information

Article

A Privacy-Preserving and Standard-Based Architecture for Secondary Use of Clinical Data

Mario Ciampi *, Mario Sicuranza and Stefano Silvestri

Institute for High Performance Computing and Networking, National Research Council of Italy, Via Pietro Castellino, 111, 80131 Naples, Italy; mario.sicuranza@icar.cnr.it (M.S.); stefano.silvestri@icar.cnr.it (S.S.)
* Correspondence: mario.ciampi@icar.cnr.it

Abstract: The heterogeneity of the formats and standards of clinical data, which includes both structured, semi-structured, and unstructured data, in addition to the sensitive information contained in them, require the definition of specific approaches that are able to implement methodologies that can permit the extraction of valuable information buried under such data. Although many challenges and issues that have not been fully addressed still exist when this information must be processed and used for further purposes, the most recent techniques based on machine learning and big data analytics can support the information extraction process for the secondary use of clinical data. In particular, these techniques can facilitate the transformation of heterogeneous data into a common standard format. Moreover, they can also be exploited to define anonymization or pseudonymization approaches, respecting the privacy requirements stated in the General Data Protection Regulation, Health Insurance Portability and Accountability Act and other national and regional laws. In fact, compliance with these laws requires that only de-identified clinical and personal data can be processed for secondary analyses, in particular when data is shared or exchanged across different institutions. This work proposes a modular architecture capable of collecting clinical data from heterogeneous sources and transforming them into useful data for secondary uses, such as research, governance, and medical education purposes. The proposed architecture is able to exploit appropriate modules and algorithms, carry out transformations (pseudonymization and standardization) required to use data for the second purposes, as well as provide efficient tools to facilitate the retrieval and analysis processes. Preliminary experimental tests show good accuracy in terms of quantitative evaluations.

Keywords: ETL architecture; secondary use of clinical data; HL7 FHIR; information retrieval; privacy laws; pseudonymization

Citation: Ciampi, M.; Sicuranza, M.; Silvestri, S. A Privacy-Preserving and Standard-Based Architecture for Secondary Use of Clinical Data. *Information* **2022**, *13*, 87. https://doi.org/10.3390/info13020087

Academic Editor: Willy Susilo

Received: 24 December 2021
Accepted: 10 February 2022
Published: 13 February 2022

1. Introduction

The large availability of health data and documents in digital format managed in Health Information Systems (HISs) facilitate the advancement of clinical research, the improvement of care services through academic training, and supports the administrative and health processes through control and governance (i.e., clinical audits for quality enhancement and clinical governance) [1]. Furthermore, health information is often further processed and linked with other datasets, such as data produced in clinical trials, allowing for the extraction of further insights, which are very useful not only for clinical practices, but also to design and improve information health systems and to contribute to the implementation or more effective policy making [2].

While the benefits of **secondary uses** of medical and clinical data are relevant for improving the quality of care, questions about how this data is accessed, by whom, and under what circumstances still raise many issues. In addition to the obligation to obtain informed consent before participating in the research [3], data sharing (especially when they contain **personal** and **sensitive** information and when data is made available outside the company or geographic area where it is generated) can lead to a loss of privacy for

individuals, such as patients and health professionals. For this reason, data security and patient's privacy are considered crucial issues to face. To guarantee the rights of the interested parties, a number of international and national laws and rules have been issued, such as the General Data Protection Regulation (GDPR) in Europe [4,5] and the Health Insurance Portability and Accountability Act (HIPAA) in the USA [6–8]. To the end of allowing the use of personal clinical data for secondary uses (i.e., for research and health planning purposes) in compliance with these laws, only de-identified data can be processed, in particular when data are shared or exchanged across different institutions. In fact, such laws, along with the ones of the various countries, establish that the secondary use of clinical data is allowed only in avoiding future associations with the people they refer to, in the case they have not provided explicit consent, or when data must be processed outside the institution/company or country that collected them [9,10].

Another challenge to processing health data for secondary purposes is the non-uniformity and variety of the information contained in clinical documents and data, due to the lack of a homogeneous data representation, considering that HISs manage structured and unstructured documents. Although there still exist many challenges that have not been fully addressed and that can hinder the use of heterogeneous and, in particular, unstructured health data, there are clear opportunities arising from their secondary use, by leveraging tools able to efficiently incorporate structured and unstructured health data into a common structured format, taking into account at the same time data sensitivity, privacy, and ethical issues [11]. Thus, with the purpose of guaranteeing the effective use of this information, unstructured data must be converted according to a homogeneous, standard, and coded representation, exploiting advanced techniques such as big data analytics [12], and effective de-identification techniques must be applied.

This paper presents a novel modular architecture capable of collecting clinical data from heterogeneous sources and transforming them into formats useful for the clinical secondary use, respecting the aforementioned requirements. The proposed architecture provides for the implementation of a pseudonymization process to make the data used in the secondary purpose a secure process, also compliant with law requirements. Furthermore, the proposed architecture applies an *Extract, Transform, and Load* (ETL) process to retrieve, transform, and make this data available and to obtain strictly necessary information from a set of heterogeneous clinical documents and data. It also leverages the ETL process to organize information in a standard format, namely HL7 FHIR, which has been adopted for its peculiarities.

The main contributions of this paper are:

- The definition of an architecture able to provide the integration of heterogeneous health and clinical data according to the FHIR interoperable standard, which also includes the capability of automatically selecting and applying the most suitable pseudonymization algorithms, enabling the secondary use of clinical data;
- The proposal of a solution to perform the analysis of the obtained large collections of FHIR resources.

After this introduction, the paper is organized as follows. Section 2 provides some background information and Section 3 describes the most recent related works. Section 4 presents the proposed architecture, highlighting the main components and the techniques used for data transformation. In Section 5, implementation details are presented. In Section 6, a use case is described, where preliminary implementation of the proposed architecture is validated. Finally, Section 7 concludes the paper.

2. Background

This Section provides an overview on the main privacy laws and aspects, standards and techniques to allow the secondary use of clinical data.

2.1. International Regulation and Law

In the United States, the HIPAA of 1996 is composed of a set of documents that establish the Privacy, Security and Patient Safety Rules to be used to protect the confidentiality of clinical information. The HIPAA privacy policy also provides definitions and standards for de-identification of clinical data. In fact, the HIPAA "*Safe Harbor*" defines 18 data items called *Protected Health Information (PHI)*, which must be removed in such a way that clinical data is considered to be de-identified.

