IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 February 2024, accepted 19 March 2024, date of publication 25 March 2024, date of current version 29 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381034

RESEARCH ARTICLE

Anonymization and Pseudonymization of FHIR Resources for Secondary Use of Healthcare Data

EMANUELE RASO[1], PIERPAOLO LORETI[2], MICHELE RAVAZIOL[3], AND LORENZO BRACCIALE[2]

[1]Department of Civil Engineering and Computer Science Engineering, University of Rome Tor Vergata, 00133 Rome, Italy
[2]Department of Electronic Engineering, University of Rome Tor Vergata, 00133 Rome, Italy
[3]Docunque SRL, 00044 Frascati, Italy

Corresponding author: Emanuele Raso (emanuele.raso@uniroma2.it)

This work was supported by the HosmartAI Project (https://www.hosmartai.eu/) funded by European Union's Horizon 2020 Research and Innovation Programme under Grant 101016834, and has been partially supported by the PRECISION project (CUP H73C22001600002) funded by Sardinia Region, Italy.

ABSTRACT Along with the creation of medical profiles of patients, Electronic Health Records have several secondary missions, such as health economy and research. The recent, increasing adoption of a common standard, i.e., the Fast Healthcare Interoperability Resources (FHIR), makes it easier to exchange medical data among the several parties involved, for example, in an epidemiological research activity. However, this exchange process is hindered by regulatory frameworks due to privacy issues related to the presence of personal information, which allows patients to be identified directly (or indirectly) from their medical data. When properly used, de-identification techniques can provide crucial support in overcoming these problems. FHIR-DIET aims to bring flexibility and concreteness to the implementation of de-identification of health data, supporting many customised data-processing behaviours that can be easily configured and tailored to match specific use case requirements. Our solution enables faster and easier cooperation between legal and IT professionals to establish and implement de-identification rules. The performance evaluation demonstrates the viability of processing hundreds of FHIR patient information data per second using standard hardware. We believe FHIR-DIET can be a valuable tool to satisfy the current regulation requirements and help to create added-value for the secondary use of healthcare data.

INDEX TERMS Anonymisation, de-identification, FHIR, healthcare, pseudonymisation, privacy.

I. INTRODUCTION

Electronic Health Records (*EHRs*) exist for the primary reason of storing patients' medical data in order to create their medical history, thus allowing doctors to provide the most appropriate care. However, this huge amount of medical data can have several secondary, extremely important missions, such as health economics, health research, or clinical or epidemiological research on a specific disease [1], [2]. Data processing for medical research typically requires the informed consent of the involved patients. However, collecting all consents may sometimes be infeasible, for instance, when analyzing data retrospectively and on a large scale. Indeed, medical research requires a tremendous amount of data to cover different kinds of people and collect as exhaustive a population sample as possible. This data can come from various, heterogeneous sources and researchers often have to deal with different data structures, or in the worst case, unstructured data. In 2011 the Health Level Seven International (HL7) health-care standards organization[1] defined a new standard, *Fast Healthcare Interoperability Resources*[2] (*FHIR*), a set of rules and specifications to ease the interoperability and exchange of healthcare data [3]. The standard describes data formats and elements

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

[1]https://www.hl7.org/index.cfm
[2]https://hl7.org/fhir/

(called *resources*) and an API for their exchange, defining a document-centric approach that uses these elements as services. The wide adoption of FHIR mitigates the problem of different data structures, facilitating the exchange of medical data between different parties. However, data protection regulations block this exchange due to the presence of patient identifiers, *Personally Identifiable Information* (*PII*), in the data. In most cases, these identifiers are not useful for the proper use of medical data for their desired purposes; on the contrary, their presence in fact hinders their use. Removing these identifiers does not completely solve the problem: along with identifiers that *directly* link the data to the related patient, there may be other information, called *quasi-identifiers* [4], which enable the reference with the patient *indirectly* and cannot be removed because they are necessary for the intended purposes. By correlating this information with the one coming from external sources, it is possible to link the medical data to the related patient even if the identifiers have been removed [5], [6].

Re-establishing the connection between identities and data exposes patients' vulnerabilities, potentially enabling the exploitation of their physical and psychological well-being [6]. Knowledge of a person's health conditions, mental health history, or sensitive medical issues can be used for targeted manipulation or coercion. For example, an individual's mental health history could be leveraged for blackmail, leading to immense emotional distress and harm. Moreover, re-identification can also result in discriminatory practices. Patients who have certain health conditions or histories may face discrimination in employment, insurance, or social settings if their private health information is exposed. Additionally, re-identification can contribute to the stigmatisation of individuals with specific medical conditions, exacerbating social and emotional burdens. It is thus fundamental to carefully assess the risk of re-identification [7], [8], an operation should be done usually case by case.

A. REGULATORY FRAMEWORK

Various regulatory frameworks, such as the *US HIPAA Privacy Rule* and the *EU General Data Protection Regulation (GDPR)*, have been defined for the proper handling of these identifiers to enable the secure use of medical data. The most common approach is the use of *de-identification*, a process that removes identifying information from a dataset so that individual data cannot be linked with specific individuals [9]. It is a set of techniques and algorithms providing different levels of effectiveness on several kinds of data. It consists of three sub-categories: i) *anonymisation*, which manipulates personal identifiers in such a way as to *irreversibly* remove the link between the individual and the data; ii) *pseudonymisation*, which *reversibly* replaces personal identifiers with artificial ones (pseudonyms); iii) *aggregation*, which replaces the information with a summary. In the HIPAA Privacy Rule,[3] in order to meet privacy requirements, *Personal Health Information* (*PHI*), health information that identifies the associated individual has to be identified and manipulated. To do this, two methods are defined, *Expert Determination* and *Safe Harbor*: the former requires the intervention of an expert, with appropriate knowledge, who assesses the risk that a certain piece of information will enable the identification of the individual; the latter requires the removal of 18 specific pieces of information (e.g., name and social security number). GDPR is rather less concrete: pseudonymisation is recognized as privacy-protective measures that reduce risk to the data subject (Art. 4 and 25), and when processing data, an appropriate level of data protection can be provided by pseudonyms (Art. 32); on the other hand, there is no definition of anonymisation, but Recital 26 states "*[. . .] The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable [. . .]*" and a guidance of anonymisation techniques is issued by the EU expert group Article29 [10].

