systems

MDPI

Article

Health Data Sharing towards Knowledge Creation

Luís B. Elvas [1,2,*], João C. Ferreira [1,2], Miguel Sales Dias [1] and Luís Brás Rosário [3]

[1] ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal; jcafa@iscte-iul.pt (J.C.F.); miguel.dias@iscte-iul.pt (M.S.D.)
[2] Inov Inesc Inovação—Instituto de Novas Tecnologias, 1000-029 Lisbon, Portugal
[3] Faculty of Medicine, Lisbon University, Hospital Santa Maria/CHULN, CCUL, 1649-028 Lisbon, Portugal; lsrosario@medicina.ulisboa.pt
* Correspondence: luis.elvas@iscte-iul.pt

Abstract: Data sharing and service reuse in the health sector pose significant privacy and security challenges. The European Commission recognizes health data as a unique and cost-effective resource for research, while the OECD emphasizes the need for privacy-protecting data governance systems. In this paper, we propose a novel approach to health data access in a hospital environment, leveraging homomorphic encryption to ensure privacy and secure sharing of medical data among healthcare entities. Our framework establishes a secure environment that enforces GDPR adoption. We present an Information Sharing Infrastructure (ISI) framework that seamlessly integrates artificial intelligence (AI) capabilities for data analysis. Through our implementation, we demonstrate the ease of applying AI algorithms to treated health data within the ISI environment. Evaluating machine learning models, we achieve high accuracies of 96.88% with logistic regression and 97.62% with random forest. To address privacy concerns, our framework incorporates Data Sharing Agreements (DSAs). Data producers and consumers (prosumers) have the flexibility to express their prefearences for sharing and analytics operations. Data-centric policy enforcement mechanisms ensure compliance and privacy preservation. In summary, our comprehensive framework combines homomorphic encryption, secure data sharing, and AI-driven analytics. By fostering collaboration and knowledge creation in a secure environment, our approach contributes to the advancement of medical research and improves healthcare outcomes. A real case application was implemented between Portuguese hospitals and universities for this data sharing.

Keywords: information sharing; artificial intelligence; data sharing agreement; electronic health records; security; homomorphic encryption

Citation: Elvas, L.B.; Ferreira, J.C.; Dias, M.S.; Rosário, L.B. Health Data Sharing towards Knowledge Creation. *Systems* **2023**, *11*, 435. https://doi.org/10.3390/systems11080435

Academic Editor: Colette Rolland

Received: 12 July 2023
Revised: 15 August 2023
Accepted: 18 August 2023
Published: 21 August 2023

1. Introduction

Healthcare data integration is a crucial research topic for optimizing the healthcare sector [1], as accurate diagnoses and prognoses are vital for proper decision making and, consequently, fundamental for ensuring an appropriate clinical approach. Nevertheless, the integration process addresses complex and multifaceted challenges, such as safeguarding patient privacy and managing health data from multiple information systems. In this context, information sharing can present significant challenges with respect to data security and privacy, as concerns regarding trust and interoperability among institutions may arise [2].

Within this paradigm, prior to the integration of data from prosumers (i.e., producers of data who are also consumers), an effective methodology should be adopted to mitigate these trust-related issues among institutions.

While blockchain technology has the potential to represent a viable approach to mitigate these concerns, it is important to acknowledge that if not implemented and managed properly, blockchain-based technology can compromise data safeguards [3].

Alternatively, we suggest the adoption of Data Sharing Agreements (DSAs) as a potential solution to address these trust-related issues [4]. DSAs are mutual agreements between two or more parties that establish regulations for sharing and managing data, including privacy preferences and contractual requirements such as notification in case of data leakage. By clarifying each party's duties, DSAs help ensure data sharing reliability.

Given this context, a Federation of prosumers is created to collectively manage and share data in a controlled and regulated manner. Subsequently, the DSAs are established to govern the usage of information among prosumers, securing a controlled and restricted environment. The term 'Federation' in this context refers to a collaborative group of data prosumers, including healthcare providers, researchers, and relevant stakeholders, who come together to collectively manage and share data. This group is established to address trust-related issues and ensure responsible data sharing practices through mutual agreements (DSAs).

Because it is vital to ensure that the data is shared and processed in a secure and confidential manner, a virtual layer—Information Sharing Infrastructure (ISI)—assumes the responsibility of managing and collecting data from the Federation. It consists of an Artificial Intelligence Module (AIM) that operates on top of the shared and integrated data.

Once the DSAs have been established, data obtained from various sources are integrated into a centralized database. While centralizing data provides unified structured data and the optimization of operational healthcare processes, sharing data with a trusted analytics server may not always hold, and can result in potential breaches of privacy-preserving collaborative data.

To address this paradigm, an Advanced Encryption Standard (AES) algorithm has been employed in the literature [5], which involves transforming data into an unintelligible format and protecting/safeguarding its content with a secret key that only authorized parties can use to decrypt it. Furthermore, to enable operations on encrypted data without requiring its decryption, homomorphic encryption [6] has been applied to machine leaning models such as logistic regression and random forest. This approach enables these models to perform secure computations on encrypted data without requiring access to the decrypted data. These machine learning models provide predictions concerning the patient's susceptibility to specific diseases using their encrypted health data. As a result, an email alert is automatically sent to the patient's healthcare entity following every prediction based on the data received within the last hour, enabling this entity to respond appropriately to address potential critical health situations. Since these predictions rely on encrypted data, confidentiality is preserved throughout the computation process.

As mentioned, given that information sharing is a major concern in the healthcare industry, our research approach addresses the following key components:

- Data sharing: sharing of information in a controlled manner, including sensitive health data. This ensures regulatory compliance, confidentiality and integrity both while in rest and in transit.
- Artificial intelligence: usage of AI algorithms to classify and predict episodes that require immediate attention, triggering an email alert to notify the corresponding healthcare entity that action needs to be taken.
- Multi-technology: usage of a combination of technologies to enable a confidential and collaborative approach to data analysis, including homomorphic encryption. This allows computation to be carried out in a private and distributed manner.
- Streamlined access: implementation of advanced seamless access mechanisms which take advantage of the analytics and sharing infrastructure to provide continuous authentication, authorization, and privacy awareness, for privacy-aware data usage control.

For this work, a real case implementation of current DSA was established, allowing us to gain access to the data of their departments comprising 512,764 patients. In order to comply with ethical guidelines, we obtained informed consent from participants, following the principles outlined in the Declaration of Helsinki and the Oviedo Convention [7]. Furthermore, we ensured the necessary documentation, including a data dictionary, autho-

rization from the CHULN services in Cardiology, Intensive Care Medicine, and Respiratory Intensive Care Unit, as well as the CVs of the respective physicians in charge. Additionally, all members with data access signed a declaration of honor, guaranteeing adherence to GDPR regulations, which encompassed protecting sensitive information, specifying authorized personnel, defining data retention periods, establishing data disposal procedures, and preventing unauthorized utilization in other research contexts without explicit consent.