In Europe, the GDPR addresses challenges for the protection of personal data. It identifies six main principles for data protection: (i) *Lawfulness of processing*, transparency, and fairness; (ii) *purpose limitations*, according to which the data is collected only for specific and explicit purposes, legitimized by the functions to be created; (iii) *data minimization*, according to which the data must be used only for the explicit primary purposes; (iv) *accuracy*, according to which the data processed must always be updated or deleted if not correct, by guaranteeing high accuracy; (v) *storage limitations*, for which data must be deleted immediately after it has been used and no longer than necessary; and (vi) *integrity and confidentiality*, according to which whoever manages the data must guarantee full data integrity (data must not be modified in an unauthorized manner) and confidentiality (data must not be accessed without authorization). Furthermore, article 32 of GDPR provides information on the security of the personal data processing.

2.2. ETL Process

The ETL process allows the extrapolation of the minimum information of interest from big and heterogeneous sources, and it is a preparatory operation for the collection and management of data by means of a homogeneous representation format. The ETL process consists of three phases:

- **Extraction**: In this phase, the data is extracted from heterogeneous sources. The extracted data is managed in a staging area, such as a data lake;
- **Transformation**: In this phase, the collected data is transformed by applying the correct format, which is defined through the application of the following rules:
  - *Standardization*: Includes the selection of useful data, the methods, and the standard format;
  - *Deduplication*: Identifies useless duplicated data;
  - *Verification*: Eliminates incorrect data;
  - *Sorting*: Groups and sorts data;
  - *Other activities:* Depend on the context and the purposes of the ETL process (i.e., de-identification).
- **Loading**: In this phase, the extracted and transformed data are loaded into a new destination to the end of managing and analyzing the data, such as a data management system, data lake, or any kind of data repository. The upload of the data can be full or incremental.

2.3. Data De-Identification

De-identification may be implemented by using anonymization or pseudonymization techniques. Data are considered anonymous when the data subjects to which they refer to are no longer identifiable. The privacy norms do not prescribe any particular technique for anonymization; it is therefore up to the individual data controllers to ensure that any anonymization process chosen is sufficiently robust. With pseudonymization, data are processed in a way that it cannot be attributed to a specific individual without the use of additional information. Therefore, it is necessary to make such additional information inaccessible and separate. Pseudonymized data still permit identifying a person, while anonymized data can no longer be associated with specific individuals. Pseudonymization aims to ensure that an individual is not identified on personal data, which are replaced by one or more artificial identifiers, or pseudonyms, which cannot be linked directly to

their corresponding nominative identities, making the data record less identifiable while remaining suitable for data analysis and data processing [13]. A person is considered identifiable, directly or indirectly, by using personal data, such as a name, identification number, location data, or one or more characteristic elements concerning their physical, physiological, genetic, economic, cultural, or social identity.

Although 100% anonymization is the most desirable objective from the point of view of protection of personal data, in some cases it is not possible to remove useful data and therefore a residual risk of re-identification must be taken into account [14]. It is not always possible to lower the risk of re-identification under a previously defined threshold by maintaining a useful dataset for a specific processing. For this reason, a robust de-identification process aims to reduce the risk of re-identification under a certain threshold. This threshold depends on various factors, such as existing mitigation controls, the impact on people's privacy in the event of new identification, and the use of data. It is essential that de-identified data prove to be useful and therefore the process depends on the purpose for which such data are managed as well as on the risk of re-identification.

Furthermore, it is unlikely that a fully automated process can be used to identify every personal data in different contexts or decide how to maximize the usefulness of the data by applying specific techniques to a number of variables. Therefore, taking into account the context in the overall evaluation of the process, the intervention of a human expert is often required, depending on the purpose of data use, in order to provide information on useful and less useful data.

There are many de-identification techniques designed for the secondary use of clinical data. The main techniques known in the literature used for the de-identification of personal information are:

- *Character/record masking* [15]: Represents an anonymization technique that provides for the cancellation of main personal identifiers, such as name, date of birth, and more. This technique is used for example in legal databases.
- *Shuffling* [16]: Represents an anonymization technique that has the purpose of modifying the data in order to eliminate the relationship between the data and person by replacing the sensitive data with a different one belonging to the same type and extracted from the same corpus. There are numerous methodologies of anonymization that use this technique at different levels in the information structure [17].
- *Pseudonymization* [13]: Allows the replacement of an attribute with another value. Pseudonymization can be defined as the technique with which a unique attribute of one data is replaced with another. The person could, however, be identified indirectly.
- *Generalization* [18]: Is an anonymization technique that aims to generalize attributes associated with people. For example, the information relating to a date of birth can be generalized using only the year of birth and avoiding the indication of the day and month.

The most important techniques include [19]:

- **K-anonymity** [20]: This technique, through aggregation with k different people, tries to prevent the identification of a person. By sharing the same value with k people, it is more difficult to identify a specific person. The generalization therefore allows to share a given value of an attribute for a greater number of people. The main flaw of the k-anonymity model is that it does not protect against deductive attacks. Furthermore, with the intersection of different groups represented by different attributes, it can be even easier to identify the person.
- **L-diversity** [21]: This technique extends the k-anonymity technique to make attacks by deterministic deduction ineffective by ensuring that in each equivalence class, there are at least L different values of L attributes. L-diversity is subject to attack by probabilistic deduction.
- **T-Closeness** [20,22]: This technique represents an evolution of L-diversity, as the goal is to create equivalent classes that are similar to the initial attributes. This is useful when it is necessary that the values obtained are as close to the starting ones.

This technique requires that not only must exist at least L different values within each equivalence class, as indicated by the L-diversity technique, but also that each value is represented as many times as necessary to reflect the initial distribution of each attribute.

2.4. HL7 CDA and HL7 FHIR Standards

HL7 is an international Standards Developing Organization (SDO) that defines a standard structure for the exchange, integration, sharing, and retrieval of electronic health information. The HL7 Clinical Document Architecture (CDA) [23] standard specifies syntax rules and provides a basic structure for implementing the entire semantics of a clinical document. It thus enables the computerized exchange of clinical documents.

HL7 Fast Healthcare Interoperability Resources (FHIR) [24] is the emerging health information standard that offers a powerful and extensible data model with standardized semantics and data exchange enabling all systems using FHIR to collaborate. Converting data to the FHIR format allows quick connection to existing data. An important component of the FHIR specification is represented by the RESTful APIs, which are a collection of well-defined interfaces for making different applications able to interoperate.

Although the FHIR specification is a platform specification to implement a FHIR-based solution for a specific subdomain of healthcare that is able to consider different regulations, requirements, etc., it requires further adaptations. These ones, typically specified in Implementation Guides (IGs), include a localization of the particular standard resource elements that are used, possible additional elements, the APIs and terminologies to use, and others [25]. FHIR also enables rapid data exchange in modern mobile and web development implementations.