These regulatory frameworks, especially in Europe, give only abstract instructions without actual operative procedures to apply to the medical data. Thus, the problem of identifying what information has to be de-identified, and how, is left to the solution provider. For this reason, several major challenges arise in the anonymisation and pseudonymisation of FHIR data, such as:

- the risk of re-identification attacks: performing proper anonymisation is a really complex process. Re-identification is a well-known problem in literature and has been analysed in several works (e.g., [4], [6], [11], [12], [13]);
- the complexity of extracting PIIs/PHIs from the FHIR standards and the delicate privacy-utility trade-off: FHIR resources have many different types of IDs and identifiers that serve different purposes. Some of them stand for internal links between resources, others for correlating with external data sources. Strategies for de-identification include how to modify internal IDs consistently, and which external identifiers must be redacted according to their specific meaning and the necessity for processing the de-identified data (FHIR specs about Security[4]);
- keep the pace with HL7 FHIR update and its Ballot Process: balloting is the formal process that HL7 uses to get feedback and comments on specifications prior to publication. Fields may change with the different versions of the standard: Standard for Trial Use (STU), Draft STU (DSTU) or among different releases (e.g., R4). For a long-term project, it is necessary to make a tool upgradeable to easily accommodate FHIR modifications;

[3]https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

[4]https://build.fhir.org/security.html

- subtle pitfalls of pseudonymisation: naive implementation of pseudonymisation, albeit through a state-of-the-art cryptographic hash function (e.g., SHA256) can be disastrous. For instance, if the field belongs to a restricted set such as a date, a brute force attack can easily compute the hash corresponding to all the possible dates in the last 100 years (36500 hashes, achievable in less than a second on any common PC) to fully revert the process. Moreover, the use of inappropriate encryption schemes, such as AES-ECB, or their improper use (e.g., a constant value used as initialisation vector) could make the system vulnerable to attacks, leaking the encryption keys and exposing the protected information. The ''perfectly hiding'' feature must be guaranteed, as well as the other features recommended by international guidelines (e.g., [14], [15], [16]) and advocated by national DPAs [17].

B. TECHNICAL CHALLENGES

It is important to note that the application of de-identification techniques in healthcare data can introduce complexities and challenges for statistical analyses. This may arise from the need to remove or alter certain sensitive information, which could lead to the loss of crucial indicators for research or introduce bias into the results. For instance, the removal of demographic or geographic details could limit the ability to identify health disparities in specific populations, while the substitution of missing values with average data could affect the accuracy of statistical estimates. Therefore, it is crucial to strike a balance between safeguarding patient privacy and preserving data integrity to ensure valid and informative statistical analyses. For example the study presented in [18], illustrates the delicate balance between privacy protection and statistical validity in healthcare data de-identification. The research found that, despite significant differences in value distributions, the statistical bias introduced by anonymization can be kept relatively small. This highlights that careful implementation of anonymization techniques can preserve the statistical validity of data, supporting the need for a nuanced approach that safeguards privacy while maintaining data integrity for meaningful statistical analysis.

Furthermore, there may be situations where it is necessary to protect the privacy of individuals linked to medical data, but to prove that it comes from natural and not artificial sources to validate the results obtained from secondary uses of these data. This is the case, for instance, with the U.S. *Food and Drug Administration* (FDA), which only accepts new drugs if there is a trackback from the clinical trial to a real dataset. Another example is the Lean European Open Survey on SARS-CoV-2 Infected Patients (LEOSS) registry, a European registry used to study the epidemiology and clinical course of COVID-19. To make data available to the public in real-time while supporting evidence-generation at the rapid pace required in a pandemic, quantitative anonymisation procedures have been used to preserve the privacy of patients [19]. More generally, there are many examples in literature for secondary use of healthcare data, for instance for research and applications in quality of care and patient safety [20]. These endeavors frequently entail the implementation of de-identification methods.

In essence, the de-identification of healthcare data presents a formidable challenge, necessitating a delicate balance by specialists between leveraging the data's benefits and mitigating the risks associated with re-identification. Given that this procedure typically relies on manual intervention, the prospect of automated mechanisms for de-identifying health data appears disconcerting and potentially jeopardizes the privacy rights of individuals.

C. OUR CONTRIBUTION

De-identification presents a multifaceted challenge that demands a flexible and interdisciplinary approach. In the realm of healthcare data privacy, the task goes beyond mere anonymization; it necessitates a holistic understanding that transcends traditional boundaries. This challenge requires experts from various fields, including data science, computer security, legal compliance, and healthcare, to collaboratively navigate the complexities of de-identifying data while preserving its utility and integrity. In such a dynamic landscape, flexibility becomes paramount as it enables the adaptation of techniques and strategies to ever-evolving privacy threats and regulations, ultimately ensuring the responsible and effective use of healthcare data.

De-identifying medical data is a team effort between privacy experts (e.g., Data Protection Officers) and Data Engineers. To properly support this process, instead of introducing new de-identification techniques, in this article we present a tool that can facilitate communication between these two worlds implementing the most widely used existing techniques. With FHIR-DIET, our aim is to streamline this collaboration by providing legal experts with a range of technical solutions to carry out the de-identification process, while also offering engineers a system to implement these rules easily. FHIR-DIET is designed to operate locally on the same computer or premises as the data controller/processor, focusing on de-identifying data according to predefined rules. The mechanisms for sharing data post-de-identification are beyond the scope of this work. Our aim is to simplify the de-identification process to enable smoother and more straightforward secondary use of healthcare data.

The main contributions of the paper are:

- an analysis of the literature, identifying what are the most interesting solutions that provide anonymisation and pseudonymisation of FHIR resources (Section II);
- the presentation of FHIR-DIET, an open source solution for handling custom treatment of FHIR resources enabling a fine-grained de-identification process,
- a comparison of FHIR-DIET with other existing solutions (Section II) also providing performance evaluation of a python implementation.

The paper is organised as follows: Section II analyses existing de-identification techniques, tools and platforms; Section III describes our solution, whose implementation is discussed in Section IV; in Section V we provide the performance evaluation and finally, conclusions are drawn in Section VI.

II. RELATED WORK

A. HL7 FHIR

The Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) is a next-generation standard for healthcare data exchange, designed to improve the quality and accessibility of healthcare information. It allows healthcare systems to communicate with each other and share medical data, such as EHRs, clinical data, research data, and administrative data.

