Improving the Personal Health Train approach with Electronic Data Capture Systems

ZHIYONG ZHU, University of Twente, The Netherlands

The collecting of data is an essential component of clinical trials. Accuracy, timeliness, and consistency of data collection can significantly improve the quality of clinical trials and shorten the length of investigations. Electronic Data Collecting (EDC) is a remote electronic records management system for clinical trials that gathers clinical trial information remotely and directly from the trial center via the Internet, reducing time and improving the accuracy of the obtained data. The Personal Health Train (PHT) is a method to make a patient's data accessible to several organizations through a distributed database while maintaining patient privacy and data sovereignty. In the context of PHT, this project intends to provide a framework that enables distinct EDC systems to remotely access each other's data across international borders while maintaining data sovereignty. We evaluate this framework by analyzing metrics on how it adheres to the FAIR data principles, providing recommendations on the findings from this analysis. Then, we evaluate the framework in terms of GDPR for transnational data jurisdiction and discuss the legality of the entire system. Finally, we recommend a set of encryption techniques to enhance the security and anonymity of data flows.

Additional Key Words and Phrases: Electric Data Capture (EDC), Personal Health Train (PHT), Privacy, FAIR principles, GDPR.

1 INTRODUCTION

Early on, the majority of medical-related data, such as official medical records, prescription medicine records, X-ray records, and CT image records, were saved on paper rather than electronically. With the advent of advanced data storage, computing platforms, and mobile Internet, medical data has shown a tendency toward rapid electronic digitization. All of the above medical information is being digitized to varying degrees. Mobile Internet, big data, cloud computing and other technologies are being integrated with the medical field, new technologies and new service models are quickly penetrating into all aspects of medical care, making significant changes in the way people get medical care. For example, the Remote Data Entry (RDE) system, which was developed between the late 1980s and early 1990s, was EDC's pre[8].

EDC revolutionized the collecting of data for clinical research, radically altering the old method of data collection and the data management procedure. The typical EDC system includes not just a range of data collection functions, but also robust data querying capabilities. It also provides excellent inter-user communication solutions, allowing users of the same project to connect with EDC system data efficiently. Due to its multiple benefits, EDC has essentially replaced traditional paper case report forms in clinical trials conducted in developed nations such as the United States and Europe.

The research in this project focuses on EDC systems in the context of Personal Health Train. The Personal Health Train (PHT) is a distributed infrastructure that allows the flow of medical information between databases by accessing health record in different areas, thereby assisting healthcare professionals with data management, analysis, and medical decision-making.

The FAIR Principles are a set of guidelines for data sharing. FAIR means Findable, Accessible, Interoperable and Reusable. In this study, we give some solutions to coordinate multiple EDC systems in a PHT environment, while meeting the FAIR principles and GDPR recommendations.

2 BACKGROUND

We construct the knowledge module in this part. In section 2.1, we first discuss the concepts, features, and examples of EDC systems. We present the ideas of PHT, FAIR principles, GDPR, and CDISC in 2.2 Data Management. By connecting these ideas, we investigate how to conduct standardized, data-safe clinical trials while maintaining data sovereignty.

2.1 Electronic Data Capture systems

The traditional methodology of clinical trial data gathering relies on paper versions of medical records and afterwards requires the double-blind entry of paper information by data management employees, resulting in lengthy data collection cycles and delays in statistical analysis. With the rise of the Internet and computer technology, EDC systems have emerged, which are computer network-based systems that stress the direct capture and transfer of clinical data in electronic form via an integrated combination of software, hardware, standard operating procedures, and human resources[22]. According to the Pharmaceutical Research and Manufacturers of America and the Biometrics and Data Management Technical Group and the Clinical Trials EDC Working Group, "an EDC system is a data capture technology that transmits clinical trial data directly to the sponsor using electronic forms rather than paper forms" [25] . EDC systems collect clinical trial data using electronic case report form (eCRF) rather than paper case report form, thereby resolving the deficiencies of the old model.

The EDC system has the following advantages over traditional data collection methods: (1) real-time data entry, which can reduce data entry errors (2) data logical verification is performed simultaneously with data collection, i.e. the data entry system can detect protocol violations and out-of-range data in real-time (3) reduced collection cycle and guaranteed data traceability (4) accelerated clinical trial study progress and enhanced data quality[14, 30] .

Common EDC systems include: Castor, OpenClinica, REDCap, NowEDC and ShareCRF. A significant amount of study has been

TScIT 37, July 8, 2022, Enschede, The Netherlands

conducted on the EDC system itself, beginning with the introduction of the concept[11, 15, 21], comparing the EDC system with the traditional method[31], and examining the hazards of the EDC system itself[33]. Unfortunately, few viable remedies have been suggested.

2.2 Data management

2.2.1 Personal Health Train.

Due to legal and ethical restrictions, the majority of healthcare data cannot be extracted from hospitals or pilot facilities. PHT presents a distributed data management strategy that supports data sovereignty and prevents the export of an organization's data from its facilities. PHT enables bringing algorithms to data rather than data to algorithms.