The work is divided into five sections: (1) the introduction; (2) state of the art, where we identify the current literature work status; (3) description of our proposed framework; (4) an application case in which the CRISP-DM (Cross-Industry Standard Process for Data Mining) data mining approach is adopted [8] since we are dealing with data knowledge extraction; and (5) the conclusions.

2. State of the Art

2.1. Search Strategy and Inclusion Criteria

We undertook a systematic review of the literature using the PRISMA methodology (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [9]. Our investigation was restricted to articles written in English and published between 2016 and 2023, sourced from databases such as Scopus and Web of Science Core Collection (WoSCC), with grey literature, conference papers, workshops, books, and editorials being excluded.

To refine our search, we employed a specific search query using the domain of "Artificial Intelligence" or "Data Analytics", within the concept of "Health Data" and using the context of "Data Sharing". This approach enabled us to identify the relevant articles across both databases. Nevertheless, it is important to acknowledge that this approach may lead to the retrieval of some duplicated articles in our research results, which had to be removed.

2.2. Study Selection

The selection of articles was initially conducted on an assessment of titles and abstracts; however, in those cases where this information was insufficient or not clear, the full articles were reviewed.

2.3. Data Extraction and Synthesis

Zotero, Microsoft Excel, and the web interfaces to Scopus and WoSCC were used as tools for organizing and storing data related to the articles, according to the mentioned systematic literature review using the PRISMA method. This information included diverse categories such as the title, author, year of publication, subject area, keywords, and abstract. Furthermore, to facilitate data examination, a qualitative evaluation was conducted based on these topics.

2.4. Results

Table 1 provides details of the search criteria employed on the domain, concept, context, and limitations.

Table 1. Research conducted using Scopus and WoSCC.

Domain	Concept	Context	Limitations
"Artificial Intelligence" "Data Analytics"	"Health Data"	"Data Sharing"	2016–2023 Only journal papers Articles and reviews
281,382 Documents			
14,435 Documents			
200 Documents			

The query returned a total of 200 documents (as shown in Figure 1) by utilizing the keywords from each column. Subsequently, a manual review of each article was conducted to identify relevant research subjects and eliminate duplicates. This resulted in a set of 18 documents. Additionally, our systematic literature review methodology incorporated several parameters, such as the year of publication, research field, and the brief description of each article.

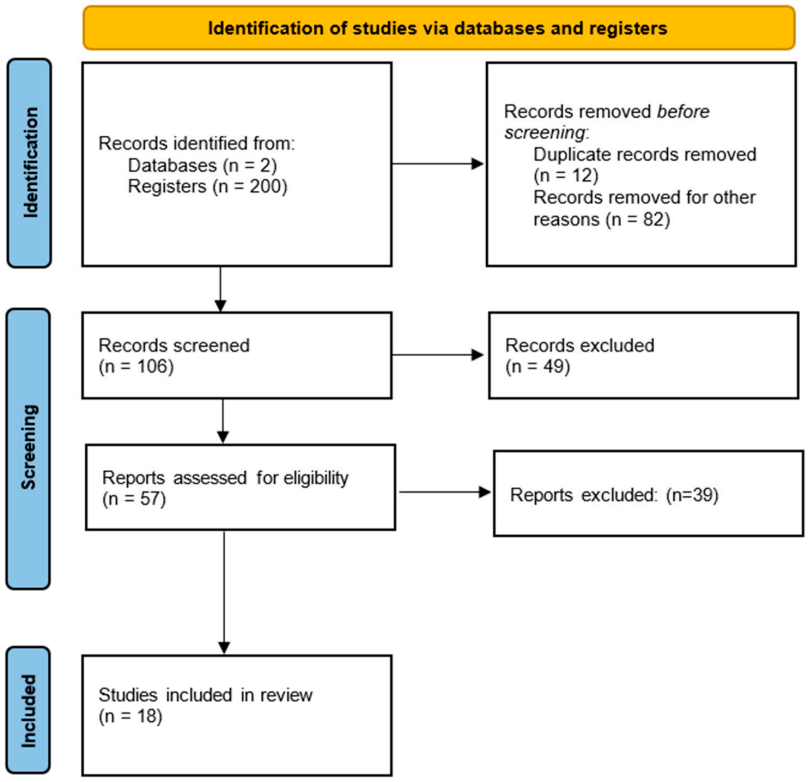

Figure 1. PRISMA workflow for systematic literature review adapted from [9].

2.5. Study Characteristics

All 18 studies included in our literature review met the search criteria mentioned above.

As illustrated by Figure 2, the research topic being addressed has gained more relevance in recent years.

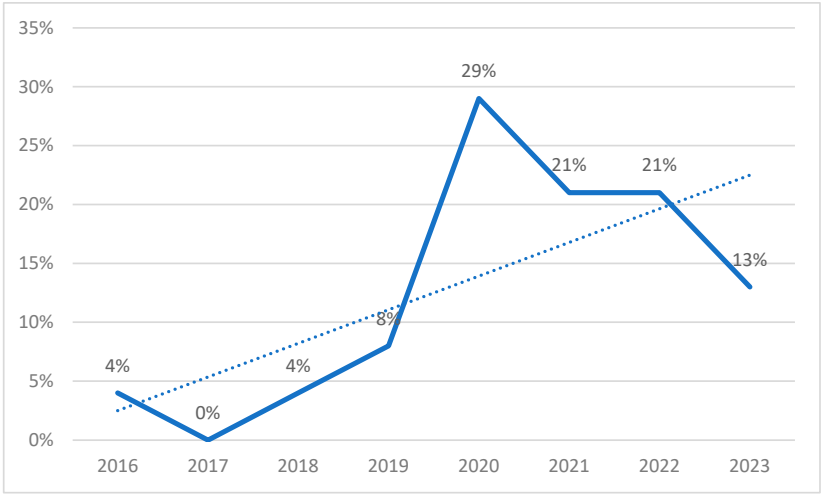

Figure 2. Percentage of articles published over the past 7 years: growth of articles in the last 7 years (2023 shows articles considered until the date of this review).

2.6. Goals and Outcomes Analysis

Within the selected and reviewed literature found to be pertinent to our study, data sharing emerged as the most recurrent subject of discussion (see Table 2). As a highly relevant topic in recent years (see Figure 2), our research focuses on the application and implementation of specific technologies to establish a robust data sharing ecosystem, mitigating the predominant concerns addressed in various articles. Given the sensitive nature of the information involved in healthcare, there are some challenges related to its sharing, including data protection and privacy issues. Within this context, articles [10–12], highlight the challenges encountered in complying with data regulations in the healthcare industry, particularly the General Data Protection Regulation (GDPR). While the GDPR is crucial to ensure privacy and personal data protection, articles [10,13] identify a disparity between its targets and its actual impacts, emphasizing the need for a trust-based framework in health data to mitigate the encountered bias. Additionally, while the authors in article [12] address the difficulties experienced by stakeholders in managing health data during the COVID-19 pandemic due to the GDPR requirements, article [14] discusses the positive impact of recent modifications in data sharing regulations in selected states of the USA on the quality of healthcare service.

