and J. W. Alexander, "A survey-based study of factors that motivate nurses to protect the privacy of electronic medical records," BMC Med. Inform. Decis. Mak., vol. 16, no. 1, pp. 1–11, 2016, doi: 10.1186/s12911-016-0254-y.

[25] M. D. Wilkinson et al., "Comment: The FAIR Guiding Principles for scientific data management and stewardship," Sci. Data, vol. 3, pp. 1–9, 2016, doi: 10.1038/sdata.2016.18.

[26] WikiPedia, AuditTrail https://en.wikipedia.org/wiki/Audit_trail

[27] B. Duncan and M. Whittington, "Enhancing Cloud Security and Privacy: The Power and the Weakness of the Audit Trail," in Cloud Computing 2016: The Seventh International Conference on Cloud Computing, GRIDs, and Virtualization, 2016, no. April, pp. 125–130.

[28] National Research Council (US) Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure. For the Record Protecting Electronic Health Information. Washington (DC): National Academies Press (US); 1997. 3, Privacy and Security Concerns Regarding Electronic Health Information. Available from: https://www.ncbi.nlm.nih.gov/books/NBK233428/

[29] Wikipedia, Attributed-based access control, https://en.wikipedia.org/wiki/Attribute-based_access_control

[30] https://www.linkedin.com/pulse/abac-xacml-where-does-complexity-come-from-borja-roux-lorenzo/

[31]https://www.f5.com/services/resources/glossary/policy-enforcement

[32] R. D. Hertin and O. I. Al-Sanjary, "Performance of hospital information system in Malaysian public hospital: A review," Int. J. Eng. Technol., vol. 7, no. 4, pp. 24–28, 2018, doi: 10.14419/ijet.v7i4.11.20682.

[33] BBC, Evaluation of Usability https://www.bbc.co.uk/bitesize/guides/zrrc2sg/revision/6

[34] Github, PHT https://github.com/PersonalHealthTrain/PHT_FHIR_Demos_German_Initiative/tree/master/PHT_Demo_FHIR/BMI

[35] Global COVID-19 Clinical Platform Case Report Form (CRF) for Post COVID condition (Post COVID-19 CRF)

[36] R. Padmanaban, M. Thirumaran, P. Anitha, and A. Moshika, "Computability evaluation of RESTful API using Primitive Recursive Function," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2018, doi: 10.1016/j.jksuci.2018.11.014.

[37] M. Antonio da Silva and M. Danziger, "The Importance Of Security Requirements Elicitation And How To Do It," PMI® Global Congress 2015, 2015.

[38] K. Goertzel et al., "Software Security Assurance," Iatac, 2007.

[39] C.Bielaszka-DuVernay,04.Feb.2008. Is Experience Always a Good Thing?, Harvard Business Review, https://hbr.org/2008/02/is-experience-always-a-good-th

[40] D.Arts, 01.Oct.2018, Castor is committed to scalable FAIR Data, Castor, https://www.castoredc.com/blog/castor-is-committed-to-scalable-fair-data/

[41]Case Study of Castor EDC https://github.com/wzh6612/Castor-eCRF

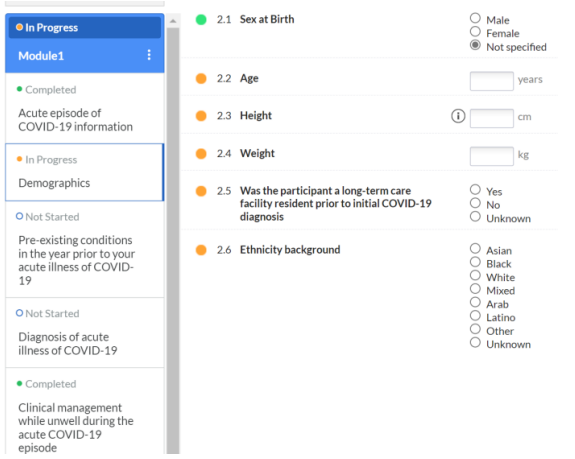
[42]REST API Security Essentials, REST API Tutorial https://restfulapi.net/security-essentials/

APPENDIX

A. Usability Testing

TIn this section, we interviewed three participants to comprehend their attitudes toward Castor and two of them work in hospitals. The three participants have different roles. There is a doctor, a hospital receptionist and a non-healthcare professional. By this, we hope to grasp a relatively comprehensive understanding of people's attitudes. As medical staff has experience recording patient's information, they can compare this experience with the way they entered data before. A participant without relative experience is set due to the knowledge gap between medical staff and non-medical staff members. Experience is not always a good thing, and sometimes it means narrow vision, stale ideas and tired habits of thought [39]. Also, in the hospital's daily routines, sometimes patients need to fill in forms themselves. That is another reason we set this non-healthcare professional role, and we expected to receive a review from unexpected perspectives.