The PHT approach is specifically founded on federated learning. Data can be highly distributed in real-world contexts, and individual data sources can be independent of one another. If we want to directly share data and combine models, privacy leakage is inevitable. Currently, the protection of data privacy has become a global hot topic, and as a result, governments throughout the world are strengthening their data security and privacy protection legislation. Legal rules unquestionably contribute to the creation of a safer society, but they also pose obstacles to clinical trials, which require vast and high-quality data, and there is an abundance of privacy-related data in hospitals, which might form data silos and hinder their development. How to legally resolve the problem of data distribution and isolation is a significant barrier for clinical trials under the concept of privacy protection. In this setting, federated learning [36] has been a popular area of research. Federated learning offers a workable answer to the problem of data stores by establishing a balance between privacy and efficiency, and it processes data through indirect data sharing. McMahan[19] and others provide the most traditional federation learning algorithm, namely (Federated Avgerage, FedAvg), which may federate data owners of all parties to execute algorithms while ensuring the privacy and security of client data.

PHT has three basic elements[4]:

Station: The station provides computational resources and integrates data from numerous data points, preserves the data, and executes duties in a secure environment, and for this reason, it evaluates the train's data queries.

Train: A component that executes an algorithm from multiple data sources and returns the results of the computation, and each train has a unique identifier.

Handler: Regulates train-station interaction, receiving trains and matching them with corresponding stations.

2.2.2 Principles of FAIR.

The draft FAIR principles were presented at the 2014 "Jointly Designing a Data Fairport"[2] conference in Leiden, the Netherlands, and formalized as scientific data management guidelines in 2016. FAIR, which stands for Findable, Accessible, Interoperable, and Reusable, has become a data management principle concentrating on the identification, traceability, sharing, and reuse of scientific data[16] . FAIR principles can be used to guide the management of medical and scientific data and to promote data sharing and reuse.

Findable: The primary prerequisite for scientific data sharing is that data can be discovered by users on time, so discoverability is the foundation for FAIR data, which provides the conditions for subsequent data access, manipulation, and reuse. An identifier is a series of characters that identifies the data, and is used to associate data names with resources through a discovery protocol. The association of data names with resources through search protocols can help users to locate and index data sets or meta locate and index datasets or metadata promptly.

Accessible: Once data is identified and discovered, it should be accessed through the services provided by the trusted repository, but there must be protocols related to accessing data resources clearly defined so that users know how to access the data, how to authenticate, how to obtain access, etc. The overall requirement of the accessibility principle is that data can be accessed not only by humans, but also by machines without any obstacle under the premise of following certain access protocols and having clearly defined authorization or authentication rules. The key words of the principle can be summarized as standardized protocols, free of charge, authorization, etc.

Interoperable: When multiple data resources are related to the same topic, users usually need to spend a lot of time to understand these data resources and think how to combine them. Therefore, interoperability is a necessary requirement for data integration, and only on the basis of interoperability can subsequent processes such as data analysis, storage and processing be performed. The overall requirement of the interoperability principle is to use standard definitions and common data elements to represent data and enable interoperability.

Reusable: The ultimate goal of data Findable, accessibility, and interoperability is to achieve extensive reuse of data resources. Data reuse not only reduces the time and financial cost of repeated data acquisition, but also verifies and enhances the reliability and reproducibility of the data, and allows users to investigate new scientific questions and make new scientific discoveries from these "secondary" data. The overall requirement of the reusability principle is that data and datasets have clear The overall requirement of the reusability principle is that the data and datasets have a clear license to be used and that accurate information about the source of the data is provided.

2.2.3 General Data Protection Regulation.

For the issue of privacy protection, the GDPR defines three most basic concepts, called data subjects, data controllers and data manipulators:

Data Subjects: The owner of private information is the data subject. Privacy information, as information and data that can be associated with a natural person, belongs to a natural person, regardless of whether it is gained directly from him/her or through analysis and processing. The consent of the data subject is required for the collection, analysis, processing, and storage of such private information, so long as it does not contradict legal requirements or compromise the public interest.

Data Controller: A data controller can be a person or an organization, and refers to the responsibility of deciding what to do with

sensitive data. What private information may be required for treatment and research, and how it is managed, are not established by the patients themselves and require particular knowledge and abilities in healthcare. The data controller is the role directly responsible to the data subject for deciding how patient privacy will be used, explaining clearly to the patient what privacy needs to be provided and what it will be used for, and obtaining informed consent from the patient as required by law, and for ensuring that the unit and research team only process and use the patient's private information to the extent that the patient (data subject) has consented.

Data Processors: The data processor, again, can be a person or an institution, not directly facing the patient (data subject), and regarding how to process and analyze the patient's private information, it is only necessary to execute the data controller's request, and as long as that request is not illegal, to follow the request and never to go beyond the scope of processing.

2.2.4 Clinical Data Interchange Standards Consortium.

Over the years, clinical trials have become increasingly electronic, as shown by the widespread usage of EDC systems. Data standards, particularly the Clinical Data Interchange Standards Consortium (CDISC) data standards, have been established to cover the full clinical trial process, to increase the efficiency and quality of drug clinical studies, and to streamline regulatory agency review[26] . CDISC data standards define clinical study data structure (metadata) and data validity values, such as data collection, storage, analysis, and submission, as well as data interchange standards.