Table 2. Detailed topics from articles.

Topic	Reference	% of Papers
Data Sharing	[10–21]	29%
AI/Deep Learning	[10–12,15–18,22,23]	22%
Data Privacy	[12,19,24]	7%
Data Governance	[13,25,26]	7%
Ethics	[13,19,22]	7%
Blockchain	[17,20,26]	7%
Big Data	[11,14,23]	7%
Data Protection	[10,25]	5%
Machine Learning	[10]	2%
Homomorphic Encryption	[27]	2%
Cloud	[27]	2%

The healthcare industry has seen a significant increase in the utilization of big data and AI in recent years, addressing several challenges and concerns. In this context, articles [19,22], highlight the importance of responsible data governance to ensure ethical research practices, and authors in [22] underscore the importance of liability in clinical application algorithms to ensure patient safety. Furthermore, in article [19], authors emphasize the risks of bias in the increasing volume of available data. From a patient´s perspective, article [16] presents a cross-sectional survey that revealed varying perceptions among patients and concluded that more public awareness and debate are necessary to ensure the acceptability of sharing their personal data. In order to address the challenges of data sharing, authors in [25] provide recommendations on establishing and managing data trusts to ensure that data sharing is conducted in a timely, fair, secure, and equitable manner.

On the other hand, articles [15,17,18,23] discuss the potential advantages of incorporating AI to standardize data sharing practices across various domains. Authors in [17,18] suggest that such approach could result in more efficient data management, improve clinical decision making, facilitate the use of supportive diagnostic tools for patient-centered treatment planning, and provide algorithms to analyze collected data. In [15,23], authors also suggest that AI could address ethical and data protection challenges related to health data sharing. Likewise, article [21] discusses how the adoption of near real-time electronic health record (EHR)-based surveillance systems and the integration of data analytics infrastructure proved instrumental for policymakers and epidemiologists in Iran during the COVID-19 pandemic.

Additionally, authors in [24] discuss how they employed an automated AI-based anonymization technique based on two heuristic principles, emphasizing privacy protection

through anonymization. The findings validate that their approach offered a more effective solution in comparison to alternative techniques.

With the increasing availability of data, the risk of data breaches has become a critical matter. In this context, articles [20,26] suggest the importance of adopting blockchain-based technologies to ensure data security and privacy. While authors in [20] propose adopting this technology for secure data exchange among healthcare entities, authors in [26] draw attention to a misalignment between the availability and the implementation of advanced technologies for medical purposes, suggesting the need for a new blockchain framework to manage big data while simultaneously promoting real-time access, data security, and patient privacy.

To ensure a secure environment for computational tasks involving health data, in [27], the authors propose a secure health cloud framework to facilitate the sharing of EHRs. This framework employs homomorphic encryption algorithms, which enable computations to be performed on encrypted data without the need for its decryption. The authors demonstrated that this approach is more efficient than conventional algorithms. Furthermore, the authors underlined several suggestions for future works, such as the application of statistical methods and machine learning algorithms for disease prediction, as well as exploration of bootstrapping techniques to enable full homomorphism.

In order to effectively manage and exchange clinical data while preserving confidentiality, the authors of [11] propose a big data analytics strategy coupled with the utilization of an AI (deep learning) algorithm for analyzing patient data and producing reports for stakeholders. The authors also recognize the need for additional research to address security concerns and suggest integration with a cloud platform to ensure scalability in the future.

In contrast to the studies found in the literature, our research introduces a novel and integrated approach to address the challenges of data sharing and privacy concerns in the healthcare industry. While previous works have explored individual aspects such as data governance, AI applications, and data anonymization techniques, our paper uniquely combines multiple technologies to create a comprehensive and secure data sharing ecosystem. The primary novelty lies in our utilization of DSAs to regulate data sharing practices responsibly, coupled with the adoption of homomorphic encryption algorithms for secure implementation of AI models. This combination ensures that data remains confidential and encrypted throughout the computational process, further safeguarding patient privacy. Additionally, our research showcases practical implementation through real-world use case validation, demonstrating the superiority and efficacy of our method in improving healthcare service quality, patient safety, and research practices. By presenting a cohesive and innovative solution, our paper contributes to advancing the field of data sharing and AI applications in healthcare, setting a new standard for secure and ethical data-driven strategies in the industry.

Beyond healthcare data sharing, which is an area where there is lack of access to data, it is interesting to explore its implementation in engineering tasks, with a particular focus on 3D applications [28,29]. This extension showcases the adaptability of our approach in diverse domains and highlights its potential to address data sharing challenges beyond the healthcare sector. By exploring the application of our method in different fields, we aim to present the broader implications and appeal of our research to a wider audience of readers and researchers.

3. A Framework for Information Sharing Infrastructure (ISI)

This chapter discusses a proposed platform that aims to facilitate data sharing among multiple prosumers by integrating data through Data Sharing Agreements (DSAs) based on agreed terms. The DSAs specify which data can be used, for what purposes, and how they can be used, and aim to capture the data sharing policies that restrict both suppliers and consumers of data while governing the flow of data between them.

To ensure trust, privacy, and compliance with GDPR, the platform uses homomorphic encryption. Prosumers define the DSAs at the time of Federation creation, based on their interests. DSAs govern the storage of prosumers' data and express constraints on shared data, such as obligations to process data before or after the data's usage, anonymize data, or perform homomorphic encryption operations.

The proposed Information Sharing Infrastructure (ISI) facilitates data sharing, ensuring continuous enforcement of policies and obligations related to the data. The ISI consists of an Artificial Intelligence Module (AIM) that operates on top of the shared and integrated data. The AIM executes the manipulation operations specified in the DSAs related to the AIM before making classifications or predictions.

This chapter further discusses the workflow involved in the creation of the proposed ISI. The ISI is a virtual layer that is deployed when Information Prosumers form a Federation by defining their DSAs to share their information. The ISI manages the Federation's information by collecting data from the prosumers and enforcing the DSA paired with the information, before the AIM executes AI operations. In these, shared data will be applied to machine learning algorithms in Python with associated keys to support decryption. Results are computed and distributed back to the Information Prosumers, with enforced DSAs, ensuring that confidentiality and privacy requirements are respected. The ISI's main components are the DSA enforcement engine and the data-protected object store, where data is encrypted at rest and stored with appropriate usage policies.