The data in the HL7 FHIR format can also be easily converted into other standards, such as the Observational Health Data Sciences and Informatics (OHDSI), Observational Medical Outcomes Partnership (OMOP), and Common Data Model (CDM) [26,27], leveraging available OHDSI open tools, such as FhirToCdm [28], or OMOP on FHIR [29,30]. FHIR resources have been designed with the purpose of improving and easing the interoperability of clinical data, allowing one to easily map them to further formats and data models, as demonstrated by [31] for the case of i2b2 data [32], or implemented by FHIR2TranSMART [33] to directly map them to the TransSMART data model [34]. The conversion of HL7 FHIR to different standards not only supports the interoperability, but enables the use of already available data warehouse systems and ETL approaches specifically tailored for different standards, improving the possibility of secondary use of data.

3. Related Works

The availability of large collections of clinical data and the need for techniques for information extraction purposes and their further reuse has led in recent years to an active research area, which has presented different and effective solutions in the literature, as well as highlighted and discussed the main issues related to those processes.

For example, the authors of [11] presented an analysis of the challenges and opportunities beyond simple structured data analysis of electronic health records, highlighting that although there exists in this area many challenges that have not been fully addressed, there are clear opportunities for the definition and implementation of effective tools that efficiently incorporate both structured and unstructured data for extracting useful information. Moreover, they also showed how the access to clinical data is still very restricted due to data sensitivity and ethical issues. The main result of their review is that a possible solution to the aforementioned issues is to allow the accessibility of unstructured data by developing tailored machine learning methods, and at the same time accelerating further research on privacy preserving techniques for the de-identification and pseudonymization of clinical texts.

In [35], the various challenges and research issues concerning the secondary use of unstructured clinical data are discussed, analyzing the methodologies to preserve the privacy, and the sensitive data of the patients. They focused their analysis on ambiguous and vague terminology and on how different legislation affects the requirements for de-identification. Moreover, the needs of adopting different approaches in the case of unstructured or structured data is underlined, also analyzing the impact of the approaches based on the named entity recognition to replace sensitive data with surrogates. Finally, the re-identification risks are discussed.

The work described in [36] presents a pseudonymization system to facilitate the exchange of data for secondary uses, designed and developed in accordance with ISO/EN 13606 [37], which has as its main objective the standardization of EHR transfers, or part of them, in a semantically operable manner. This approach allows for the total or partial anonymization of data extracted from the Electronic Health Record (EHR and HIS in general), eliminating all of the demographic references or some of the them (sex, date of birth, and place of residence) selected by the user. The elimination happens prior to storage, with the substitution of these references with identifiers that might be associated with specific demographic entities. In this way, while the transmitted data is pseudonymized and the clinical information cannot be associated with a specific entity, the eliminated demographic information is locally stored and may be recovered by means of consulting the identifiers that appear in the anonymized extract whenever the rights pertaining to access to the information are verified. This methodology requires that data have been previously extracted and structured and cannot be exploited to pseudonymized unstructured text.

The authors of [38] introduced a method for obscuring date information for clinical events and patient characteristics, preserving temporal relations in clinical data while maintaining privacy at the same time. This approach, named shift and truncate, first assigns each patient a random temporal shift value, making all dates in that patient's record shifted by that amount. In this way, temporal relations are also preserved, which can be critical for secondary uses.

The standardization of clinical data is another open issue that is also strictly related to the secondary use of health data. The Observational Health Data Sciences and Informatics (OHDSI) [26] is an international collaboration whose goal was to create and apply open-source data analytics solutions to a large network of health databases. For this purpose, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [27], which represents healthcare data from diverse sources in a consistent and standardized way, was extended and improved, also providing a collection of specifically customized software tools, not only to facilitate data exploration and evidence generation, but also to support the standardization process. Among them, it is worth mentioning WhiteRabbit [39], which helps the preparation for ETL of longitudinal healthcare databases into the OMOP CDM, by scanning the source data and providing detailed information on the tables, fields, and values that appear in a field, allowing a user to connect source data structures to the CDM data structure. Another OHDSI tool useful for the standardization is Usagi [40], which supports the process of mapping codes from source raw datasets into the standard terminologies stored in the OMOP CDM format. Although WhiteRabbit and Usagi tools can facilitate the standardization of raw clinical datasets into an OHDSI standard, they have some limits. Firstly, they both process only structured data, being not able to extract relevant information from unstructured natural language texts, which represent a large and important part of available clinical data. Furthermore, they rely on classic RDBMS or CSV and Excel formats, which are not suited for the processing of big data collections. Finally, they do not perform any task related to the anonymization or pseudonymization processes, requiring additional tools to address privacy and ethical issues related to the processing of clinical and medical data.

A framework for processing unstructured clinical documents of EHRs and for their integration within standardized structured data in the OMOP CDM format is presented in [41]. In detail, the authors described a framework named Staged Optimization of

Curation, Regularization, and Annotation of clinical Text (SOCRATex), which first extracts clinical notes for the target population and preprocesses the data. Then, the framework defines an annotation schema with a hierarchical structure and applies a document-level hierarchical annotation by means of a machine learning approach. Finally, it indexes the annotations for a search engine system. The tests performed on real clinical EHRs showed that the system was able to integrate the annotated documents into the OMOP CDM database [27].

Big Data Analytics (BDA) approaches have also been leveraged for the analysis of unstructured clinical data. The authors of [12] described an approach to exploit the health data from the Italian interoperable EHR platform, allowing for the extraction of useful data through a combination of BDA techniques and Natural Language Processing (NLP) methods, supporting in this way the process of secondary use of information.

In [42], a complex framework of BDA systems and tools has been presented, with the purpose of addressing some of the issues related to the analysis of big biomedical unstructured data collections.

In summary, as described in the following of this paper, our proposed approach addresses many of the issues of the methods and systems presented in the literature and briefly described in this Section and provides many innovative aspects. It provides a single architecture able to support pseudonymization processes that can automatically select the most appropriate de-identification approaches. It also supports the standardization of health data leveraging the recent FHIR format by extracting data from both structured, semi-structured, and unstructured documents and datasets. Furthermore, it is capable of dealing with very large structured, semi-structured, or unstructured clinical datasets as well as integrating ETL and BDA functionalities to facilitate the secondary use.

4. System Architecture

This Section describes the overall architecture of the proposed system. The aim of this architecture is to facilitate the secondary use of health data by structuring them in a standardized way and in compliance to the privacy norms.

The proposed architecture uses a pseudonymization action in the transformation phase of the ETL process. This action allows transforming the data, keeping the information content of interest for secondary clinical use, and eliminating the direct link between the data and interested party (that is, the patient).

The architecture uses transformation features that make use of ad-hoc profiles, depending on the purpose of use of the data [43]. Figure 1 shows the system architecture. The modules of the architecture are described in detail in the following Section.