CDSIC has established a series of standards for the exchange of clinical research data, the Clinical Data Acquisition Standards Harmonization (CDASH) identifies the basic data recording fields required from clinical, research, and regulations perspectives, making data collection in research centers more efficient and consistent. It defines a basic set of "highly recommended and suggested/conditional" data gathering fields in the early stages of clinical research. The Study Data Tabulation Model (SDTM) establishes the conditions for data submission, and the usage of CDASH data collection fields (or variables) makes mapping the SDTM structure easier[13]. The CDISC part of the standard is introduced as shown in Table 1.

Standard name	Abbreviations	Function
Study Data Tabulation Model	SDTM	Includes all standard vocabulary and coding sets addressed by the CDISC model/standard
Clinical Data Acquisition Standards Harmonization	CDASH	Content standard for the base data collection fields in the case report form
Operational Data Model	ODM	Content and format standards for capturing, exchanging, reporting or submitting, and archiving clinical study data based on case report forms
CDISC Terminology Dictionary		A dictionary of terms used to explain the electronic capture, exchange, and reporting of clinical research information terms and their definitions

Table 1. Introduction of some CDISC standards

3 PROBLEM STATEMENT

Several researches have been done about the functions of EDC systems, but there is a lack of research on the issues of data security and data standardization that may arise from interactions between different EDC systems. In this study, We provide solutions for data privacy and security issues in the PHT-based environment and evaluate these answers.

We can break the problem down into three questions:

(1). How to set up a PHT environment that has the participation of multiple organizations from different data jurisdictions, considering that each organization uses a different EDC software systems (Castor and OpenClinica)?

(2). How to apply the recommendations on GDPR in a such PHT environment?

(3). How the FAIR data principles can be addressed in such PHT?

4 SOLUTION DESIGN

In this section, we set up the framework in the PHT environment, in 4.1 we design and execute experiments, through 4.2 we analyze the FAIRness of the framework with metrics, 4.3 we establish data management guidelines, in 4.4 we analyze the authorship of data sovereignty in the context of GDPR, and we provide some legal suggestions for the framework in 4.5.

4.1 Experiment setup and execution

We conceived a cross-regional clinical trial, involving the Netherlands and the United States, to study the efficacy of pill X for the treatment of heart disease. Patients in both countries enrolled were randomized to the pill X group and the placebo group, on the basis of the heart disease medication they were receiving. Due to the long treatment period of heart disease, the experimenters needed to follow up the treatment for many years. To maximize data protection, the institutions in both countries have developed algorithms and intend to to run these algorithms with each other's data. The Dutch institution uses OpenClinica, a clinical research data management system, while the US institution uses Castor.

We deployed OpenClinica, an EDC system to simulate the Dutch clinical trial institution, on a local computer. OpenClinica's data stations are configured on AmazonWebServices (AWS). Similarly, the US institution utilized CastorEDC, with the exception that Castor did not require deployment, because it is accessible via the Internet. Meanwhile, Castor's data station is also deployed on AWS. We create EC2 instances on AWS and select the server node as the US.

The most basic connection between the two organizations is that they each run the data acquired by the other on their own algorithms. eCRF, implemented by EDC, is a significant tool for clinical research data collection and a substantial source of clinical research data[10]. In the PHT environment, institutions upload their respective eCRFs to their respective AWS s3 buckets; the filename of each eCRF file acts as its ID, and each institution has a different bucket in which to keep these files. When a copy of the eCRF must be retrieved, the script train is initiated, the eCRF number is searched. When the needed eCRF data is located, the query train will bring back the

query results to the bucket of your own institution, and the results can be downloaded from the data station.

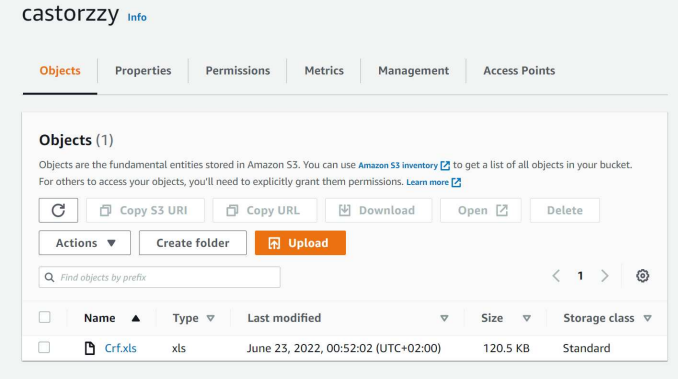

Fig. 1. eCRF data from the Netherlands accessed in the US in a PHT environment

4.2 FAIRness analysis

To assure the uniformity and generalizability of the indicator design, this study reviewed the assessment methods and assessment frameworks of a range of countries. Representative ones are the Go FAIR Metric Group's (GFMG) FAIR Common Metrics Framework[34] , the Dutch Data Archiving and Networked Services (DANS)[35] , the EU Horizon 2020[7, 27] and the Australian Research Data Commons' (ARDC)[3] FAIR principles assessment metrics, the Australian Common wealth Scientific and Industrial Research Organisation's (CSIRO) 's 5-star data assessment too[12] . CSIRO has 14 indicators, each with its own detailed interpretation.

Based on the common indicator framework designed by the GFMG group, we created the assessment framework for the FAIR application in the PHT environment (see Table 2). Under the four basic indicators of findable, accessibility, interoperability, and reusability, eleven secondary indicators and twelve tertiary indicators are used to evaluate the use of FAIR in a PHT setting.