Figure 3 shows the logic architecture of the ISI, composed of a "DSA Usage Control System for Data", which represents an integral part of the overall "Information Sharing Infrastructure Diagram" that facilitates secure and controlled data access. The system consists of several interconnected components. At the core of the system is the Policy Decision Point (PDP), which receives data access requests and evaluates the relevant policies to determine whether access should be granted or denied. The PDP relies on contextual information to make informed decisions. This contextual information is provided by the Context Handler, which gathers details such as the user's role, time of access, and location. The Session Manager handles the management of user sessions within the DSA system. It handles tasks such as authentication and session termination, ensuring secure and authorized access to the data. In response to the access decision made by the PDP, the Obligation Manager enforces any obligations or actions that need to be performed. This can include tasks such as logging access events, generating audit reports, or executing specific actions based on the access request. To assist in the policy evaluation process, the PDP relies on Policy Information Points (PIPs). These PIPs serve as information sources that provide additional attributes or contextual information required for policy evaluation. In the diagram, three PIP boxes are shown, representing different sources of policy information that may include external systems, databases, or services. Additionally, the system incorporates "External Attributes" that are obtained from external sources. These attributes provide supplementary information that enhances the context for access control decisions. External attributes can be fetched from external databases, APIs, or other systems to make more informed policy evaluations.

The "DSA Usage Control System for Data" diagram is connected to the database and the anonymization toolbox within the broader "Information Sharing Infrastructure Diagram." This connection signifies that the usage control system governs access to the data stored in the database and ensures that the anonymization toolbox adheres to the established access policies. Together, these interconnected components and connections form a robust framework for enforcing access control, policy evaluation, and data protection within the Information Sharing Infrastructure.

The chapter also discusses how the proposed approach enables Information Prosumers to selectively share their information with a specific subset of members within the Federation. They can also perform pre- or post-processing manipulation operations on their information and apply AI algorithms. Moreover, Information Prosumers can disclose the analysis results only to certain Information Prosumers and under specific conditions.

Figure 4 shows that the ISI serves as a data source for the AIM, providing the necessary input for the AI algorithms and models to process. The ISI consists of various components and connections that enable secure and controlled data access. Within the ISI, the data required for the AIM are obtained from multiple sources. This includes structured data from patients, such as physiological information, which is processed using a machine learning toolbox specifically designed for structured data analysis. This toolbox employs statistical correlation techniques and machine learning algorithms to extract meaningful insights from the structured data.

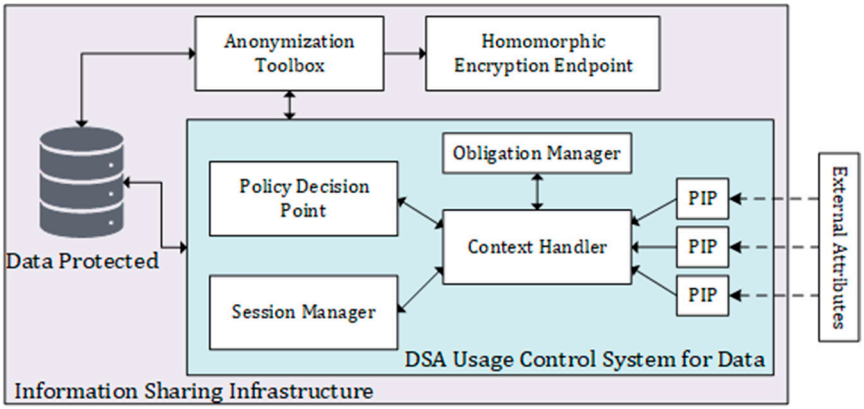

Figure 3. Proposed Information Sharing Infrastructure (ISI).

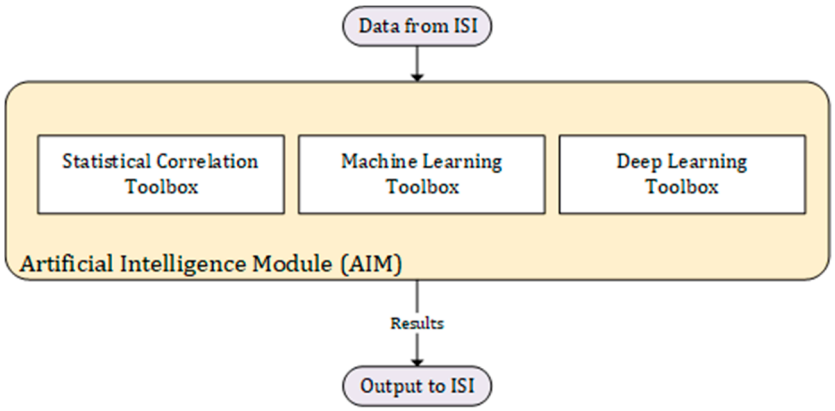

Figure 4. Artificial Intelligence Module (AIM).

In addition to structured data, the ISI also incorporates image data such as MRIs, CT-scans, and echocardiographies. To process these image data effectively, a deep learning toolbox is utilized. Deep learning algorithms specifically designed for image analysis are employed to extract features and patterns from the images, enabling the AIM to make accurate image classification and perform tasks related to image interpretation.

The outputs generated by the statistical correlation toolbox, machine learning toolbox, and deep learning toolbox collectively form the results of the AIM module. These outputs may include predictions, classifications, feature representations, or any other relevant insights derived from the data. Finally, the output from the AIM module is sent back to the ISI. This allows the results to be integrated back into the broader system, enabling further analysis, decision making, or sharing with authorized parties within the ISI.

By connecting the AIM module to the ISI, the system leverages the power of artificial intelligence and machine learning techniques to extract valuable information and knowledge from the data available within the Information Sharing Infrastructure. This integration enables advanced data analysis, predictive modeling, and image interpretation, ultimately enhancing the overall capabilities and potential benefits of the system.

The goal of the data sharing platform is to enable the creation of the proposed ISI, which involves a four-step workflow: (1) identification of data sharing needs and elaboration of DSAs; (2) secure data sharing through the Information Prosumer encrypting their data and sending it to the ISI; (3) manipulation operations specified in the DSAs being executed on the data by the AIM before making predictions; and (4) results returned to all Federation members who can take appropriate actions.

Figure 5 illustrates an example with four prosumers (i.e., a hospital, a clinic, a research institute, and a home care center), where all data are sent to a server and encrypted using the AES encryption algorithm. Each prosumer has a unique key, allowing them to encrypt and decrypt their data.

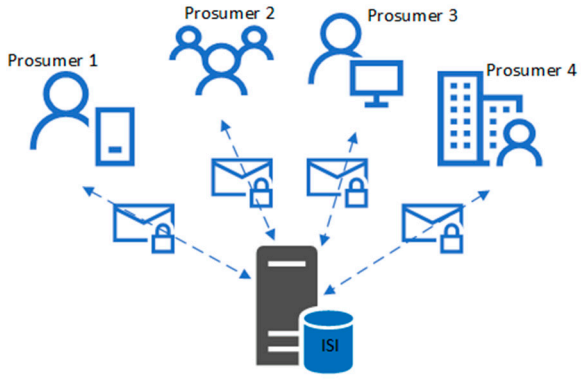

Figure 5. Information Sharing Infrastructure, with a central database to store health-encrypted data.