FHIR is based on modern web standards, such as RESTful APIs, JSON, and XML, making it suitable for native integration into modern systems so as to simplify the development of healthcare applications. It has been designed to be scalable, enabling the sharing of data across various platforms, ranging from mobile applications to cloud-based systems. FHIR uses a resource-based approach to data modeling, where every piece of medical data is defined as a ''resource''. The standard defines numerous resources to represent real-world concepts in the healthcare system such as Patient, Practitioner, etc. or releted to the healthcare process, comprising Clinical (Allergy, Problem, Procedure, CarePlan/Goal, Family History, RiskAssessment, etc.), Diagnostics (Observation, Report, Specimen, ImagingStudy, Genomics, etc), Medications (Request, Dispense, Statement, Immunization, etc.) and Financial aspects (Claim, Account, Invoice, ChargeItem, etc.) (full list of resource types[5]). FHIR latest specification, v5.0.0 (*R5*), provides 157 different resource types categorised in different sets.

These resources can be easily accessed and manipulated to share medical data across different platforms. FHIR is strongly focused on implementation providing developers with simple interfaces and it is in constant evolution to adapt to different healthcare processes, being as flexible as possible. To ensure consistency and interoperability of medical data, a set of standard vocabularies, *SNOMED CT* (Systematized Nomenclature of Medicine – Clinical Terms), *LOINC* (Logical Observation Identifiers Names and Codes) and *ICD* (International Classification of Diseases), has also been included into the standard. FHIR is being widely adopted by healthcare organizations and technology providers, and has been integrated by many EHR vendors. Also, it is being used by healthcare providers and technology companies, and being used in research initiatives and clinical studies, where interoperability of medical data is crucial. It is a powerful and flexible standard for healthcare data exchange, paving the way for improved patient care, better research, and more efficient healthcare delivery.

[5]https://www.hl7.org/fhir/resourcelist.html

B. DE-IDENTIFICATION TECHNIQUES

The application of anonymisation and pseudonymisation techniques to generic EHRs has been extensively discussed in the literature, and several papers offer an overview of the proposed solutions [21], [22], [23], [24], [25], [26]. For instance, in [27], various pseudonymisation architectures are proposed to foster the secondary use of medical data. Re-identification risk issues have also been extensively addressed in various articles [4], [6], [11], [12], [13], where risk assessment methods are also proposed.

Several works propose and analyse the anonymisation and pseudonymisation of FHIR resources. In [28] Kim et al. present a platform to store and transfer data collected from Parkinson's patients. This data is de-identified by simply removing quasi-identifiers. In [29] Neto et al. present a disease surveillance middleware platform. They propose an anonymisation solution that uses encryption to secure personal information of patients. In [30] Dimopoulou et al. propose a library to anonymise and pseudonymise FHIR data on mobile devices showing how these operations are negligible. The library allows the identifiers and quasi-identifiers to be deleted when they are not required, and to replace them, when they cannot be removed, with random values (anonymisation) and pseudonyms (pseudonymisation). In the last case, the mapping between values to modify and pseudonyms has to be stored. Speaking of techniques we can distinguish two important dimensions. The first is the reversibility and non-reversibility of the procedure, which then maps onto pseudonymization and anonymization techniques. The most widely used techniques for pseudonymization are based on the presence of mapping tables (Trusted Third Party) or encryption/decryption systems. Here, standard symmetric or asymmetric encryption techniques are often used (e.g., AES, RSA, ECIES) with their known strengths and weaknesses [31] and on whose performance there are many dedicated works [32], [33]. Another important dimension relates to techniques that work individually on each piece of data independently, and techniques that, while working on individual data, offer enhanced anonymization features. In fact, the three conditions for ensuring proper anonymization are the prevention of (i) isolating a person in a group (single-out); (ii) linking an anonymized data to data referable to a person in a distinct dataset (linkability); and (iii) inferring new information referable to a person from an anonymized data (inference). This is why techniques such as k-anonymity are used, which offers a ''hiding in the crowd'', and differential privacy [34] that add optimized noise to avoid single-out.

C. FHIR DE-IDENTIFICATION TOOLS AND PLATFORMS

In [35] the author extensively analyses the state-of-the-art and provides a tool, called *Data Privacy Tool*, to de-identify FHIR data. The tool allows users to define the sensitivity of a FHIR field by labeling it as *identifier*, *quasi-identifier*, *sensitive* or *insensitive*. The first one is removed, while the last one remains as it is. For the ones labeled as quasi-identifier or

sensitive, the tool provides different operations according to their data types (e.g., redaction, substitution, generalisation, fuzzing).

Microsoft also proposed its solution, called *FHIR Data Anonymization*, to anonymise FHIR data [36]. It allows users to choose among different operations to modify the information regarded as personal. The tool is accessible via i) command line (can be on-premises or in the cloud), ii) an Azure Data Factory pipeline or iii) operation in the FHIR server for Azure.

Google developed its own solution, called *Cloud Healthcare API*, to support the FHIR ecosystem [37]. Among others, it allows to anonymise FHIR data choosing among different operations. The set of supported operations is much the same as Microsoft's. The solution is meant to be used within Google Cloud.

D. COMPARISON WITH EXISTING SOLUTIONS

In this work, we focus only on de-identification techniques that do not imply any kind of aggregation, detailed in Section III. However, the proposed architecture is extensible, hence other techniques can be added as modules.

With respect to the library presented in [30], our solution is more flexible, providing several operations that allow users to customise the anonymisation/pseudonymisation process according to their needs. With respect to the Data Privacy Tool proposed in [35], our solution does not allow users to label the FHIR fields (a time-consuming operation when the amount of data is large from our point of view); however, users can efficiently select, at the same time, the fields they want to process and the related transformations to apply, using customisable configuration files. Finally, Microsoft's FHIR Data Anonymization tool [36] and Google Healthcare API [37] are designed to operate into their respective clouds and so it is difficult to integrate them into on premises healthcare systems. On the contrary, our solution is cloud-independent since it is designed for native integration with other FHIR services through Web APIs and it can be easy deployed using containers. Moreover, FHIR-DIET allows users to pseudonymise data using external management entities such as GPAS. To the best of our knowledge, it is the first time that a FHIR de-identification solution is cloud-agnostic and supports the use of external tools to pseudonymise its data.

III. FHIR-DIET DESIGN AND ARCHITECTURE

We propose a tool, called *FHIR-DIET*, for de-identification of FHIR data: specifically, the tool provides anonymisation and pseudonymisation. It does not provide aggregation since we are interested in data as individuals and not as aggregated values.