4.1. Extraction Module

The purpose of this module is to extract all information of interest from heterogeneous sources. In detail, it analyzes the input in order to identify which extraction rules must be applied. At that point, it uses the most appropriate extraction algorithm to obtain this information. The heterogeneous sources of information taken into consideration are: (i) Documents structured in standard format, such as the HL7 CDA; (ii) documents structured through proprietary rules and formalisms; (iii) documents in free text, for which an extraction algorithm obtained by means of a rule-based approach or through appropriately trained hybrid approaches is envisaged [12]. Finally, this component is also capable of extrapolating and aggregating information even from single data.

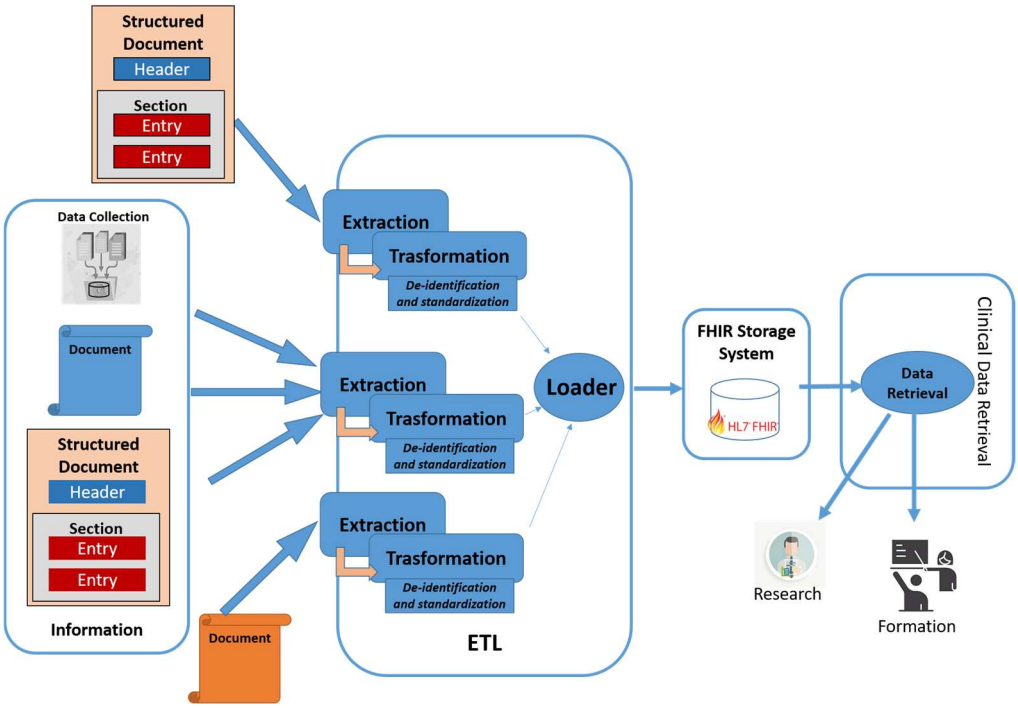

Figure 1. Proposed architecture.

The extraction module also allows for the association of particular tags with the data to classify the obtained information and permits to organize the information adequately during the standardization phase. A possible approach is through NLP techniques, which enable the extraction of structured data from unstructured documents [44,45], or even approaches based on deep learning, which have a great ability to analyze different types of data and extract information with specific characteristics [46]. In detail, the extracted information with the same value from different sources can be aggregated with an indication of whether or not it intersects. The aggregation process can also retrieve information and aggregate it into report documents, or suggest changes to a specific clinical document, such as a patient summary.

4.2. Transformation Module

This module performs various elaborations with different purposes, exploiting the data obtained from the *extraction module* in the previous phase. The most important processing steps are *pseudonymization* and *standardization* of the obtained data. Through the *pseudonymization* phase, personal data are thus not directly attributable to a specific person. In this phase, some information, such as the patient's identification, name, surname, genetic information, etc., is deleted. Other types of information are transformed (pseudonymized), depending on the use of the data: For instance, the residential address is generalized and does not include the street and house number, but only the province or the municipality. Other data will be merged with each other, realizing aggregated data.

The module is able to apply different pseudonymization techniques (previously described in Section 2), selecting the most appropriate one depending on the different kinds of data to process, following the algorithm shown in the flow chart depicted in Figure 2.

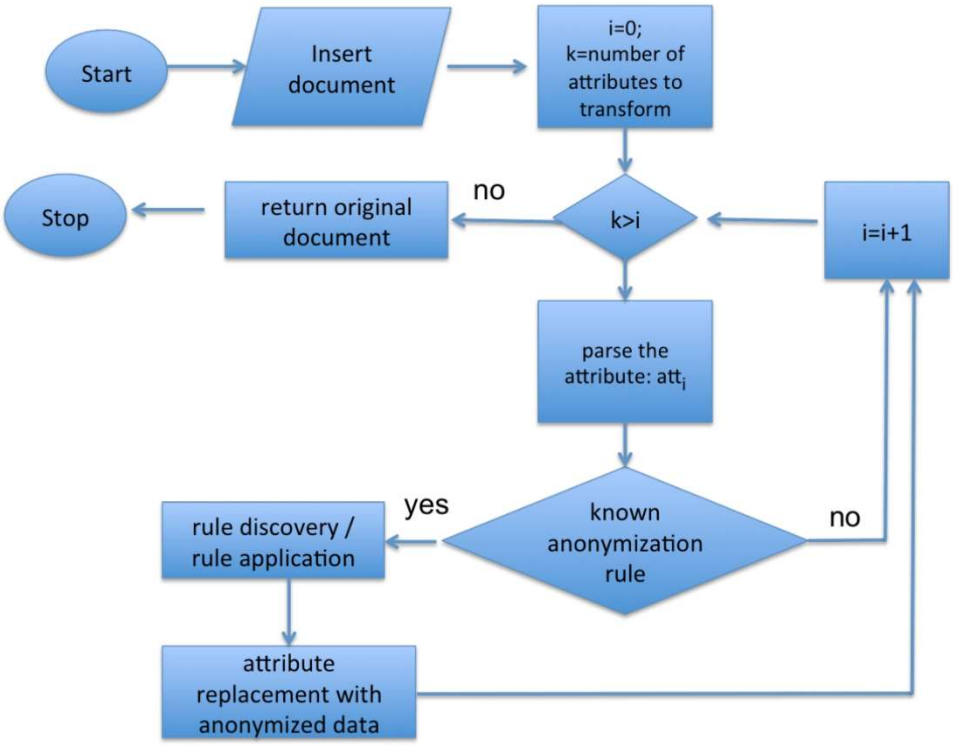

Figure 2. Flow chart of an anonymization algorithm.