Our logical model with integrated data includes a MySQL database where each file is loaded into a separate table with the data being encrypted using the Advanced Encryption Standard (AES). Our MySQL database engine was able to properly handle the loading of data records with different encodings.

AES is a specification for the encryption of electronic data established by the US National Institute of Standards and Technology (NIST) in 2001 [30]. Encrypting turns the data into "human-unreadable" text referred to as cipher text instead of plaintext, which means the data is in its original form.

The algorithm can use keys of 128, 192, or 256 bits to encrypt and decrypt data in groups of 128 bits of data called blocks. This means it takes 128 bits as input and outputs 128 bits of encrypted cipher text as output. AES relies on the substitution–permutation network principle. This means it performs a few rounds, including substituting and shuffling the input data. The key size defines the number of rounds being 10, 12, or 14 for 128, 192, or 256 bits, respectively. Other authors have identified AES as a good encryption mode for homomorphic encryption.

Homomorphic encryption (HE) is a form of encryption that allows performing computations over encrypted data without access to the secret key. The result of such a computation remains encrypted. HE enables cloud services to process users' data without compromising privacy or security. HE can also be used as part of a secure multi-party computation protocol.

The current encryption algorithms force the data to be decrypted prior to processing it. This, however, means that data privacy laws are not complied with. Furthermore, data become less defended against unauthorized access. By enabling the computing of encrypted data, HE assures that data privacy and integrity are kept while the data are processed and ensures an extra layer of data protection. A fully homomorphic encryption (FHE) algorithm allows unlimited ciphertext operations while producing a valid result.

Figure 6 depicts where some of the risks lie and why FHE helps mitigate those risks.

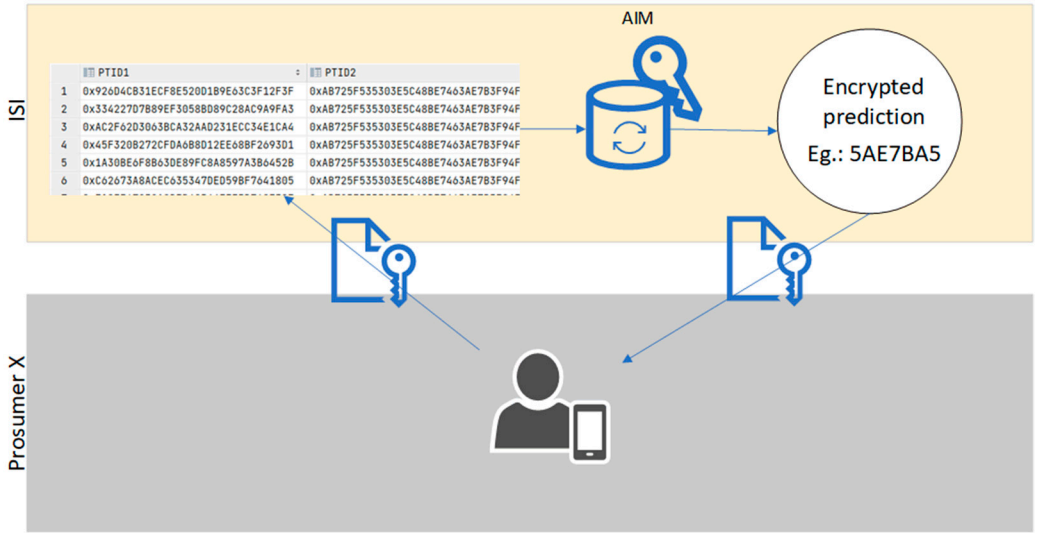

Figure 6. Data flow, from after the data are sent from the prosumer until it receives the output.

There might be a security threat between the prosumer's computer and the cloud service (server for AI processing) if the data is in plaintext while in transit.

Having the data encrypted locally, at the prosumer's side, before sending it to the cloud service and decrypting upon receiving the prediction means that it is secure while in transit within the network, thus limiting security risks.

In other words, predictions may be performed on cloud services without compromising data privacy, which means that medical records are not exposed to unauthorized parties.

This approach has been developed based on data that were made accessible by a Portuguese hospital from Lisbon and originated from five entities (prosumers). To obtain these data, DSAs were signed among all entities. The description of all the work and methodology that followed is depicted in Section 4.

4. Use Case Validation

Our work can be applied to several cases, but we validate it with an encrypted multi-syndrome dataset of clinical data collected at Hospital Santa Maria, the largest Portuguese public hospital, located in Lisbon. Health data were collected under the framework of the FCT project DSAIPA/AI/0122/2020 AIMHealth—Mobile Applications Based on Artificial Intelligence, co-coordinated by two of the authors, aiming to contribute with a preventive approach for public health strategies in facing the COVID-19 pandemic situation. The access to the dataset for research purposes was approved by the Ethical Committee of the Faculty of Medicine of Lisbon, one of the project partners. The dataset is currently being accessed by the authors (belonging to the ISTAR research center), Iscte, and the Faculty of Medicine researchers, under a DSA and within the scope of the mentioned FCT AIMHealth project. Nonetheless, and because this is a real implementation case, DSAs were implemented between Hospital Santa Maria (HSM), Faculty of Medicine of Lisbon, ISTAR, and Iscte.

In this section, we describe this application's work using a CRISP-DM methodology [31]. CRISP-DM is well-suited for addressing real-world data mining challenges and provides a structured approach for data mining projects. It is comprised of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [31]. This methodology allows a comprehensive and iterative analysis of the data and requires close collaboration between data scientists and domain experts. We successfully addressed a specific data mining problem using CRISP-DM and delivered actionable insights to the organization. To support this approach, we developed the ISI described in Figure 3.

4.1. Business and Data Understanding

This is the phase where data was first accessed, with business and variable understanding. Data relate to 512,764 patients and contain real-time clinical signals such as temperature, blood oxygen level (SpO_2), and heart rate. Data were extracted from a number of different information systems and encrypted at the hospital before delivery. The schema of this dataset includes 138 tables (from an identical number of files) and occupies around 75 Gbyte of data. We loaded the collected data into our secure local database taking into account the DSA. It was extracted and transferred to a secure Iscte server and encrypted, and was then ready for the application of ML algorithms.

The significant number of files (138) reflected, in part, the dispersion of hospital databases with 1594 variables, some with identical meanings. Their content was formatted as comma-separated values (CSVs), where all the data belong to the different prosumers. This is a real-life scenario in some hospitals and, in this case, within the department of cardiology, involving several small independent databases that lack interoperability. We found that some of the source hospital databases in production were somehow loosely coupled with the clinical processes and workflow. This resulted in tables with multiple fields that are not filled, and important clinical data that are not properly organized or structured but are instead introduced in free text. This posed difficulties in organizing and analyzing each variable in the context of overwhelming amounts of information and fragmented data across the 138 tables.