A. DE-IDENTIFICATION PROCESS

In Figure 1 we describe the envisage de-identification process. A close collaboration between lawyers and engineers leads to the identification of the information in the FHIR data that has to be de-identified to respect the legal privacy requirements. This includes the proper selection of the techniques that have to be applied to each field of the FHIR resources and the mapping of such rules into a configuration file. After this preparation phase, the data can be processed by the system, which outputs a de-identified version, e.g., suitable for statistical analysis. These data can be used directly or combined with others to create large datasets that can also be outsourced by the company that generated them and eventually made publicly available.

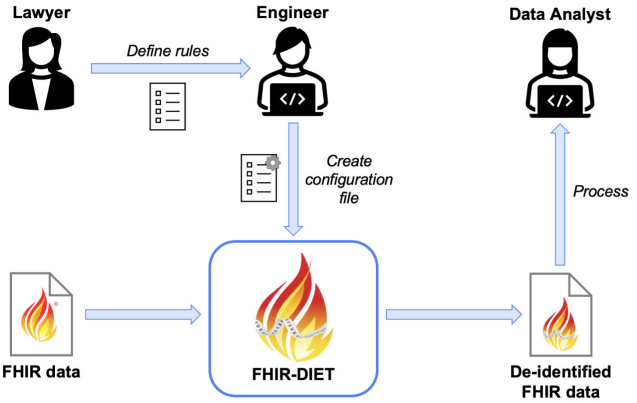


FIGURE 1. De-identification process of the FHIR-DIET tool.

B. ARCHITECTURE

Figure 2 shows a high-level representation of the architecture of our solution. FHIR-DIET accepts FHIR resources as input and returns de-identified/pseudonymized data as output. Resources are processed according to the rules written in a *configuration file* and expressed in the form of a set of *selection-action rules*. The selection is based on rules defined to identify the fields of the FHIR data to process using standard FHIR mechanisms; the actions define the operation to apply and can be chosen in a wide set of transformations such as redaction, perturbation, hashing and substitution for the anonymisation. For pseudonymisation, the system supports both state-of-the-art and advanced cryptographic schemes with perfectly hiding properties as well as external *Trusted Third Party Pseudonym Management Platforms*. The presence of several actions allows more expressiveness in terms of privacy-utility trade-offs so that it can be adapted to different contexts and needs.

The main impact of FHIR-DIET is the ability to generate added-value with a secondary use of data, i.e. using FHIR data for different goals such as health economy and healthcare research, or disease-specific clinical or epidemiological research. In the current EU privacy legislation framework (i.e., GDPR), FHIR-DIET provides data controller, processor and subject important benefits. Indeed, de-identifying data is in the interests of both the data controller/processor, as it reduces the risk of involuntary data disclosure (e.g., data leak), and the data subject, who wants to protect the privacy of her data. Moreover, according to the Recital 26 of GDPR,

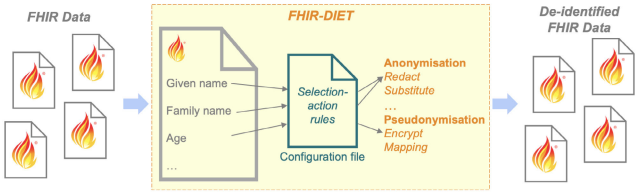


FIGURE 2. FHIR-DIET architecture.

anonymised data is no longer considered personal data, so the regulation does not apply to it.

We point out that FHIR-DIET does not store any kind of FHIR data, so it can be seen as a *function*: people who have access to files containing personal information can use it to de-identify them when necessary. Thus, FHIR-DIET does not provide any access restriction that has to be enforced at the FHIR platform level, i.e., regulating the access to the FHIR data.

C. METHODOLOGY AND APPROACH

Anonymisation and pseudonymisation operations are defined by rules expressed using a selection-action approach: the selection phase identifies the fields of the FHIR resource, while the action phase specifies the operation to perform on such fields. For the selection phase, FHIR-DIET makes use of the navigation system of *FHIRPath*[6] (an expression language defined by FHIR) to select classes of data or individual elements. Such a standardised method to select data ensures full compatibility with the standard while granting the ability to perform flexible filters.

FHIR-DIET supports different operations for each one of the three categories, anonymisation, pseudonymisation and de-pseudonymisation (details in Section III-E). After careful examination, taking into account the most widely adopted techniques in the literature, the supported operations have been chosen to cover most of the desired de-identification behaviors (complete deletion, one-way and two-way obfuscation, numerical perturbation). The selection-action rules are specified through configuration files (e.g., YAML-based) and custom rules can be specified by the user, according to the specific application needs and to extend the compliance, for example, to GDPR.

The correct behavior of the developed operations was tested on official sample FHIR resources[7] and artificially generated FHIR data produced by Synthea,[8] an open-source synthetic patient generator that models the medical history of synthetic patients [38]. The application of de-identification techniques makes data losing informative content, possibly arising biases in the distribution of values and/or limitations to their use. This is a well-know concern and has been address in different works, e.g. [18] where authors demonstrate how the bias is small, preserving validity of statistical results in relatively low-dimensional data.

TABLE 1. Example of FHIRPath application.

FHIRPath string	Result
`age`	`37`
`name.family`	`[Chalmers, Windsor]`

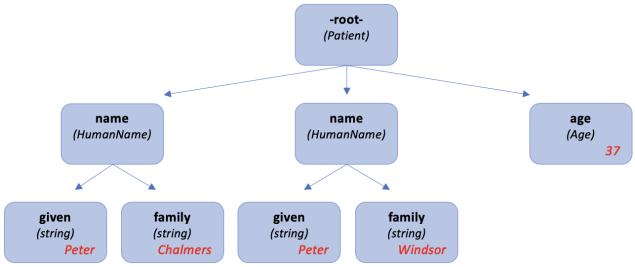


FIGURE 3. Example of the tree of a FHIR resource.

D. SELECTION

During the application of a selection-action rule, the first, selection phase is used to identify the fields of the FHIR resource that have to be altered in the second, action phase. This identification is done using *FHIRPath*, a path-based graph-traversal language developed by HL7. It allows the navigation and extraction of nodes from FHIR resources, almost like *XPath*[9] with XML. Users can easily navigate the tree defining concatenations of node labels to select the desired nodes. Figure 3 shows a (partial) representation of the tree of an FHIR resource: it consists of two `name` nodes and an `age` node; each `name` node contains a `given` node and a `family` node. Table 1 reports some examples of the application of FHIRPath on this FHIR resource. If a user wants to retrieve the age, she has to pass to FHIRPath the navigation string `age`, and it will return the value `37`. Instead, using the navigation string `name.family`, FHIRPath will return the values of all the `family` nodes whose parent node is a `name` node (in the example, the values will be `[Chalmers, Windsor]`).