In detail, the algorithm analyzes each attribute of the extracted data and leverages a rule-based approach to implement an attribute-to-anonymization technique association, in case that attribute belongs to a known rule. An example of these association rules is shown in Table 1, where the anonymization (deletion of that value) or the pseudonymization technique (L-diversity) is associated with different attributes of the extracted data. This approach is very flexible and can be used in different operational contexts.

Table 1. Example of the rules for the selection of the de-identification techniques.

Attribute	De-Identification Technique
Address	L-diversity algorithm
Date	L-diversity algorithm
Identifier	Deletion of the value

The other main task of the transformation module is the standardization phase, which organizes the obtained data by means of standard structures. The module collects information through HL7 FHIR resources. For example, measurements and simple assertions made about a patient, device, or other subjects is organized as the Observation FHIR resource.

The transformation module aims to create technical mapping that follows a preliminary conceptual mapping from the source document to the FHIR resources. In general, two types of conceptual mappings in FHIR can be performed on the basis of the necessary output: (i) A new FHIR document can be created in a semantically equivalent way to the source document, which in this case, a composition resource can be used; or (ii) some information of interest is extracted from the source document and entered in the FHIR resources, which in this case, the data are stored in the correct FHIR resources based on the meaning of the extracted data (e.g., resources such as observation, condition, etc. can be used).

The proposed approach is based on the second methodology: Once information has been extracted using the extraction module, it will be stored in FHIR resources useful for the

intended purpose. More specifically, in order to formalize the mapping between documents in the HL7 CDA format and FHIR resources (also with defined profiles but not currently published in Implementation Guides), the extraction module uses a configuration file that contains XPath expressions pointing to the sections and fields of the source document represented in HL7 CDA and the elements and attributes of a target FHIR resource.

The architecture uses specific resource profiles, in order to apply further constraints and encoding related to the represented data. This phase is facilitated by the tagging carried out in the extraction module.

4.3. Loader Module

This module implements the storage and management of the FHIR resources obtained from the transformation module. In detail, the resources built through the transformation module are sent to a FHIR server used to store and manage the data, in order to make clinical data available for the second purposes. The use of FHIR services allows to exchange data through RESTful APIs based on the standard. In this case, various tools can be adopted, such as the *Azure API for FHIR* [47] or the *HAPI FHIR* [48].

The success of the emerging FHIR standard as a next generation standard for clinical interoperable data is also highlighting the focus of the scientific community on the research and development of FHIR-based data access approaches [49], which have demonstrated their feasibility in answering queries [50] or semantic queries [51] on large collections, allowing for the implementation of advanced Personal Health Record (PHR) systems [52,53], enabling for both primary and secondary use of the data.

4.4. Data Retrieval Module

This module acts as an interface to the FHIR server, which manages the clinical data stored for the second purpose, allowing to retrieve the information of interest. The data retrieval module also allows data retrieval for research purposes in a specific application domain or field, or to retrieve data for training purposes in specialized courses.

5. Implementation Details

In this Section, a proof of concept of the implemented architecture is illustrated, providing further implementation details with respect to the modules of the architecture.

The **extraction module** is implemented using the *Apache Spark* BDA framework [54], which acts as a data collector and integrator. The use of a big data approach permits the processing of the large size of information produced in clinical environments and stored in the HISs. The implementation follows the principles detailed in [12,55].

The **transformation** and **load modules** use specific Spark User Defined Functions (UDFs) [56] for data normalization and integration and a MongoDB NoSQL database for data storage [57]. The data extracted and stored in the MongoDB database is accessed through Spark SQL [58]. These data can be also accessed through the FHIR API and FHIR Repository, as shown in [25].

Data management, obtained through the load module, allows for information retrieval in an interoperable and simple way and it is implemented by HAPI FHIR [18]. HAPI FHIR is a complete implementation of the HL7 FHIR standard for healthcare interoperability in Java. The FHIR API is based on the JAXB and JAX-WS APIs. Therefore, by means of FHIR transactions, the proposed architecture is capable of making the resource profiles managed by the FHIR repository available. More in detail, the RESTful API FHIR transactions used in the implemented platform are:

- *Read*, to get the status of a specific resource. It is used by information retrieval applications interested in a specific resource profile;
- *Search*, to search for a specific FHIR resource and obtain information of interest;
- *Update* and *patch*, to update an existing resource. It is used to modify some statistical data or to update specific clinical observations;
- *Delete*, to remove a specific resource.

Additional capabilities for second purpose analyses with the proposed platform are provided by the possibility of applying the dynamic data warehousing approach described in [12], which could be leveraged to easily create advanced statistics and analytics.

6. Preliminary Tests and Discussion

The implemented system has been tested on about 100 GB of heterogeneous real clinical data, used within the Italian national her interoperability system, with the purposes of testing the functionalities of the various modules of the architecture, namely the information extraction capabilities, the pseudonymization methods, and the integration of the data into HL7 FHIR resources. In detail, this document collection includes EHRs, patient summaries, admission and discharge notes, diagnostic and medical reports, medical prescriptions, and treatment plans coming from the EHR systems of Italian hospitals and clinical facilities. Such kinds of data allow us to test the proposed approach in a real context and on real-world data.

Moreover, additional tests have been performed on the transformed data after the load phase, verifying the analytics capabilities on the obtained structured and de-identified data.

The first set of tests analyzed the extraction module of the platform, with the purpose of demonstrating that the proposed architecture is able to extract and tag the information required by the transformation module. The experiments proved that it was possible to obtain information related to diseases from the unstructured parts of EHRs, such as the disease name, its category, and the corresponding category code, or the ICD-9 CM code used to classify diagnoses. Moreover, the tests allowed the extraction of the data associated with the diagnoses and the extraction of the clinical and patient's data, such as date of birth, place of birth, gender, etc. In Figure 3, the Entity-Relationship (E-R) diagram of a sample of the extracted information is depicted.

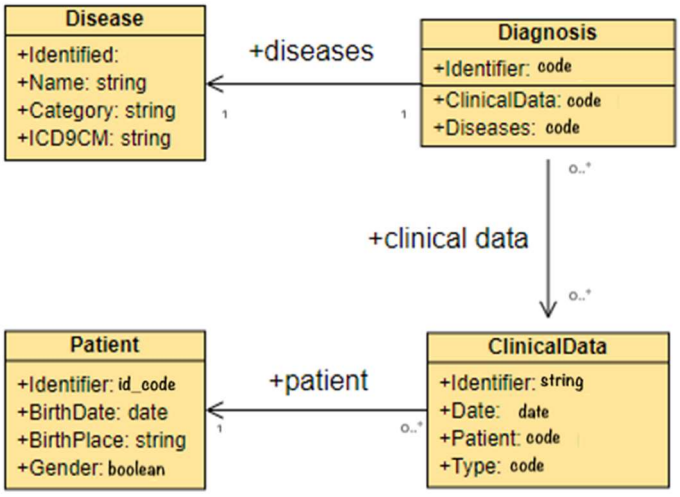

Figure 3. E-R schema of extracted data case study.