Most of the files (116 out of 138) are relatively small, with less than 100 thousand records. For instance, the file containing the types of precautions has 17 lines (i.e., intoxications, infections, etc.). However, the remaining 22 files are comparatively large, having anywhere between 280 thousand and 68 million records.

During our data understanding analysis, we identified several important variables, including gender, blood group, birthdate, and ethnicity of the patients. In addition, we found 52 different types of diagnosis, ranging from circulatory system illnesses to infectious and parasitic diseases and various pediatric-related diagnoses. Real-time data were identified in one specific table containing 657 real-time data variables, including systolic blood pressure, mean arterial pressure, and aortic pulse rate.

Because we aimed to predict whether patients will suffer from specific events (abnormal values of physiological variables), we had the valuable assistance of the mentioned cardiologist specialist to help to focus our analysis on the most significant variables. Considering his business knowledge, we considered 85 of 657 real-time measures available in the table. These 85 variables included patient height, aortic pulse rate, blood pressure, and heart rate, which were considered the independent variables used to predict the dependent variable (diagnoses as seen in the "admission diagnosis" table).

4.2. Data Preparation

In this phase, data were prepared for data fusion and encryption. Data in our dataset were made available in CSV (comma-separated values) text format. Each line in a CSV file is equivalent to a record, with the variables (columns) being separated by a comma. This method of distinguishing variables may create problems with descriptive values that often include commas within them (for example, open text reports). Furthermore, in Portugal, the decimal point in numbers is not a point, but a comma. Such cases were an extensive issue while analyzing the structure and content of some files, as it became difficult to identify the commas that represent column separations. To address this, a Python script was developed in-house and utilized to automatically replace the problematic commas with semicolons.

As mentioned, due to the existence of several information systems in production at the hospital, which were the source of our collected dataset, relevant data are scattered, not integrated, and sometimes duplicated across our various tables, leading to added challenges while analyzing, understanding, and integrating data.

The critical tasks of data cleaning, record deduplication, and data integration were performed, over the mentioned 6-month period, on the premises of the hospital based on the signed DSA. Confidentiality agreements do not allow us to describe the process.

The output of this first stage was a set of 138 clean files, which were loaded and populated our MySQL database in an encrypted form. Whilst the number of files is the same as that of the raw dataset, their contents were, at this stage, cleaned, reduced in terms of their number of records, and able to be integrated.

The created metadata that allowed an easy understanding of the data were also important. Since the data were encrypted, these metadata were fundamental for researchers to know which variables to use for the prediction or other data mining processes. Data were then ready to be used by machine learning (ML) algorithms.

In this section, we exposed some weaknesses of having data dispersed in multiple, isolated databases, in the context of the various hospital information systems, often in different formats. While each individual information system may fulfil its intended purpose, such segregation makes a full overview extremely difficult to accomplish.

The work developed for this section—by loading the raw data into the same format in a single, clean set of records, without duplicates, without unnecessary extra data, with consistent metadata, and while needing considerable amount of work—sets the stage for new insights to be gained, new analysis to be undertaken, new knowledge to be created, and new conclusions to be drawn. This is especially important in the context of hospital health as it might help save lives.

4.3. Modeling and Evaluation

In this section, we describe how this approach can be used, with an example. Since the data are available, it is possible to use and create knowledge. In this case, since we have the mentioned data from the COVID lockdown period, the goal was the detection of abnormal patient data.

Detecting abnormal values in clinical data can be challenging and requires expert knowledge and experience. In cardiology, the ability to predict abnormal values of specific variables can provide valuable insights into patient outcomes and disease progression. In this use case, we developed a machine learning model that can accurately predict abnormal values (outliers) of specific variables in cardiology data. Moreover, an email alert system was developed to notify health practitioners every time the model predicts an outlier value, facilitating real-time patient data tracking and prompt interventions when necessary. The following section describes how the abnormal value prediction and email alert generation were made, the results, and their implications for cardiology investigation.

For the modeling phase, we employed a supervised machine learning approach to predict outliers in patient data, specifically focusing on variables such as oxygen saturation, pulse, and heart rate identified by the cardiologist specialist. In this context, we collect all the vital signs from a medical machine that monitors the physiological signs of the patients every 5 min. For feature selection, we built a function that automatically calculates the Pearson correlation between the dependent variable (y—physiological variable chosen by the physician) and all the other variables selected by the physician, such as blood pressure, respiratory rate, age, height, jugular saturation of O_2, esophageal temperature, ST segments, room temperature, body temperature, invasive blood pressure—diastolic and systolic, hemoglobin, arterial O_2 saturation, pH, pulse, and heart rate. Once the Pearson correlation is calculated, only the variables with a correlation outside the interval of $]-0.2;0.2[$ are chosen to integrate the set of independent variables. To achieve a more accurate result, the autocorrelation of the dependent variable is also calculated.

Heart rate was selected as a dependent variable in this paper to showcase the proposed solution's effectiveness and provide a practical example of its implementation.

We created an autocorrelation plot to identify a threshold autocorrelation for detecting outliers in heart rate. The plot revealed an autocorrelation value of 0.7 seven hours before an abnormal value was registered, as illustrated in Figure 7. This suggests that heart rate

data showing an autocorrelation of 0.7 or higher seven hours before an event could be a useful predictor of future outliers. The autocorrelation in Figure 7. was measured as the correlation between the heart rate variable and its lagged values within the same time series. Autocorrelation quantifies the degree of correlation between a variable and its past values at different time lags. Temporal dependence of the heart rate data can impact the model's performance. High autocorrelation may suggest a strong temporal relationship between heart rate values at different time points, implying that the current heart rate value might be dependent on its past values. In predictive modeling, autocorrelation can influence the accuracy of the predictions. In this way, by using past values, we can make future predictions.

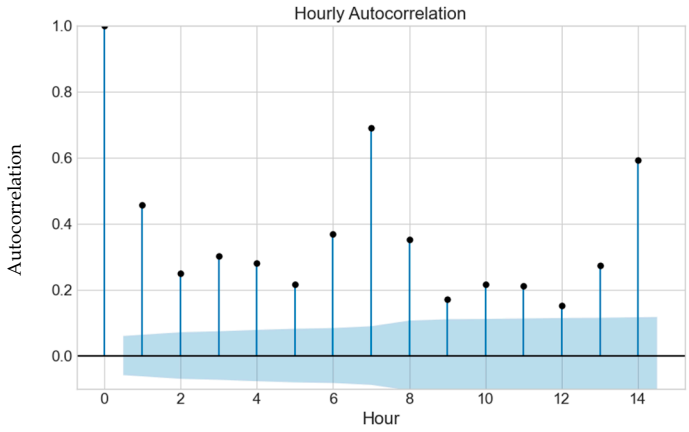

Figure 7. Heart rate autocorrelation.