The evaluation focused on Usability Testing, and the participants expressed their feelings about learnability, efficiency and error handling. It was designed like the following: The interviewee login the test account we registered for beforehand. We skipped the process of asking participants to register their accounts as this may cause their worries about their privacy. Data-Entry roles were assigned to the test accounts. Participants needed to fill in the form step by step.

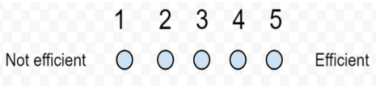


Pic 2. Demographics section of the eCRF to be filled by participants.

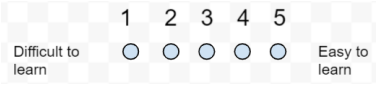
After this, participants were requested to try to create a simple eCRF themselves. As guidance, we printed the relevant tutorial for participants to follow. This included the introduction of the sections and steps to build an eCRF. In the end, we collected feedback from the participants with a survey form. The survey contains the following questions:

1. Do you have experience with data entry in an information system? If yes, have you ever faced any problems?

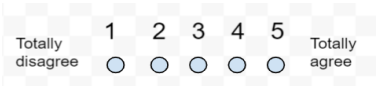
2. How efficient is the system?



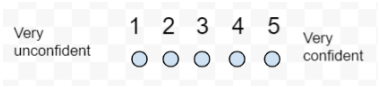
3. How easy is it to learn?



4. You find this system more convenient to record data than paper



5. How confident are you that you can build an eCRF only with the system's built-in guidance?



6. Which feature surprised you the most? Why is it?
7. Which feature caused you frustration?
8. Describe your overall attitude towards the system.
9. Possible improvements

3.4.2 *Result Analysis*

The table below is a quantitative analysis of the attitudes towards the efficiency and learnability of the system.

Table 4. Quantitative result of use-case experiment

	Doctor	Hospital receptionist	non-healthcare professional
Data-entry experience	Yes	Yes	Yes
Efficiency	5	4	3
Learnability	4	2	1
Q4	5	5	5
Q5	4	2	1

From this table, we can see that the doctor, who has experience with EDC systems, tends to give a high appraisal. Because the receptionist only has experience with simple information systems to register patients, this EDC system confused her due to its complicated structures. To give a clearer visualization of how the attitudes towards the system vary between different roles, we made the line chart below.

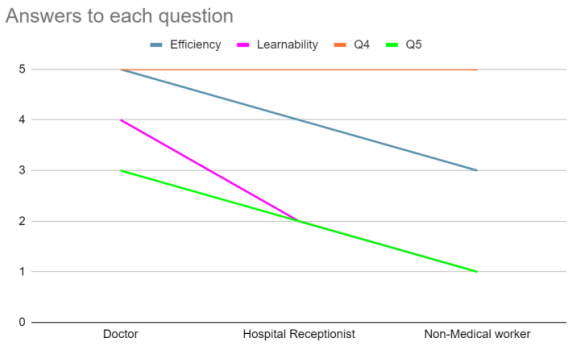


Chart 1. Flowchart to show the difference between roles

This chart illustrates how the satisfaction varies for each role. Except for data entry, we can see that the satisfaction of the system has a positive relevance with the relevant experiences that the participants have. However, all of them felt it was convenient to use Castor to fill the form.

Q5 is asking about how confident the participants are to build an eCRF without the guidance we provide. Without the guidance we provided, it means the participants can only explore the system with the built-in guidance. This series has nearly the same direction as learnability. This means the built-in guidance directly influences the learnability of a system.

Participants' answers to Q6, Q7, Q8 are listed in the following table:

	Doctor	Hospital receptionist	non-healthcare professional
Most surprising feature	Error-detection	Data-entry is convenient	Data-entry is convenient
Most frustrating feature	Cannot update the value of a submitted form	Question types are not named in an understandable way	The process to set up an eCRF is complicated
Overall attitude	It can improve the work efficiency	Not user-friendly	Easy to fill the form but hard to build one. Built-in guidance for beginners is not enough.

Table 5. Qualitative results of use-case experiment

The general attitude of the receptionist and the non-healthcare professional is that it is convenient to record patients' information with this system, but the building process is not concise. And their suggested improvements are also about reducing the redundancy in the eCRF building. In contrast, the doctor felt the system was convenient to use and it could improve his efficiency.

Due to the limited number of participants, we cannot conclude that this system is difficult for beginners to operate and useful for experienced users, but we get a first impression of how people with different backgrounds react to the system. Although the non-healthcare professional thought it was difficult to build an eCRF, it is enough that he felt it was convenient to fill in forms with the EDC system. In general, Castor is a useful tool for experienced medical staff. However, it needs to improve by adding more guidance for beginners.