This kind of data can be easily transformed into HL7 FHIR resources in the transformation module, as shown by the further experiments performed, which tested the main tasks of the second module of the platform. The extraction and transformation process allows to address the issues related to inconsistent and not informed data items in the target structured repository.

Furthermore, when the attribute of the extracted data corresponds to a de-identification rule, the module is also able to successfully apply the required techniques for the de-identification of the data, following the algorithm and the association rules described in Section 4.2.

After the transformation phase, the obtained pseudonymized data according to the standard HL7 FHIR format have been further analyzed, simulating in this way their

secondary use. The data stored in the final FHIR repository have been processed through the dynamic data-warehousing approach proposed in [12], producing advanced statistics and analytics related to the correlation among the available data along the time and space dimensions. These tests easily demonstrated that the data stored can be leveraged to monitor and analyze a large set of parameters, such as the incidence of a disease, or the number and type of hospitalizations in a region, or the number of drugs used in a city in a specific time slot. Figure 4 shows an example of this kind of analysis, where the statistics of the hospitalizations in the Campania region of Italy have been produced.

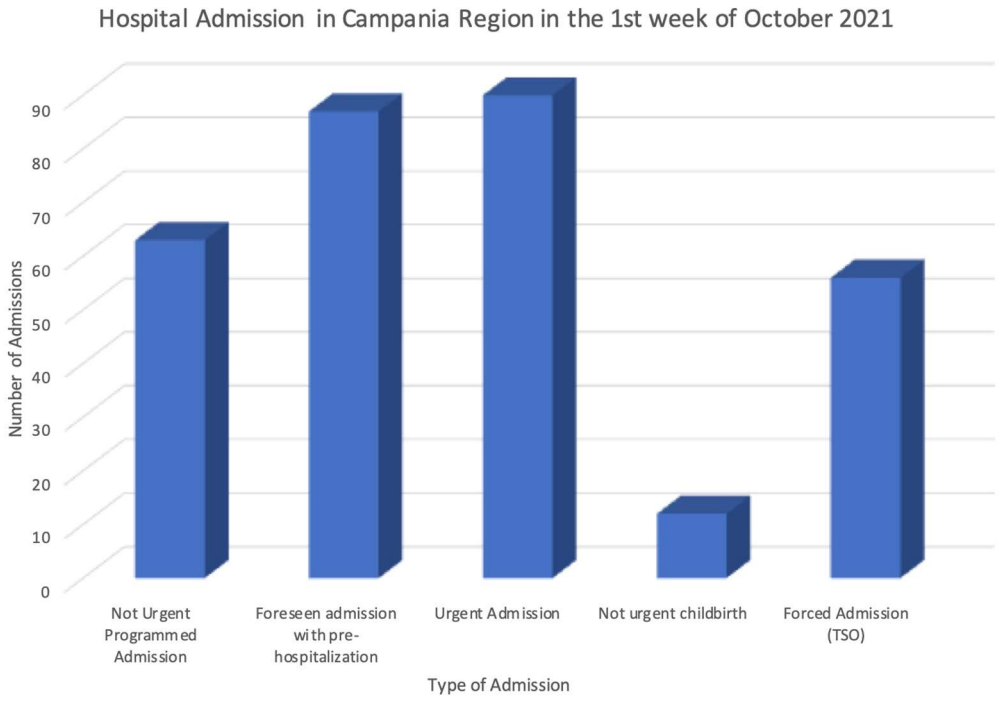

Figure 4. Example of statistics obtained through the secondary use of the data.

The six main principles highlighted in Section 2.1 represent the security context in which the proposal is implemented. Addressing these GDPR principles depends on how the proposed architecture is used and implemented. More in detail:

1. *Lawfulness of processing:* It depends on the type of processing performed by an authorized professional, for example, data could be collected through a patient's consent or for public interest.
2. *Purpose limitations:* The proposal is limited to the purpose of secondary use (research purposes, etc.), and so when the data are then processed, they will be managed ensuring this principle. In addition, the proposed architecture provides specific retrieval capabilities depending on the context of the request.
3. *Data minimization:* The information retrieval functionalities return only the requested FHIR resources that depend on the purpose of use and type of operation.
4. *Accuracy:* This principle is reached by making the transformation and loading process with a high frequency.
5. *Storage limitations:* The proposed architecture stores only the data necessary for the type of processing.
6. *Integrity and confidentiality:* The access to resources can be regulated by access control mechanisms to avoid unauthorized access and reduce related cybersecurity vulnerabilities and risks.

7. Conclusions

This work presented a modular architecture capable of collecting clinical data from heterogeneous sources and transforming them into a standard format useful for secondary use for research, governance, and medical training purposes, as well as applied the necessary de-identification techniques in order to respect privacy and ethics law requirements. The adequate and efficient use of data from heterogeneous sources was obtained by making the representation of the processed information compliant and standard, leveraging the HL7 FHIR format. Moreover, the national and international laws establish that the secondary use of clinical data is allowed only if the information exchanged is previously de-identified, in order to avoid future associations with people and to respect privacy and ethics requirements. In this work, the GDPR and HIPAA laws have been considered, which suggest to carry out a pseudonymization processes. Thus, the proposed architecture is also able to perform the further processing on the data to de-identify, allowing the secondary analyses in compliance with law requirements.

The proposed architecture leverages a combination of different approaches to implement the transformation of the data. The main transformation tasks are the pseudonymization, obtained with different techniques depending on the data type and secondary use purposes, and the standardization of the information, implemented through the HL7 FHIR standard. The architecture exploits BDA techniques, NLP, and deep learning methods for the analysis of free text, rule-based approaches for the selection of the required de-identification method, selection of the most appropriate de-identification technique, and HL7 FHIR standard resources, storage, and retrieval tools.

In this way, the platform is able to carry out the required transformations which can enable the extraction of data for secondary use in compliance with law requirements. Moreover, the platform is capable of efficiently using the extracted data and processing the information to facilitate the retrieval process, through the FHIR server. The experiments confirmed that the implemented platform is able to effectively integrate data into a standard HL7 FHIR format, as well as to apply the most appropriate de-identification technique and, finally, to support and facilitate the retrieval of required information for secondary analyses. Moreover, the flexibility of the proposed rule-based de-identification approach allows one to easily define and implement new rules, addressing in this way those cases where specific classes of data are not correctly transformed.