FHIRPath is an extremely powerful language to navigate and extract data from FHIR resources, and being the official standard released by HL7 itself there are no concerns about compatibility. Both FHIRPath and XPath allow the user to traverse with ease the tree of FHIR and XML resources respectively; however, they are not meant for *updating* the resources. While XML has some unofficial updating languages (e.g., *XUpdate*[10]), to the best of our knowledge, the same does not go for FHIR. Therefore, some ad-hoc solutions have to be implemented.

[6]https://build.fhir.org/ig/HL7/FHIRPath/
[7]https://build.fhir.org/patient-example.html
[8]https://syntheticheath.github.io/synthea/
[9]https://www.w3schools.com/xml/xpath_intro.asp
[10]https://github.com/xuexiangjys/XUpdate

TABLE 2. Supported actions.

Action	Category	Description
`keep`	Anonymisation	Keep the element as is
`cryptohash`	Anonymisation	Apply a hash function on the element (default, SHA3-256)
`perturb`	Anonymisation	Change a numerical value adding a random positive/negative value (available with integer, float and date)
`redact`	Anonymisation	Remove the element
`substitute`	Anonymisation	Substitute the element with a given value
`encrypt`	Pseudonymisation	Encrypt the element using the given cryptographic scheme (default, RSA)
`ttp_gen_list`	Pseudonymisation	Return the list of elements selected by FHIRPath
`ttp_pseudonymize`	Pseudonymisation	Substitute the element with the related pseudonym according to the given pseudonym mapping file
`decrypt`	De-pseudonymisation	Decrypt the encrypted element using the given cryptographic scheme (default, RSA)
`ttp_depseudonymize`	De-pseudonymisation	Substitute the pseudonym with the related original value according to the given pseudonym mapping file

E. SUPPORTED ACTIONS

The list of actions supported by FHIR-DIET are described in Table 2. They are divided into the three categories discussed in Section III-C and are detailed as follows.

1) ANONYMISATION

The system offers the following operations to anonymise the selected data:

- `keep`: leave the data as is,
- `redact`: remove the data,
- `perturb`: add a random noise to the data,
- `cryptoHash`: substitute the data with the digest resulting from the application of a hash function to it,
- `substitute`: substitute the data with a constant value.

FHIR-DIET provides a default configuration that applies to common PHIs found in a FHIR resource. Specifically, the anonynisation operations are applied to 18 identifiers of PHI listed in the Safe Harbor method under the HIPAA Privacy Rule de-identification standard [39].

2) PSEUDONYMISATION AND DE-PSEUDONYMISATION

To support pseudonymisation and de-pseudonymisation in the most generic way, the system provides the two operations:

- `encrypt`: use encryption to substitute data with encrypted one. In this way it is possible, if provided with the decryption key, to reverse the process;
- `mapping`: substitute data with dynamically generated values, keeping track of the association to reverse the operation.

Both the methods replace the considered data with an artificial identifier and a de-pseudonymisation procedure can be used to invert the process.

Specifically, the `encrypt` action uses cryptographic methods to implement pseudonymisation. FHIR-DIET allows users to choose among several cryptographic schemes,

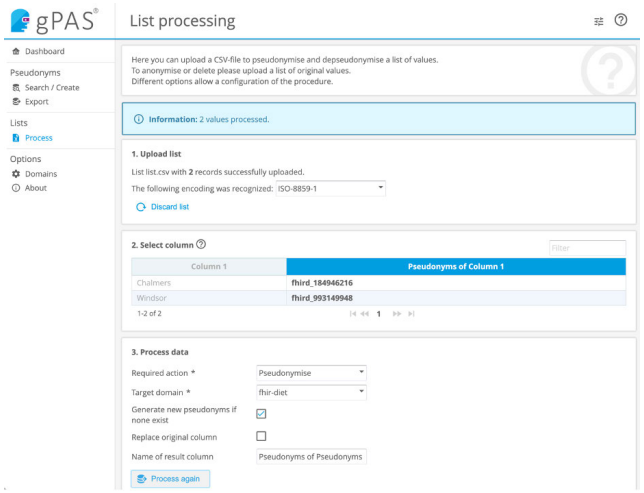


FIGURE 4. Example of interaction with gPAS: i) a file containing the strings whose pseudonyms are required is uploaded on gPAS; ii) the column with the strings involved is selected; iii) the required action (e.g., pseudonymise), the domain used for the generation of pseudonyms and other options are set; iv) the pseudonyms are generated and the resulting file can be downloaded.

all of which offer a perfectly hiding property and can be fed to the tool through the configuration files. In particular, following the guidelines [14], the choice is around literature approaches based on asymmetric cryptographic algorithms (e.g., [40]) or more recent approaches based on cryptographic accumulators for pseudonym generation [41], which inherits the security properties of a Merkle tree (also envisioned by the recent ENISA report [15]) achieving post-quantum security. The reference is the ENISA best practices [14] and techniques [15] and the ISO 25237 standard on pseudonymisation and specifically the Clause 6 [16].

The `mapping` action substitutes a personal identifier with dynamically-generated values, keeping track of the association to reverse the operation. Using the very same

TABLE 3. Example of pseudonyms generated by gPAS.

Original	Pseudonym
Chalmers	fhird_184946216
Windsor	fhird_993149948

actions (described in the following), pseudonym management can be addressed locally or can be passed to external entities. For example, a popular TTP Pseudonym Management Platform that currently manages over 22 million pseudonyms is *gPAS*.[11] FHIR-DIET integrates with gPAS to manage the matching between pseudonyms and identifiers.

a: PSEUDONYMISATION WITH TRUSTED THIRD PARTY

The idea is to use a Trusted Third Party (TTP) Pseudonym Management Platform (e.g., gPAS) providing the mapping between pseudonyms that it generates and their related original values. Our approach to the integration is based on the three actions, `ttp_gen_list`, `ttp_pseudonymize`, `ttp_depseudonymize`, detailed in the following.

Generate list (`ttp_gen_list`)

This action will generate a file with the data to be pseudonymised. For instance, following the example described at the end of Section III-D, using the selection string "*name.family*", we obtain a file containing `Chalmers` and `Windsor`.

As shown in Figure 4, we can upload this file on gPAS to obtain a list of pseudonyms related to the given original values. Finally, this list can be downloaded as a mapping file, whose content appears like the one shown in Table 3.