For sample selection, five eCRFs generated by Castor were simulated in this study, and the descriptions corresponding to each indicator were evaluated for samples, with a check mark for compliance and a blank for non-compliance.

4.3 Guidelines for data management

Based on the above issues, we provide the following recommendations.

4.3.1 Clarify the access terms and user review mechanism.

To protect the intellectual property rights of medical science data, based on the findings of this study, an approval system should be implemented in both data stations to evaluate the user status and divide the authority levels, distinguishing between ordinary and advanced users, or public interest free use and commercial paid use users. It is advantageous for both institutions to manage their data uniformly for efficient use. The data stations should also allow human or automated review and evaluation, as well as the granting of access privileges to users, actively maintain users, and investigate

Level 1 Indicators	Level 2 Indicators	Level 3 Indicators	Explanation of indicators	
Findable	Identifier	Identifier Type	Whether the dataset uses a uniqueformal, persistent identifier	✓
	Resource Identification	Metadata inclusion identifier	Whether the dataset identifier is included in all metadata records/files describing the data	✓
	Indexed in searchable	Discoverable by search engines	Whether to register in the platform	✓
Accessible	Access Agreement	Access Registration	With or without access data user terms	
	Access authorization	User audit mechanism	Whether data access rights correspond to user registration status	✓
	Data lifetime	Data lifetime commitment	Does it guarantee to provide stable storage	
Interoperable	Language	Data file format	Whether to use the common machine-readable standard format representation	✓
	FAIRized vocabulary	Description of data elements	Whether data elements are associated with standard vocabularies	
	Qualified references	Status of data association	Data is associated with other data in a way that explicitly and contextually enhances the relationship	✓
Reusable	License Statement	Clarity of the license statement	Whether the conditions of reusability are clearly expressed	
		Clarity of the restriction statement	Whether the condition of non-reusability is clearly expressed	
	Traceability specification	Clarity of traceability information	Whether to describe the traceability information of the data	

Table 2. FAIR application evaluation metrics system based on PHT environment

the development of various technical support mechanisms for user review. The terms and conditions should be clearly stated.

4.3.2 Ensure the long-term utilization of metadata.

Due to the timeliness, privacy, and quality concerns, data may be blocked from open sharing. Both EDC systems must publish their storage commitments for data stability, for the long-term use of data and for the traceability of medical science datasets. At the end of the dataset lifecycle, inform users of data unavailability and provide information on data context, creator, and creating institution through metadata.

4.3.3 Reference to standard vocabularies and data exchange standards.

Although both institutions use electronic medical records, the standards for data entry may not be equivalent. Without standardization and uniformity in terms of format and semantics, it is difficult to merge the vast volume of real-time clinical medical record data,

limiting its maximum utilization. Faced with the challenge of clarifying the semantics of data sets from various sources, the transferred data should support the CDSIC standard.

The eCRF fields are defined directly using the CDASH data standard, while SDTM defines the data submission standard. Employing CDASH data collection fields (or variables) allows the transformation of data into a standard SDTM format. In the CDASH data standard, the Demographics (DM) and Subject Characteristics (SC) fields are used to define basic subject information such as name, age, sex, and date of birth, and the AE event field is used to define adverse occurrences[13] . Taking the adverse event collection field as an example, some of the fields correspond to CDASH and SDTM variables as shown in Table 3.

Collection Data	CDASH Variable Name	SDTM variable	Description
Adverse event number	aespid	AESPID	Record the unique identifier for each adverse event for the subject.
Adverse event name	aeterm	AETERM	Name of the adverse event.
Date of start of adverse event	aestdat	AESTDTC	Time when the adverse event started.
Adverse event end date	aeendat	AEENDTC	Time when the adverse event was resolved.
Severity of adverse event	aesev	AESEV	Describe the severity of the adverse event.
Is the adverse event related to a congenital anomaly or birth defect?	aescong	AESCONG	Record whether the "serious" adverse event was related to a congenital anomaly or birth defect.
Is the adverse event related to the study treatment?	aerel	AEREL	The clinician/investigator determines whether there is a causal relationship between the study treatment and the adverse event.
Outcome of adverse event?	aeout	AEOUT	Describe the status of the subject(s) associated with the adverse event.

Table 3. Correspondence between the adverse event section fields and the CDASH and SDTM variables

Other data table fields that cannot be found in the SDTM or CDASH standards, can be designed according to the SDTM standard. In addition, the design of the data dictionary is based on the CDISC Terminology (CT) standard [9], for example, the gender is set to male and female, and the blinding modes are single-blind, double-blind and non-blind.

4.3.4 *Permission to follow explicit dataset reuse conditions.*

Uncertain access permission or reuse statements of datasets will prevent users from using their own clinical data in a reasonable and lawful manner. Consequently, the EDC platform must define the scope of its authority, which can refer to the international standard reuse statement license, and can add additional terms such as data reuse and citation format statement in the protocol used.

4.3.5 *Adopt standard traceability format.*

Traceability information facilitates the evaluation of the application of clinical data at another institution. Accurate and extensive machine-readable traceability information can provide researchers with credentials and support for evaluating datasets. According to the conclusions of this study, Castor and OpenClinica in the PHT environment lacked machine-readable traceability information, and the production process of the dataset, the creator, the data generation device, and the data processing process should be extensively documented.