For future work, we plan to further optimize the architecture modules and test different implementation approaches in order to improve the performances and functionalities of the architecture. Moreover, we plan to improve the Spark UDFs for the transformation and load modules, by implementing configurable parameters that are able to extend and/or customize their functionalities. Additional tests will also be carried out on heterogeneous data. Finally, tools for the conversion and/or integration of FHIR resources to other standards, such as OMOP CMD could be integrated within the architecture.

Author Contributions: Conceptualization, M.S., M.C. and S.S.; methodology, M.C. and M.S.; software, M.S. and S.S.; validation, M.S. and S.S.; formal analysis, M.C.; investigation, M.C., M.S. and S.S.; resources, M.S.; data curation, M.S.; writing—original draft preparation, M.S.; writing—review and editing, M.C. and S.S.; visualization, M.C., M.S. and S.S.; supervision, M.C.; project administration, M.C.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Commission, grant number 883273, AI4HEALTHSEC-A Dynamic and Self-Organized Artificial Swarm Intelligence Solution for Security and Privacy Threats in Healthcare ICT Infrastructures.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Teasdale, S.; Bates, D.; Kmetik, K.; Suzewits, J.; Bainbridge, M. Secondary uses of clinical data in primary care. *J. Innov. Health Inform.* **2007**, *15*, 157–166. [CrossRef] [PubMed]
2. Hutchings, E.; Loomes, M.; Butow, P.; Boyle, F.M. A systematic literature review of attitudes towards secondary use and sharing of health administrative and clinical trial data: A focus on consent. *Syst. Rev.* **2021**, *10*, 1–44. [CrossRef] [PubMed]
3. ICH Harmonised Guideline Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice ICH E6(R2) ICH Consensus Guideline. Available online: https://ichgcp.net (accessed on 12 December 2021).
4. European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46, General Data Protection Regulation*; European Commission: Brussels, Belgium, 2016.
5. Albrecht, J.P. How the GDPR will change the world. *Eur. Data Prot. Law Rev.* **2016**, *2*, 287–289. [CrossRef]
6. Carrion, I.; Aleman, J.L.F.; Toval, A. Assessing the HIPAA standard in practice: PHR privacy policies. In Proceedings of the Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, Boston, MA, USA, 30 August–3 September 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 2380–2383. [CrossRef]
7. Summary of the HIPAA Privacy Rule. Available online: http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary (accessed on 14 October 2021).
8. United States Congress. *Health Insurance Portability and Accountability Act of 1996, Accountability Act*; United States Congress: Washington, DC, USA, 1996.
9. West, S.L.; Blake, C.; Zhiwen, L.; McKoy, J.N.; Oertel, M.D.; Carey, T.S. Reflections on the use of electronic health record data for clinical research. *Health Inform. J.* **2009**, *15*, 108–121. [CrossRef] [PubMed]
10. Katulic, T.; Katulic, A. GDPR and the reuse of personal data in scientific research. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1311–1316. [CrossRef]
11. Tayefi, M.; Ngo, P.; Chomutare, T.; Dalianis, H.; Salvi, E.; Budrionis, A.; Godtliebsen, F. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdiscip. Rev. Comput. Stat.* **2021**, *13*, e1549. [CrossRef]
12. Silvestri, S.; Esposito, A.; Gargiulo, F.; Sicuranza, M.; Ciampi, M.; De Pietro, G. A Big Data Architecture for the Extraction and Analysis of EHR Data. In Proceedings of the 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, 8–13 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 283–288. [CrossRef]
13. Bolognini, L.; Bistolfi, C. Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation. *Comput. Law Secur. Rev.* **2017**, *33*, 171–181. [CrossRef]
14. Dankar, F.K.; El Emam, K.; Neisa, A.; Roffey, T. Estimating the re-identification risk of clinical data sets. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, 1–15. [CrossRef] [PubMed]
15. Data Masking and Encryption are Different. Available online: https://www.iri.com/blog/data-protection/data-masking-and-data-encryption-are-not-the-same-things (accessed on 10 December 2021).
16. Deleger, L.; Lingren, T.; Ni, Y.; Kaiser, M.; Stoutenborough, L.; Marsolo, K.; Kouril, M.; Molnar, K.; Solti, I. Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *J. Biomed. Inform.* **2014**, *50*, 173–183. [CrossRef] [PubMed]
17. Tomashchuk, O.; Van Landuyt, D.; Pletea, D.; Wuyts, K.; Joosen, W. A data utility-driven benchmark for de-identification methods. In Proceedings of the International Conference on Trust and Privacy in Digital Business, Linz, Austria, 26–29 August 2019; Springer: Cham, Switzerland, 2019; pp. 63–77. [CrossRef]
18. Naldi, M.; D'Acquisto, G. *Big Data and Privacy by Design. Anonymization, Pseudo-Anonymization and Security*; Giappichelli, G.: Torino, Italy, 2019.
19. Kayaalp, M. Modes of De-identification. In *American Medical Informatics Association Annual Symposium (AMIA) 2017, Washington, DC, USA, 4–8 November 2017*; AMIA: Bethesda, MD, USA, 2017; p. 1044.
20. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In Proceedings of the 23rd International Conference on Data Engineering ICDE 2007, Istanbul, Turkey, 17–20 April 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 106–115. [CrossRef]
21. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3. [CrossRef]
22. Dutta, A.; Bhattacharyya, A.; Sen, A. Comparative Analysis of Anonymization Techniques. In *Privacy and Security Issues in Big Data. Services and Business Process Reengineering*; Das, P.K., Tripathy, H.K., Mohd Yusof, S.A., Eds.; Springer: Singapore, 2021; pp. 69–78. [CrossRef]
23. HL7 Clinical Document Architecture (CDA). Available online: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7 (accessed on 8 December 2021).
24. HL7 Fast Healthcare Interoperability Resources (FHIR). Available online: https://www.hl7.org/fhir/ (accessed on 8 December 2021).
25. Ciampi, M.; Marangio, F.; Schmid, G.; Sicuranza, M. A Blockchain-based Smart Contract System Architecture for Dependable Health Processes. In Proceedings of the Italian Conference on Cybersecurity ITASEC 2021, Virtual Event, Italy, 7–9 April 2021; pp. 360–373. Available online: https://www.rheagroup.com/event/itasec-2021/ (accessed on 8 December 2021).