While attempting to predict an outlier using encrypted data, we compared different algorithms for AI predictions, such as logistic regression (1) and random forest (2). The random forest achieved the best results.

$$f(enc(x)) = \frac{1}{1 + e^{-enc(x)}} \tag{1}$$

where *enc x* is the encrypted data to which the logistic regression is applied.

$$\sum_{i=1}^{C} -f_i \cdot \log(f_i) \tag{2}$$

f is the frequency of label i at a node and C is the number of unique labels.

The data were split into two groups: one for training the algorithms, consisting of 75% of all data, and another for testing the models, with the remaining 25%. After training the algorithms on the training data, we utilized the trained models to predict outcomes on the test data. The logistic regression model achieved an accuracy of 96.88%, and the random forest model outperformed it slightly, attaining an accuracy of 97.62%. These accuracies were calculated by comparing the model's predictions to the actual outcomes in the test data, ensuring a comprehensive evaluation of the predictive performance.

Our approach was trained and evaluated using appropriate techniques, including preprocessing steps and hyperparameter tuning. Regarding hyperparameter tuning, several variations of random forest were tested, namely by varying the number of estimators from 1 to 15. All fifteen variations produced comparable results, the difference being smaller than 0.01 percent. Once an outlier was predicted, we emailed the relevant entities to alert them that the patient needed to be observed because an event was predicted in 7 h; see the example in Figure 8.

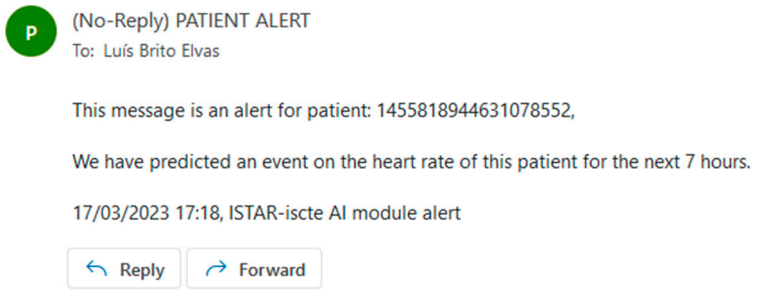

Figure 8. Patient alert from the running AI process.

Our modeling approach has significant implications for patient outcomes and clinical decision making, as it enables timely interventions that can improve patient care and potentially save lives. Every time a patient event is predicted, an email is sent to the patient's hospital or health center warning about said event. The authors consider the results satisfactory and within the expected range of values for the algorithms used (random forest and logistic regression).

4.4. Deployment

The deployment phase is crucial to the CRISP-DM process as it involves implementing the developed solution into the operational environment. We deployed the developed outlier detection model to a hospital setting in this use case.

Our research center, ISTAR-IUL at Iscte University in Lisbon, is one of the prosumers who signed the DSA between the Hospital de Santa Maria and the other entities. This agreement has enabled us to access and analyze cardiology data to develop and deploy the outlier detection model.

In this work, we present a novel approach that seamlessly integrates diverse elements, including cutting-edge technologies such as artificial intelligence, Data Sharing Agreements, and homomorphic encryption. This unique integration enables us to develop a robust and secure framework for sharing and managing medical data while harnessing the power of artificial intelligence for advanced data analysis and prediction. Our proposed method offers significant contributions compared to existing approaches, where real-world scenarios and practical data were employed for validation. The results show its potential to improve performance, efficiency, and data security in the healthcare domain. By further emphasizing the technical novelty and advantages of our method, we aim to underscore its relevance and potential impact in the field of data sharing and healthcare analytics.

Since deploying the model, we have observed several positive outcomes. The model is achieving good results and has proven to be an asset for the hospital's medical staff in monitoring patient health. Furthermore, the email alert system implemented as part of the model provides real-time patient data updates, enabling the medical staff to take prompt actions when necessary.

In addition, the deployment of our model has sparked several ongoing research endeavors aimed at exploring various aspects of cardiology data analysis. These research works involve the application of cluster algorithms to patients diagnosed with infarcts, pneumonia, and myocarditis. In particular, our future work will include efforts to develop predictive models for infarcts. Through these research initiatives, we aim to deepen our understanding of patient outcomes and disease progression, ultimately contributing to the advancement of cardiology research.

Overall, the deployment of the outlier detection model has been successful, enabling us to gain valuable insights into cardiology data and improve patient care in the hospital setting.

5. Conclusions

In the context of health data exchange, our proposed platform utilizes DSAs to regulate the secure and confidential transfer of medical data between healthcare providers, patients, and relevant stakeholders. The agreements establish clear guidelines on data scope, purpose, and security measures to ensure compliance with privacy regulations and protect sensitive information. This collaborative framework facilitates the implementation of contractual agreements, all while preserving data privacy and confidentiality through encryption techniques.

We proposed a platform composed of an ISI that consists of various components and connections that enable secure and controlled data access, providing the necessary data for the AIM, in a secure and collaborative way. The AIM is responsible for all of the AI component, where predictions can be made and knowledge may be created. While previous works have explored individual aspects such as data governance, AI applications, and data anonymization techniques in isolation, our paper uniquely combines multiple technologies to create a comprehensive and secure data sharing ecosystem. This novel combination allows for a comprehensive approach that effectively addresses the challenges and concerns of health data sharing while preserving privacy and confidentiality. By synergizing these diverse technologies, our research offers a holistic solution that sets a new standard for responsible and efficient data-driven strategies in the healthcare industry.

This research aims to provide a flexible, secure, and privacy-aware framework allowing sharing of confidential, distributed information in health entities. This allows knowledge creation based on shared services, and DSAs are one of the first steps towards data and information sharing in the health sector. We implemented an information analysis infrastructure using DSA and developed an AI module that uses encrypted data to make predictions. The experimental results demonstrate the accuracy and efficacy of our approach, with the logistic regression model achieving 96.88% accuracy and the random forest model slightly outperforming it with an accuracy of 97.62%.

To the authors' knowledge, this is one of the first approaches to use a DSA to share information in the health sector.