4.4 GDPR analysis

We discuss the legal aspects of the framework. The classification of three roles(data subjects, data controllers, data processors) clearly identifies the source of the validity of the privacy work, namely the patient's consent (data subject). Second, the person accountable for decision management and ensuring that the patient's privacy is fully respected and protected is identified, i.e., the data controller. Lastly, for the data processor, the responsibility boundary is also specifically defined from the perspective of depriving him/her of decision-making authority, i.e. he/she is only responsible for the execution and does not need to consider external privacy as long as it is not illegal, maximizing efficiency.

The data controllers and processors have different specific identities based on various scenarios, e.g., at the time of the visit, it could be the attending physician who, as controller, informs the patient about which hospital activities may involve privacy, whereas other departments such as laboratory, examination, and information must not exceed the commitment of the attending physician to the patient. In this clinical trial scenario, the study leader is the controller and is responsible for deciding what patient information is collected and how the study will be conducted, while the subject members are expected to strictly adhere to the study leader's requirements and refrain from any misuse and sharing beyond their requirements. In such a multi-center, collaborative study, the patient's hospital is the controller, and the other hospitals can only handle patient privacy according to the patient's hospital's clear instructions.

4.5 Guidelines for data protection legislation

We have observed a few legal issues in the framework and have provided some guidance.

4.5.1 *Not providing patients with copies of their data.*

Under the GDPR, citizens have the right to seek a copy of their personal healthcare data from a healthcare provider [20]. The copy they receive must be aggregated, widely accessible, and machine-readable. The right to receive a copy increases the data subject's control over his or her personal information. Consequently, according to this provision, patients in the medical field can access their personal health care data and request a copy of their personal health care data at the appropriate medical institution whenever they need access to their personal health care data, regardless of whether they maintain a paper or electronic record of their health care data at that time. In addition to preventing wasteful duplication of examinations and conserving medical resources, citizens' access to their own health care records can allow them to obtain copies of the information. Therefore, the patient role should be added to both

the Castor and OpenClinica systems so that patients always have access to their own data.

4.5.2 GDPR and U.S. law intersect.

In fact, due to the intricacy of the cross-regional data issue, the GDPR itself discusses whether the use of data in different circumstances is subject to the GDPR. We analyze this study as an example.

1. If the data collected by the Dutch institution is entrusted to a U.S. institution for processing, the Dutch institution is the data controller and the U.S. institution is the data processor. The controller is required to enter into a contract with the processor to ensure that the processor processes the data in accordance with the GDPR [28]. To explain the responsibilities of the U.S. processor, a data processing agreement is required between the Dutch data controller and the American processor. The GDPR applies to both the data controller and the processor in this instance.

2. If the U.S. institution entrusts a Dutch institution with the processing of its data, where the U.S. institution is the data controller and the Dutch institution is the data processor, then the Dutch institution, i.e. the processor, will be subject to Article 3(1) of the GDPR. This does not imply, however, that the U.S. organization is equally subject to Article 3(1) of the GDPR. In other words, an out-of-EU data controller's selection of an in-EU data processor to process data on its behalf does not automatically subject the controller to the GDPR [28]. Moreover, a processor's "establishment" in the EU cannot be considered the controller's "establishment" in the EU just because the controller has delegated the processing of data to the processor, thereby making the controller liable to the GDPR.

3. The data controller, i.e., the patient, is located outside the EU. In this case, it is also determined whether the data processing act involves the provision of goods, services, and surveillance in the EU. If so, the data processing activity performed by the data processor in relation to the targeting falls under the GDPR[29] . The Focus should be placed on the significance of the data processing acts done by the data processor to the controller's goal-directed operations[5] . In this instance, data from U.S. patients are subject to the GDPR since they are compared and combined with data from European patients and eventually influence the outcome of this clinical research involving the Netherlands.

The Clarifying Lawful Overseas Use of Data Act (CLOUD Act)[23] is the primary data protection law in the United States. It grants US authorities the authority to obtain data kept abroad from service providers subject to US jurisdiction. The European Data Protection Board (EDPB) believes that the United States government may evade the Agreement on Mutual Legal Assistance between the European Union and the United States (MLAT), which is now in existence between the European Union and the United States[1] . So that EU data subjects are forced to disclose personal data subject to the GDPR at the request of U.S. law enforcement or authorize service providers to intercept, disclose, or listen to the content of their wire or electronic communications in real time at the request of a foreign government if that foreign government has entered into an administrative agreement with the United States, indicating that the Cloud Act has extraterritorial jurisdictional effect[24] . Consequently, if processors and controllers of patient data are subject to GDPR or EU member state regulations, they may encounter a conflict between U.S. law and GDPR or other EU or member state laws, placing data subjects in a difficult position. Before beginning clinical trials, contracts should be signed and the government should be consulted.

4.5.3 Lack of information on data processors and controllers.

In this PHT system, the transnational processing of data necessitates the notification of the international data controller. According to the GDPR, when a controller collects data from a data subject, it must disclose the identity and contact information of the representative to the data subject. If a data controller outside the EU fails to tell the data subject of the identity of its representative, it violates the GDPR's transparency requirement[6] . On the official website or clinical trial sheets, the contact information of the responsible party should be listed.