26. Hripcsak, G.; Duke, J.D.; Shah, N.H.; Reich, C.G.; Huser, V.; Schuemie, M.J.; Suchard, M.S.; Park, R.W.; Wong, I.C.K.; Rijnbeek, P.R.; et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. *Stud. Health Technol. Inform.* **2015**, *216*, 574–578.
27. Overhage, J.M.; Ryan, P.B.; Reich, C.G.; Hartzema, A.G.; Stang, P.E. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 54–60. [CrossRef] [PubMed]
28. OHDSI FhirToCdm Github Repository. Available online: https://github.com/OHDSI/FhirToCdm (accessed on 18 January 2022).
29. Pfaff, E.R.; Champion, J.; Bradford, R.L.; Clark, M.; Xu, H.; Fecho, K.; Krishnamurthy, A.; Cox, S.; Chute, C.G.; Overby Taylor, C.; et al. Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study. *JMIR Med. Inform.* **2019**, *7*, e15199. [CrossRef] [PubMed]
30. OMOP on FHIR. Available online: https://omoponfhir.org (accessed on 18 January 2022).
31. Murphy, S.; Wilcox, A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2). *EGEMS* **2014**, *2*, 1074. [CrossRef]
32. Boussadi, A.; Zapletal, E. A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 120. [CrossRef] [PubMed]
33. FHIR2TranSMART. Available online: https://github.com/thehyve/python_fhir2transmart (accessed on 18 January 2022).
34. TranSMART Project. Available online: https://github.com/transmart (accessed on 18 January 2022).
35. Berg, H.; Henriksson, A.; Fors, U.; Dalianis, H. De-identification of Clinical Text for Secondary Use: Research Issues. In *HEALTHINF 2021*; Online Streaming, 11–13 February 2021; SCITEPRESS; 2021; pp. 592–599. Available online: https://www.scitepress.org/Papers/2021/103187/103187.pdf (accessed on 23 December 2021).
36. Somolinos, R.; Muñoz, A.; Hernando, M.E.; Pascual, M.; Cáceres, J.; Sánchez-de-Madariaga, R.; Fragua, J.A.; Serrano, P.; Salvador, C.H. Service for the Pseudonymization of Electronic Healthcare Records Based on ISO/EN 13606 for the Secondary Use of Information. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1937–1944. [CrossRef]
37. *ISO 13606-1*; Electronic Health Record Communication Part 1: Reference Model. International Organization for Standardization: Geneva, Switzerland, 2008.
38. Hripcsak, G.; Mirhaji, P.; Low, A.F.H.; Malin, B.A. Preserving temporal relations in clinical data while maintaining privacy. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 1040–1045. [CrossRef] [PubMed]
39. WhiteRabbit for ETL Design. Available online: https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design (accessed on 16 January 2022).
40. OHDSI Usagi. Available online: http://ohdsi.github.io/Usagi (accessed on 16 January 2022).
41. Park, J.; You, S.C.; Jeong, E.; Weng, C.; Park, D.; Roh, J.; Lee, D.Y.; Cheong, J.Y.; Choi, J.W.; Kang, M.; et al. A Framework (SOCRATex) for Hierarchical Annotation of Unstructured Electronic Health Records and Integration into a Standardized Medical Database: Development and Usability Study. *JMIR Med. Inform.* **2021**, *9*, e23983. [CrossRef]
42. Ciampi, M.; De Pietro, G.; Masciari, E.; Silvestri, S. Health Data Information Retrieval For Improved Simulation. In Proceedings of the 2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP), Västerås, Sweden, 11–13 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 364–368. [CrossRef]
43. Bender, D.; Sartipi, K. HL7 FHIR: An agile and RESTful approach to healthcare information exchange. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical System, Porto, Portugal, 20–22 June 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 326–331.
44. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Inform.* **2017**, *73*, 14–29. [CrossRef]
45. Alicante, A.; Corazza, A.; Isgrò, F.; Silvestri, S. Unsupervised entity and relation extraction from clinical records in Italian. *Comput. Biol. Med.* **2016**, *72*, 263–275. [CrossRef]
46. Osmani, V.; Li, L.; Danieletto, M.; Glicksberg, B.; Dudley, J.; Mayora, O. Processing of electronic health records using deep learning: A review. *arXiv* **2018**, arXiv:1804.01758.
47. Azure Healthcare APIs, A Unified Solution That Helps Protect and Combine Health Data in the Cloud and Generates Healthcare Insights with Analytics. Available online: https://azure.microsoft.com/en-us/services/healthcare-apis/#overview (accessed on 1 December 2021).
48. The HAPI FHIR Library, an Implementation of the HL7 FHIR Specification for Java. Available online: https://hapifhir.io (accessed on 1 December 2021).
49. Ayaz, M.; Pasha, M.F.; Alzahrani, M.Y.; Budiarto, R.; Stiawan, D. Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Med. Inform.* **2021**, *9*, e21929. [CrossRef] [PubMed]
50. Khalique, F.; Khan, S.A. An FHIR-based Framework for Consolidation of Augmented EHR from Hospitals for Public Health Analysis. In Proceedings of the 2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT), Moscow, Russia, 20–22 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4. [CrossRef]
51. Jiang, G.; Xiao, G.; Kiefer, R.C.; Prud'hommeaux, E.; Solbrig, H.R. Building an FHIR Ontology based Data Access Framework with the OHDSI Data Repositories. In Proceedings of the American Medical Informatics Association Annual Symposium (AMIA) 2017, Washington, DC, USA, 4–8 November 2017; AMIA: Bethesda, MD, USA, 2017.

52. Lee, Y.L.; Lee, H.A.; Hsu, C.Y.; Kung, H.H.; Chiu, H.W. Implement an international interoperable phr by FHIR—A Taiwan innovative application. *Sustainability* **2021**, *13*, 198. [CrossRef]
53. Hong, J.; Morris, P.; Seo, J. Interconnected Personal Health Record Ecosystem Using IoT Cloud Platform and HL7 FHIR. In Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, 23–26 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 362–367. [CrossRef]
54. Apache Spark. Available online: https://spark.apache.org/docs/latest (accessed on 10 December 2021).
55. Gargiulo, F.; Silvestri, S.; Ciampi, M.; De Pietro, G. Deep neural network for hierarchical extreme multi-label text classification. *Appl. Soft Comput.* **2019**, *79*, 125–138. [CrossRef]
56. Scalar Used Defined Functions (UDFs). Available online: https://spark.apache.org/docs/latest/sql-ref-functions-udf-scalar.html (accessed on 2 December 2021).
57. MongoDB: The Application Data Platform. Available online: https://www.mongodb.com (accessed on 10 December 2021).
58. Armbrust, M.; Xin, R.S.; Lian, C.; Huai, Y.; Liu, D.; Bradley, J.K.; Meng, X.; Kaftan, T.; Franklin, M.J.; Ghodsi, A.; et al. Spark SQL: Relational Data Processing in Spark. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD' 15), New York, NY, USA, 31 May–4 June 2015; ACM: New York, NY, USA, 2015; pp. 1383–1394. [CrossRef]