Pseudonymise (`ttp_pseudonymize`)

This action uses a mapping file to apply pseudonymisation to a FHIR resource: in particular, the data selected in the selection phase is replaced with the related pseudonyms contained in the mapping file.

De-pseudonymise (`ttp_depseudonymize`)

This action uses a mapping file to reverse the pseudonymisation on a FHIR resource: in particular, the data (pseudonyms) selected in the selection phase is replaced with the related original values contained in the mapping file.

IV. PROOF OF CONCEPT

We developed a PoC of the system using Python available on GitHub[12] and released it as open-source.

A. ARCHITECTURE

Figure 5 shows the architecture of the developed PoC, detailing the components and the adopted technologies. The service is accessible through one simple API, `process`, which, according to the required action, triggers one of the three main APIs: `anonymize`, `pseudonymize` and `de-pseudonymize`. All the APIs operate on JSON data and are provided through OpenAPI definition, using Swagger.

We provide the system as a module that can be installed on-premises locally or as a microservice on a private infrastructure. It allows users to run a Web service whose REST APIs are implemented using FastAPI[13] (Figure 6). An example of execution of the Web interface is shown in Figure 7. Since the de-identification operations to be performed are specific to the type of data and the type of use to be made of it, we allow this information to be written in a configuration file. This file is loaded at the start of the service and used for each data processing request. The deployment of the system is available also through a Docker container: the containerisation technology allows the tool to be easily pluggable in deployment configuration either in Cloud or inside orchestrators, such as Kubernetes, or other microservice architectures.

The service is also accessible through *command line interface* (CLI) as shown in Figure 8.

B. CONFIGURATION FILE

Configuration files are used to define the selection-action rules to apply to the given FHIR resources. Each configuration file can contain multiple rules expressed using the YAML[14] format. Figure 9 shows an example of a configuration file with multiple selection-action rules: the *match* label specifies the string to use in the selection phase, while the *action* label specifies the action to perform on the selected data on the FHIR resource; the *params* label contains additional parameters required to perform the desired action.

V. PERFORMANCE EVALUATION

FHIR-DIET provides two modes to access the service: Web interface and CLI. To assess the performance of the Web Interface we resort to the measure of conventional metrics for Web Services such as requests processed per second (req/s), throughput, and latency. Concerning the CLI usage, we show the number of operations executed per second. All the tests have been conducted on a commodity laptop (Intel i7, quad-core, 16GB RAM).

A. WEB INTERFACE

The performance of the Web interface was evaluated using stress/load tests executed by *bombardier*,[15] a HTTP(S) benchmarking tool written in Go. The Web interface was running on a *Uvicorn*[16] web server, and the execution of bombardier has been performed with the following parameters:

- maximum number of concurrent connections: *125*,
- request timeout: *2 seconds*,
- test duration: *60 seconds*.

[11] https://www.ths-greifswald.de/en/researchers-general-public/gpas/
[12] https://github.com/docunque/fhir-diet
[13] https://fastapi.tiangolo.com/
[14] https://en.wikipedia.org/wiki/YAML
[15] https://github.com/codesenberg/bombardier
[16] https://www.uvicorn.org/

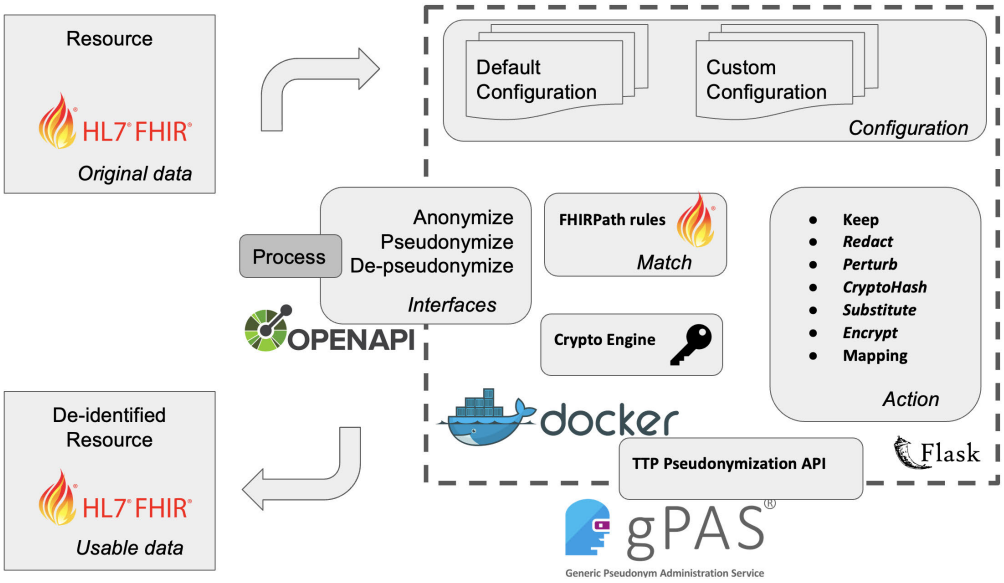


FIGURE 5. PoC Architecture.

FIGURE 6. Web interface.

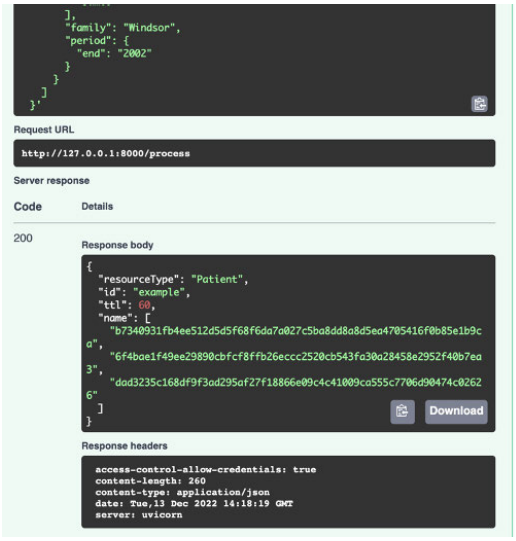


FIGURE 7. Example of usage.

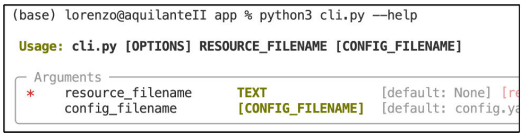


FIGURE 8. Command line interface.

Table 4 shows the results of the tests. The tool is able to handle around 400 requests per second. Obviously, it depends on the complexity of the operations to perform (for these tests, we asked the tool simply to perform the hash of the name of the patient) and on the computational capacity of the server. Still, the results provide insight into the feasibility of using the system in Cloud contexts, within data pipelines and parallel executions, to process data even in near-real time.