4.5.4 Lack of assurance of data anonymity and transmission security.

The GDPR mandates that data controllers (processors) in EU member states be evaluated for the transfer of personal data to third countries in order to ensure that personal data moved outside the EU is secured to the same extent as personal data inside the EU. This necessitates both the security of the data transfer procedure and the anonymity of the data itself. Since data transmission in trains is exposed to the network in the PHT environment, an encryption system and a concealment system must be developed.

Each eCRF image can generate a unique 128-bit hexadecimal string of numbers (abstract) using the MD5 algorithm, and the abstract generated by the MD5 algorithm is identical each time for the same image. However, even a small change in the electronic medical record will result in a substantial change in the MD5 value. In this approach, as the eCRF is being created, an abstract is produced for each record, which is comparable to generating a "fingerprint" for each medical record to identify the medical record's legitimacy and any modifications made to it. The "fingerprint" will change dramatically if the medical record is altered.

Using a combination of encryption and concealment techniques, the encrypted eCRF is first encrypted using Henon map to produce the encrypted eCRF picture X. The image holding the case record number and the MD5 value of the eCRF is then chosen as the carrier image Y. The size of the carrier picture can vary. The carrier picture Y is scaled to match the size of the confidential eCRF, and then the encrypted eCRF is embedded in the carrier image using the LSB technique. The carrier image contained in the eCRF is the composite image Z. The illegal person does not know if this carrier information conceals other information, and even if they did, it would be difficult to extract or erase the hidden information[17]. After the hidden carrier reaches the receiver by PHT, the receiver utilizes the key to recover or identify the hidden secret information from it[18] , and then recovers the original confidential eCRF image from the synthetic image.

The eCRF image encryption and hiding scheme proposed in this paper has the following features. 1. each eCRF is generated with a "fingerprint" by MD5 algorithm, which can be used to prevent the record from being illegally modified. 2. The image with the MD5 value of the medical record is used as the carrier image, and the chaotic encrypted electronic medical record is hidden in the

carrier image, which is transmitted as a composite image in the transmission, effectively protecting the privacy of the eCRF.

5 DISCUSSION

In a PHT scenario, it is entirely feasible to construct a clinical study involving various national organizations. This naturally throws a substantial burden on data storage and processing. It requires a robust cloud server, a standard data registry, numerous legal contracts and notifications, a method to encrypt and conceal the data, etc. Access to data within EDC systems is controlled by roles and permissions. In clinical trials, the duties and permissions for each function are relatively consistent. The system should record the access history of users, including the access time, IP address, role, and modification status. If the system detects a cluster of unauthorized access, it must notify the system administrator to guarantee data security.

But even if the software staff does a great job, the most significant factor is the system operator's awareness of operating standards. The effectiveness of the process is dependent on the skill of the employees utilizing the resources and tools, and the usage of the EDC system demands training by qualified specialists, while the system itself should provide a testing environment. The system must give staff with extensive and detailed operational manuals, an evaluation mechanism to decide which staff pass the test and which require additional training, and a rapid query feature for frequently asked questions to speed the learning process.

Even with the EDC system and PHT environment, paper forms may still be required in clinical trials. Therefore, the format of the form and the layout of the questions should ease the recording and entry of data by physicians. After the first draft of the form design has been finished, the clinical doctors, data managers, and statisticians could meet to review the form's logic and then finalize it through changes. Prior to the implementation of the clinical trial, a copy of the instructions for completing the form must be prepared to ensure that it is accurately filled out. Each column of the form is explained separately. This includes each variable, the description of the variable, the type (number, character, date), the length (how many digits) and the number of decimal places of the number, the range of values (e.g., age 18-35) and the skip rule (e.g., if no medication was taken, the following medication name can be skipped and the next question can be entered directly) as well as the variable's coding (e.g., side effects 1 - none, 2 - yes).

In many instances, patient privacy information collected by a research project may be accessed by multiple institutions or stored in a repository and reanalyzed by subsequent researchers, a process that frequently results in the loss of control over the information once it leaves the initial research team. This is a severe problem in traditional privacy protection, especially for electronic material, whose content is replicable and may be fully uncontrollable once it leaves the first controller. The GDPR specifies in greater detail how patient privacy information should be legally, securely, and effectively shared between different research institutions, who should inform patients of this sharing, and how to secure personal privacy data during and after research. In the case of cross-institutional collaborative research, based on the definitions of data controllers and data processors, strict constraints can be imposed through agreements between controllers and processors, which impose control requirements on the receivers and processors of patient privacy data through contractual means and bind their behavior by law. Any processing beyond the boundaries of the agreement, and if the partnering institution persists in doing so after the data controller has written it down and created undesirable effects, can be disputed by legal means on behalf of the institution and the patient. However, the disagreement and conflict between the laws of the two countries is difficult to resolve and requires additional confirmation from the government.

In addition, for many medical and scientific researchers who are concerned that too much privacy protection will have a negative impact on their research, the GDPR provides corresponding additional provisions, regulating how to use personal privacy information in these areas, through a series of management and technical measures designed to reduce the adverse impact on individuals, while taking into account the advancement of medical research. For instance, valuable scientific data can still be saved and reused after wiping information that can be used to identify specific individuals, so that the FAIR principles can also be met.