Author Contributions: Conceptualization, J.C.F. and L.B.E.; methodology, L.B.E.; Validation, L.B.R., M.S.D.; formal analysis, M.S.D.; writing—original draft preparation, L.B.E.; writing—review and editing, J.C.F., M.S.D. and L.B.R.; supervision, J.C.F., M.S.D. and L.B.R.;. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially funded by national funds through FCT—Fundação para a Ciência e Tecnologia, I.P., under the projects FCT UIDB/04466/2020, and FCT DSAIPA/AI/0122/2020 AIMHealth—Mobile Applications Based on Artificial Intelligence. Luís Elvas holds a Ph.D. grant, funded by FCT with UI/BD/151494/2021.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mirzaei, A.; Aslani, P.; Schneider, C.R. Healthcare data integration using machine learning: A case study evaluation with health information-seeking behavior databases. *Res. Soc. Adm. Pharm. RSAP* **2022**, *18*, 4144–4149. [CrossRef]
2. Esposito, C. Interoperable, dynamic and privacy-preserving access control for cloud data storage when integrating heterogeneous organizations. *J. Netw. Comput. Appl.* **2018**, *108*, 124–136. [CrossRef]
3. Siyal, A.A.; Junejo, A.Z.; Zawish, M.; Ahmed, K.; Khalil, A.; Soursou, G. Applications of Blockchain Technology in Medicine and Healthcare: Challenges and Future Perspectives. *Cryptography* **2019**, *3*, 3. [CrossRef]
4. Caimi, C.; Gambardella, C.; Manea, M.; Petrocchi, M.; Stella, D. *Legal and Technical Perspectives in Data Sharing Agreements Definition*; Springer International Publishing: Cham, Switzerland, 2016; pp. 178–192. [CrossRef]
5. Zhao, N. Improvement of Cloud Computing Medical Data Protection Technology Based on Symmetric Encryption Algorithm. *J. Test. Eval.* **2022**, *52*. [CrossRef]

6. Chillotti, I.; Gama, N.; Georgieva, M.; Izabachène, M. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In *Advances in Cryptology—ASIACRYPT 2016*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2016; p. 3. [CrossRef]
7. Andorno, R. The Oviedo Convention: A European Legal Framework at the Intersection of Human Rights and Health Law. *J. Int. Biotechnol. Law* **2005**, *2*, 133–143. [CrossRef]
8. Shearer, C. The CRISP-DM model: The new blueprint for data mining. *J. Data Warehous.* **2000**, *5*, 13–22.
9. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ Online* **2009**, *339*, 332–336. [CrossRef]
10. Schneider, G. Health data pools under european policy and data protection law: Research as a new efficiency defence? *J. Intellect. Prop. Inf. Technol. E-Commer. Law* **2020**, *11*, 49–67.
11. Ithayan, J.; Sundar, C. A Secured Healthcare Management and Service Retrieval for Society Over Apache Spark Hadoop Environment. *IETE J. Res.* **2023**, *69*, 684–703. [CrossRef]
12. McLennan, S.; Rachut, S.; Lange, J.; Fiske, A.; Heckmann, D.; Buyx, A. Practices and Attitudes of Bavarian Stakeholders Regarding the Secondary Use of Health Data for Research Purposes During the COVID-19 Pandemic: Qualitative Interview Study. *J. Med. Internet Res.* **2022**, *24*, e38754. [CrossRef]
13. Bak, M.; Ploem, M.; Tan, H.; Blom, M.; Willems, D. Towards trust-based governance of health data research. *Med. Health Care Philos.* **2023**, *26*, 185–200. [CrossRef] [PubMed]
14. Batarseh, F.A.; Latif, E.A. Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare. *Big Data Res.* **2016**, *4*, 13–24. [CrossRef]
15. Townend, D. Conclusion: Harmonisation in genomic and health data sharing for research: An impossible dream? *Hum. Genet.* **2018**, *137*, 657–664. [CrossRef] [PubMed]
16. Aggarwal, R.; Farag, S.; Martin, G.; Ashrafian, H.; Darzi, A. Patient Perceptions on Data Sharing and Applying Artificial Intelligence to Health Care Data: Cross-sectional Survey. *J. Med. Internet Res.* **2021**, *23*, e26162. [CrossRef]
17. Joda, T.; Waltimo, T.; Probst-Hensch, N.; Pauli-Magnus, C.; Zitzmann, N.U. Health Data in Dentistry: An Attempt to Master the Digital Challenge. *Public Health Genom.* **2019**, *22*, 1–7. [CrossRef] [PubMed]
18. Allam, Z.; Jones, D. On the Coronavirus (COVID-19) Outbreak and the Smart City Network: Universal Data Sharing Standards Coupled with Artificial Intelligence (AI) to Benefit Urban Health Monitoring and Management. *Healthcare* **2020**, *8*, 46. [CrossRef]
19. Car, J.; Sheikh, A.; Wicks, P.; Williams, M.S. Beyond the hype of big data and artificial intelligence: Building foundations for knowledge and wisdom. *BMC Med.* **2019**, *17*, 139. [CrossRef]
20. Elvas, L.B.; Serrão, C.; Ferreira, J.C. Sharing Health Information Using a Blockchain. *Healthcare* **2023**, *11*, 170. [CrossRef]
21. Sheikhtaheri, A.; Tabatabaee Jabali, S.M.; Bitaraf, E.; TehraniYazdi, A.; Kabir, A. A near real-time electronic health record-based COVID-19 surveillance system: An experience from a developing country. *Health Inf. Manag. J.* **2022**. [CrossRef]
22. Morley, J.; Machado, C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; Floridi, L. The ethics of AI in health care: A mapping review. *Soc. Sci. Med.* **2020**, *260*, 113172. [CrossRef]
23. Wang, S.; Pershing, S.; Lee, A. AAO Taskforce AI AAO Med Big data requirements for artificial intelligence. *Curr. Opin. Ophthalmol.* **2020**, *31*, 318–323. [CrossRef] [PubMed]
24. Arca, S.; Hewett, R. Analytics on Anonymity for Privacy Retention in Smart Health Data. *Future Internet* **2021**, *13*, 274. [CrossRef]
25. Alison Paprica, P.; Sutherland, E.; Smith, A.; Brudno, M.; Cartagena, R.G.; Crichlow, M.; Courtney, B.K.; Loken, C.; McGrail, K.M.; Ryan, A.; et al. Essential requirements for establishing and operating data trusts: Practical guidance co-developed by representatives from fifteen canadian organizations and initiatives. *Int. J. Popul. Data Sci.* **2020**, *5*, 1353. [CrossRef]
26. Ismail, L.; Materwala, H.; Karduck, A.P.; Adem, A. Requirements of health data management systems for biomedical care and research: Scoping review. *J. Med. Internet Res.* **2020**, *22*, e17508. [CrossRef] [PubMed]
27. Indra Priyadharshini, S.; Vigilson Prem, M. Secure e health cloud framework for patients' EHR storage and sharing for indian government healthcare model. *Proc. Est. Acad. Sci.* **2020**, *69*, 266–276. [CrossRef]
28. Yang, Y.; He, F.; Han, S.; Liang, Y.; Cheng, Y. A Novel Attribute-Based Encryption Approach with Integrity Ver-ification for CAD Assembly Models. *Engineering* **2021**, *7*, 787–797. [CrossRef]
29. Wu, Y.; He, F.; Zhang, D.; Li, X. Service-Oriented Feature-Based Data Exchange for Cloud-Based Design and Manufacturing. *IEEE Trans. Serv. Comput.* **2018**, *11*, 341–353. [CrossRef]
30. *NIST FIPS 197-upd1*; Advanced Encryption Standard (AES). National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2001. [CrossRef]
31. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000.