B. CLI

The performance of the CLI was evaluated using the script *kpi.py* available in the GitHub repository. The script computes the execution time of a specified anonymisation/pseudonymisation operation applied to the given FHIR data file. The final execution time is calculated as an average over 100 executions. Finally, we invert the execution time, obtaining the speed of execution of the anonymisation and pseudonymisation in terms of *operation*/*sec* (s^{-1}). Table 5 shows the resulting performance. Compared with the Web interface, it is evident that the CLI performs worse. This is mainly due to the overhead introduced by the creation of a process to execute the single execution of the CLI test. In fact, while the Web interface is made available by the Web server

```yaml
general:
  appname: FHIR-DIET
rules:
  - match: Patient.name
    action: cryptohash
    params:
      hash_type: sha3_256
  - match: Patient.birthDate
    action: perturb
    params:
      max: 10
      min: -5
  - match: Patient.address
    action: encrypt
    params:
      algorithm: RSA
      public_key: test/keys/id_rsa.pub
  - match: Patient.address
    action: decrypt
    params:
      algorithm: RSA
      private_key: test/keys/id_rsa
  - match: Patient.id
    action: substitute
    params:
      substitute_with: foo
```

FIGURE 9. Example of a configuration file with multiple rules.

TABLE 4. Performance of the Web interface.

Statistics	Avg	Stdev	Max
Reqs/s	393.22	384.23	2768.95
Latency [s]	0.31915	0.08619	1
Statistics	**Value**		
HTTP codes	1xx - 0, 2xx - 23559, 3xx - 0, 4xx - 0, 5xx - 0, others - 0		
Throughput [KB/s]	398.64		

TABLE 5. Performance of the CLI.

Operation	Avg execution speed $[s^{-1}]$
Anonymisation	2.63
Pseudonymisation	2.93

running thanks to the underlying, always-on process, a new process must be created each time the CLI is used, increasing the time required to complete the execution.

VI. CONCLUSION

Current regulatory frameworks, especially in Europe, do not provide explicit instructions on how to process the personal information contained in healthcare data, making the operative part of processing them very complicated and non-standardised, often requiring the intervention of a human expert. What can be done currently is to simplify the operative part after identifying the data to be processed and their processing mode by providing support tools to perform the selected operations in their desired modalities. This provides a high degree of flexibility and could become a valuable tool in the hands of users for de-identification on FHIR. We believe FHIR-DIET can help by providing a valuable tool to meet current legislative requirements and offer added-value for the secondary use of healthcare data. In this paper we presented our solution, describing its architecture and implementation. FHIR-DIET facilitates collaboration between legal and IT experts in defining de-identification rules to be applied automatically to FHIR data, using a human-readable syntax. It offers various actions for data processing, encompassing a range of desirable behaviors. Specifically, anonymisation and pseudonymisation operations are supported and can be configured through flexible parameters. Additionally, the pseudonymisation process can be integrated with external Trusted Third Party Pseudonym Management Platforms, such as gPAS. We evaluated the performance of the tool using stress/load tests, and the results show that the tool can provide efficient anonymisation and pseudonymisation fostering the secondary use of medical data. The code of the PoC is released as open-source on GitHub as already stated in Section IV.

REFERENCES

[1] S. Zenker, D. Strech, K. Ihrig, R. Jahns, G. Müller, C. Schickhardt, G. Schmidt, R. Speer, E. Winkler, S. G. von Kielmansegg, and J. Drepper, "Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio)medical research: Towards a new German national standard," *J. Biomed. Informat.*, vol. 131, Jul. 2022, Art. no. 104096.

[2] J. Yoon, L. N. Drumright, and M. van der Schaar, "Anonymization through data synthesis using generative adversarial networks (ADS-GAN)," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2378–2388, Aug. 2020.

[3] A. Roehrs, C. A. da Costa, R. da Rosa Righi, S. J. Rigo, and M. H. Wichman, "Toward a model for personal health record interoperability," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 867–873, Mar. 2019.

[4] Y. J. Lee and K. H. Lee, "What are the optimum quasi-identifiers to re-identify medical records?" in *Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2018, pp. 1025–1033.

[5] J. Henriksen-Bulmer and S. Jeary, "Re-identification attacks—A systematic literature review," *Int. J. Inf. Manage.*, vol. 36, no. 6, pp. 1184–1192, Dec. 2016.

[6] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PLoS ONE*, vol. 6, no. 12, Dec. 2011, Art. no. e28071.

[7] A. Antoniou, G. Dossena, J. MacMillan, S. Hamblin, D. Clifton, and P. Petrone, "Assessing the risk of re-identification arising from an attack on anonymised data," 2022, *arXiv:2203.16921*.

[8] R. Ratra, P. Gulia, and N. S. Gill, "Evaluation of re-identification risk using anonymization and differential privacy in healthcare," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, pp. 563–570, 2022.

[9] S. Garfinkel, "De-identification of personal information," U.S. Dept. Commerce, NIST, Comput. Secur. Division, Inf. Technol. Lab., Gaithersburg, MD, USA, 2015.

[10] Article 29 Data Protection Working Party. (2014). *Opinion 05/2014 on Anonymisation Techniques*. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[11] M. Scaiano, G. Middleton, L. Arbuckle, V. Kolhatkar, L. Peyton, M. Dowling, D. S. Gipson, and K. E. Emam, "A unified framework for evaluating the risk of re-identification of text de-identification tools," *J. Biomed. Informat.*, vol. 63, pp. 174–183, Oct. 2016.

[12] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 2, pp. 169–177, Mar. 2010.

[13] K. El Emam, F. K. Dankar, A. Neisa, and E. Jonker, "Evaluating the risk of patient re-identification from adverse drug event reports," *BMC Med. Informat. Decis. Making*, vol. 13, no. 1, pp. 1–14, Dec. 2013.