6 CONCLUSION

The fact that all data is stored in a single "data center" makes extensive sharing and monitoring difficult. PHT, the distributed data management approach, enables the decentralized storing of medical records and allows hospitals to process data without remote access to it. In addition, PHT can make medical records more "visible," but this "transparency" is confined to the sharing units and not the entire society and network. Utilizing PHT approach, each node of medical records sharing, i.e., hospitals and health service centers, may process data more efficiently, thereby decreasing the expense of communication and time in medical records sharing. Moreover, as medical record file sharing is limited to each node, its security will also be enhanced.

The presented investigated a specific scenario of the Personal Health Train involving two countries, We simulated a clinical trial in the Netherlands and the United States. The first part of the study consisted in preparing the simulation scenario and deploy them in the two participant-countries, we built the PHT environment, and decentralized the data management using AWS, Castor, and OpenClinica. We implemented the PHT approach to enable the algorithm to access the data stations of both countries while simultaneously ensuring data sovereignty. In addition, we evaluated this system by dividing the knowledge of FAIR into eleven secondary and twelve tertiary measures. Furthermore, we analyze the PHT system from a legal perspective by evaluating the three data roles of the GDPR and discussing the legal constraints on data subjects in different environments by situation.

Our results show that it is feasible to build a framework for data processing between two countries iin the PHT environment. However the FAIR principle is not well satisfied, and the problem is focused on the difficulty of data access by data subjects and the reuse of the data. To adhere to the FAIR principles, user access terms and user review procedures must be implemented to ensure that

patients have access to their data, and the data storage cycle must be provided to enable data traceability. In addition, we propose a set of CDISC data standards applicable to global clinical research. By standardizing everything from collecting standards to variable name to data submission standards, we expect to decrease the issues related to data merging caused by disparate EDC systems.

Furthermore, the results of the study also showed that the frame was missing a copy of the patient's personal health data, while the system does not have a patient role. Contact information for data processors and data controllers was also absent from the system, making it more difficult for the patient to obtain the data. By comparing the relevant U.S. data laws with the GDPR, we discovered a degree of incompatibility, which required that project organizers consult with both governments before moving further. To maintain the security of data transfer and anonymity of data, we created a combined encryption and chaos system that satisfies the applicable GDPR criteria via MD5 encryption and Henon map.

This study builds on our earlier results[32] and has been progressed by building a basic PHT environment and performing cross-regional data access. Analyzing the limitations of this study, first, the train component is only implemented in the script train, which searches for specific data between S3 buckets and without implementing a queuing scenario at the data stations and without deep processing of the data. In addition, the amount of data is quite small, all of the data is manually entered by us, and since we are not involved in real clinical data processing, we cannot evaluate the framework's performance. The data processing criteria have also not been applied. These limitations will be investigated in future work. Our future work should build on this research, and the PHT approach should be extended to investigate how to maintain stable operations when several trains and stations interact, as well as to validate the stability and efficiency of PHT with huge data volumes using real data or in collaboration with real companies. In addition, for FAIRness assessment of data, appropriate countermeasures should be taken and solutions should be tried to be implemented to identify problems in the implementation phase.

REFERENCES

[1] Michael Abbell. 2010. Agreement On Mutual Legal Assistance Between The European Union And The United States. In *Obtaining Evidence Abroad in Criminal Cases 2010*. Brill Nijhoff, 403–420.

[2] M Axton, A Baak, N Blomberg, JW Boiten, LB da Silva Santos, PE Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3 (2016), 15.

[3] Michelle Barker, Ross Wilkinson, and Andrew Treloar. 2019. The Australian research data commons. *Data science journal* 18, 1 (2019).

[4] Oya Beyan, Ananya Choudhury, Johan Van Soest, Oliver Kohlbacher, Lukas Zimmermann, Holger Stenzhorn, Md Rezaul Karim, Michel Dumontier, Stefan Decker, Luiz Olavo Bonino da Silva Santos, et al. 2020. Distributed analytics on sensitive medical data: the personal health train. *Data Intelligence* 2, 1-2 (2020), 96–107.

[5] Maja Brkan. 2016. Data protection and conflict-of-laws: A challenging relationship. *Eur. Data Prot. L. Rev.* 2 (2016), 324.

[6] Niamh Clarke, Gillian Vale, Emer P Reeves, Mary Kirwan, David Smith, Michael Farrell, Gerard Hurl, and Noel G McElvaney. 2019. GDPR: an impediment to research? *Irish Journal of Medical Science (1971-)* 188, 4 (2019), 1129–1135.

[7] European Commission. 2016. Guidelines on fair data management in horizon 2020. *Tech. Rep.* (2016).

[8] Khaled El Emam, Elizabeth Jonker, Margaret Sampson, Karmela Krleža-Jerić, Angelica Neisa, et al. 2009. The use of electronic data capture tools in clinical trials: Web-survey of 259 Canadian trials. *Journal of medical Internet research* 11, 1 (2009), e1120.