[14] ENISA. (2019). *Pseudonymisation Techniques and Best Practices*. [Online]. Available: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices/

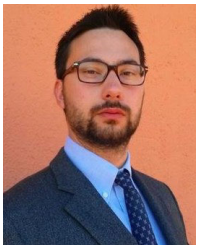
[15] ENISA. (2021). *Data Pseudonymisation: Advanced Techniques and Use Cases*. [Online]. Available: https://www.enisa.europa.eu/publications/data-pseudonymisation-advanced-techniques-and-use-cases/

[16] ISO. (2017). *ISO 25237:2017 Health Informatics—pseudonymization*. [Online]. Available: https://www.iso.org/standard/63553.html

[17] AEPD & EDPS. (2019). *Introduction To the Hash Function As a Personal Data Pseudonymisation Technique*. [Online]. Available: https://edps.europa.eu/data-protection/our-work/publications/papers/introduction-hash-function-personal-data_en
[18] C. E. Koll et al., "Statistical biases due to anonymization evaluated in an open clinical dataset from COVID-19 patients," *Sci. Data*, vol. 9, no. 1, p. 776, 2022.
[19] C. E. Jakob, F. Kohlmayer, T. Meurers, J. J. Vehreschild, and F. Prasser, "Design and evaluation of a data anonymization pipeline to promote open science on COVID-19," *Sci. Data*, vol. 7, no. 1, p. 435, 2020.
[20] D. R. Schlegel and G. Ficheur, "Secondary use of patient data: Review of the literature published in 2016," *Yearbook Med. Informat.*, vol. 26, no. 1, pp. 68–70, Aug. 2017.
[21] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," *Trans. Data Privacy*, vol. 7, no. 3, pp. 337–370, 2014.
[22] M. Simi, K. S. Nayaki, and M. S. Elayidom, "An extensive study on data anonymization algorithms based on k-anonymity," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 225, Jan. 2017, Art. no. 012279.
[23] C. A. Kushida, D. A. Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin, "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies," *Med. Care*, vol. 50, no. 2, p. S82, 2012.
[24] S. Gokila and P. Venkateswari, "A survey on privacy preserving data publishing," *Int. J. Cybern. Informat.*, vol. 3, no. 1, pp. 1–8, 2014.
[25] B. Eze and L. Peyton, "Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing," *Proc. Comput. Sci.*, vol. 63, pp. 348–355, Jan. 2015.
[26] I. E. Olatunji, J. Rauch, M. Katzensteiner, and M. Khosla, "A review of anonymization for healthcare data," *Big Data*, Mar. 2022.
[27] K. Pommerening and M. Reng, "Secondary use of the EHR via pseudonymisation," in *Medical and Care Compunetics 1* (Studies in Health Technology and Informatics). Amsterdam, The Netherlands: IOS Press, 2004, pp. 441–446.
[28] D.-Y. Kim, S. Hwang, M.-G. Kim, J.-H. Song, S.-W. Lee, and I. K. Kim, "Development of Parkinson patient generated data collection platform using FHIR and IoT devices," in *Proc. MEDINFO*, vol. 245, 2017, p. 141.
[29] S. Neto, F. S. Ferraz, and C. A. G. Ferraz, "Towards identity management in healthcare systems," in *Proc. Int. Conf. Internet Comput. (ICOMP) World Congr. Comput. Sci.*, 2016, p. 157.
[30] S. Dimopoulou, C. Symvoulidis, K. Koutsoukos, A. Kiourtis, A. Mavrogiorgou, and D. Kyriazis, "Mobile anonymization and pseudonymization of structured health data for research," in *Proc. 7th Int. Conf. Mobile Secure Services (MobiSecServ)*, Feb. 2022, pp. 1–6.
[31] G. Narula, B. Gandhi, H. Sharma, S. Gupta, D. Saini, and P. Nagrath, "A novel review on healthcare data encryption techniques," in *Proc. ICICC*, vol. 3. Cham, Switzerland: Springer, 2022, pp. 489–498.
[32] P. Dixit, A. K. Gupta, M. C. Trivedi, and V. K. Yadav, "Traditional and hybrid encryption techniques: A survey," in *Networking Communication and Data Knowledge Engineering*. Cham, Switzerland: Springer, 2018, pp. 239–248.
[33] E. Raso, L. Bracciale, P. Gallo, G. Bernardinetti, G. Bianchi, E. R. Sanseverino, and P. Loreti, "Performance evaluation of cryptographic schemes for blockchain security of smart grids," in *Proc. Workshop Blockchain Renewables Integr. (BLORIN)*, Sep. 2022, pp. 113–117.
[34] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang., Program.* Cham, Switzerland: Springer, 2006, pp. 1–12.
[35] E. Şimşek Yılgın, "Preserving privacy of health data residing in HL7 FHIR repositories through de-identification," M.S. thesis, Dept. Comput. Eng., Middle East Tech. Univ., Ankara, Turkey, 2022.
[36] Microsoft. (2022). *Tools for Health Data Anonymization*. [Online]. Available: https://github.com/microsoft/Tools-for-Health-Data-Anonymization
[37] Google. (2021). *API Cloud Healthcare*. [Online]. Available: https://cloud.google.com/healthcare-api?hl=it
[38] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan, "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 3, pp. 230–238, Mar. 2018.
[39] (2012). *Guidance on De-Identification of Protected Health Information*. [Online]. Available: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
[40] J. Lehnhardt and A. Spalka, "Decentralized generation of multiple, uncorrelatable pseudonyms without trusted third parties," in *Proc. Int. Conf. Trust, Privacy Secur. Digital Bus.* Cham, Switzerland: Springer, 2011, pp. 113–124.
[41] G. Kermezis, K. Limniotis, and N. Kolokotronis, "User-generated pseudonyms through Merkle trees," in *Proc. Annu. Privacy Forum*. Cham, Switzerland: Springer, 2021, pp. 89–105.

EMANUELE RASO received the M.S. degree in computer science engineering from the Department of Civil Engineering and Computer Science Engineering, University of Rome Tor Vergata, in 2019, where he is currently pursuing the Ph.D. degree. He was a Researcher for the EU H2020 "BRP4GDPR" Project. His research interests include cybersecurity, particularly in applied cryptography, data privacy, and confidentiality.

PIERPAOLO LORETI is currently an Associate Professor of telecommunications with the University of Roma Tor Vergata. His research interests include wireless and mobile networks, the IoT systems and platforms, framework design, analytic modeling, and performance evaluation through simulation and test-bedding.



MICHELE RAVAZIOL is currently a Medical Doctor with more than 20 years of experience. He is the Chief Executive Officer with Docunque SRL, a company that develops innovative healthcare technology with a new kind of medical practice management software.

LORENZO BRACCIALE is currently an Assistant Professor with the Department of Electronic Engineering, University of Rome Tor Vergata. His research interests include distributed systems, communication systems, and privacy-preserving technologies.

Open Access funding provided by 'Università degli Studi di Roma "Tor Vergata"' within the CRUI CARE Agreement