[9] Rhonda Facile, Erin Elizabeth Muhlbradt, Mengchun Gong, Qingna Li, Vaishali Popat, Frank Pétavy, Ronald Cornet, Yaoping Ruan, Daisuke Koide, Toshiki I Saito, et al. 2022. Use of Clinical Data Interchange Standards Consortium (CDISC) standards for real-world data: expert perspectives from a qualitative Delphi survey. *JMIR medical informatics* 10, 1 (2022), e30363.

[10] Li G., Li X. Yan, and Wen Z. Huai. 2014. Application of the Clinical Data Interchange Standards Association standards in the design of case report forms for clinical studies in Chinese medicine. *Journal of Guangzhou University of Traditional Chinese Medicine* 31, 1 (2014), 138–141.

[11] Paul A Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G Conde. 2009. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics* 42, 2 (2009), 377–381.

[12] Ali Hasnain and Dietrich Rebholz-Schuhmann. 2018. Assessing FAIR data principles against the 5-star open data principles. In *European Semantic Web Conference*. Springer, 469–477.

[13] Vojtech Huser, Chandan Sastry, Matthew Breymaier, Asma Idriss, and James J Cimino. 2015. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). *Journal of biomedical informatics* 57 (2015), 88–99.

[14] Wen Jingran and Zhang Xiaoyan. 2010. *Application and prospect of electronic data acquisition system in clinical trials*. Ph. D. Dissertation.

[15] Irene Katzan, Micheal Speck, Chris Dopler, John Urchek, Kay Bielawski, Cheryl Dunphy, Lara Jehi, Charles Bae, and Alandra Parchman. 2011. The Knowledge Program: an innovative, comprehensive electronic data capture system and warehouse. In *AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 683.

[16] Rebecca Daniels Kush, D Warzel, Maura A Kush, Alexander Sherman, Eileen A Navarro, R Fitzmartin, Frank Pétavy, Jose Galvez, Lauren B Becnel, FL Zhou, et al. 2020. FAIR data sharing: the roles of common data elements and harmonization. *Journal of Biomedical Informatics* 107 (2020), 103421.

[17] Qi Li, Xin Liao, GQ Qu, Guo-yong CHEN, and Jiao DU. 2016. Adaptive steganography algorithm in digital image based on Arnold transform. *J. Commun* 37, 6 (2016), 192–198.

[18] Ting Liang and Peng XU. 2013. Image steganography algorithm based on human visual system and nonsubsampled contourlet transform. *Journal of Computer Applications* 33, 01 (2013), 153.

[19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[20] Robert Merrick and Suzanne Ryan. 2019. Data privacy governance in the age of GDPR. *Risk Management* 66, 3 (2019), 38–43.

[21] Emily F Patridge and Tania P Bardyn. 2018. Research electronic data capture (REDCap). *Journal of the Medical Library Association: JMLA* 106, 1 (2018), 142.

[22] Ivan Pavlović, Tomaž Kern, and Damijan Miklavčič. 2009. Comparison of paper-based and electronic data collection process in clinical trials: costs simulation study. *Contemporary clinical trials* 30, 4 (2009), 300–316.

[23] Miranda Rutherford. 2019. The CLOUD Act: Creating Executive Branch Monopoly over Cross-Border Data Access. *Berkeley Tech. LJ* 34 (2019), 1177.

[24] Jessica Shurson. 2020. Data protection and law enforcement access to digital evidence: resolving the reciprocal conflicts between EU and US law. *International journal of law and information technology* 28, 2 (2020), 167–184.

[25] Stephen A Sonstein, Jonathan Seltzer, Rebecca Li, Honorio Silva, C Thomas Jones, and Esther Daemen. 2014. Moving from compliance to competency: a harmonized core competency framework for the clinical research professional. *Clinical Researcher* 28, 3 (2014), 17–23.

[26] Tammy Souza, Rebecca Kush, and Julie P Evans. 2007. Global clinical data interchange standards are here! *Drug discovery today* 12, 3-4 (2007), 174–181.

[27] Daniel Spichtinger and Jarkko Siren. 2017. The development of research data management policies in Horizon 2020. *Research Data Management-A European Perspective* (2017), 11–23.

[28] Anni-Maria Taka. 2017. Cross-Border Application of EU's General Data Protection Regulation (GDPR)-A private international law study on third state implications.

[29] W Gregory Voss. 2017. First the GDPR, now the proposed ePrivacy regulation. *Journal of Internet Law* 21, 1 (2017), 3–11.

[30] Brigitte Walther, Safayet Hossin, John Townend, Neil Abernethy, David Parker, and David Jeffries. 2011. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. *PloS one* 6, 9 (2011), e25348.

[31] Brigitte Walther, Safayet Hossin, John Townend, Neil Abernethy, David Parker, and David Jeffries. 2011. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. *PloS one* 6, 9 (2011), e25348.

[32] Zihan Wang, Wallace Ugulino, and João Luiz Rebelo Moreira. 2021. An information security diagnostic of Electronic Data Capture Systems for the Personal Health Train. In *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*. IEEE, 216–225.

[33] James A Welker. 2007. Implementation of electronic data capture systems: barriers and solutions. *Contemporary clinical trials* 28, 3 (2007), 329–336.
[34] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.
[35] Mark D Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for FAIRness. *Scientific data* 5, 1 (2018), 1–4.
[36] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.