Facilitating health and academic research through a pseudonymisation service mutualizing the Luxemburgish eHealth platform and its identity management IT infrastructure

AGENCE eSanté LUXEMBOURG

Agence nationale
des informations partagées
dans le domaine de la santé

Raffaella Vaccaroli, PhD *
Francois Wisniewski *
Christophe Pinon*
Philippe Delebarre Debay*
Frédéric Markus *
Heiko Zimmermann *
Hervé Barge*

* Agence eSanté G.I.E., Agence nationale des informations partagées dans le domaine de la santé

List of abbreviations

CCPA	California Consumer Privacy Act
ENISA	European Union Agency for Cybersecurity
EU	European Union
SFTP	Secure File Transfer Protocol
GDPR	General Data Protection Regulation
IHE	Integrating the Healthcare Enterprise
IHE-PIX	IHE-Patient Identifier Cross-Referencing
IS	Information System
ISO	International Organization for Standardization
IT	Information Technology
MPI	Master Patient Index
MPI-SPS	Master Patient Index for the SPS
OID	Object IDentifier
SOAP	Simple Object Access Protocol
SPS	Service de Pseudonymisation en Santé / Health Pseudonymisation Service
SSH	Secure SHell protocol
SSL	Secure Sockets Layer standard
TTP	Trusted Third Party
USA	United States of America
WS	Web Service
WSDL	Web Services Description Language

Abstract

Background: Since its inception, the European General Data Protection Regulation (GDPR) has imposed to data controllers the use of data protection safeguards in order to process personal data, also known as "identifiable data" [1]. Anonymisation and pseudonymisation are two of the most discussed techniques responding to these requirements. Anonymisation foresees the permanent removing of personally identifiable information from data sets. In contrast, pseudonymisation occurs by replacing identification traits with fabricated identifiers such as random alphanumeric strings. The advantage of the latter technique lies in its reversibility. Pseudonymisation not only secures the conformity to GDPR requirements, but also prepares systems to the inevitable handling of big data from the fields of research and health.

Here, we aim to describe the implementation of a service dedicated to identity pseudonymisation in the health sector. This pseudonymisation service mutualizes the Luxemburgish eHealth identity management IT infrastructure.

Methods: With this publication we describe the modifications that have been implemented into the national eHealth platform in order to respond to the growing interest of research as well as medical institutions in regards to the digital exchange of de-identified health data in Luxembourg. Firstly, the implementation of a service able to perform pseudonymisation within various contexts (Biobanks, national registries, academic research) [2, 3]. Secondly, as identity management represents a prerequisite for pseudonymisation, the connection of this service to the national master patient index.

Results: With this paper, we describe the application of the approach proposed by Roth [2-4] in the establishment of a pseudonymisation service mutualizing the national eHealth IT infrastructure.

Conclusion: The described service allows for a national experience of collaborative convergence between the fields of healthcare and academic research and the national IT eHealth platform, with the aim to mutualize efforts in order to make progress in translational research. Although the implementation is a laborious process, the resulting privacy by design approach fosters the potential secondary use of big data in health as well as in biomedical research. Indeed, pseudonymisation, when supported by a reliable identity management system, offers the possibility to conduct larger scale investigations by aggregating data from different sources as well as from distributed data bases.

Keywords: Pseudonymisation, eHealth, Data protection, Privacy, Anonymisation, Re-identification, General Data Protection Regulation, GDPR

Background

Anonymisation and pseudonymisation of personal health data

Since its adoption in 2018, the GDPR of the European Union (EU) strongly encourages the application of data protection safeguards while processing personal health data such as genetic data (defined by Art 4.13), biometric data (defined by Art 4.14) or more generally data concerning health (defined by Art 4.15) [1]. Anonymisation and pseudonymisation represents the most popular techniques applied within this scope.

The need of privacy protection measures, allowing for a secure and ethic way in the handling and use of health data, is growing in the scientific community. However, there is a certain level of confusion in perceiving and applying the concepts of anonymisation and pseudonymisation as well as in their respective applications. Although this is not the case, pseudonymised data are commonly perceived as anonymous data. While being collected, datasets are composed by two linked outputs: the identity traits of the person and the medical data that can relate directly or indirectly to a person. Thanks to the association of identity and medical data, a dataset is assigned to a specific person. The process of isolating permanently medical from the related identity traits is defined as de-identification (anonymisation). Differently, pseudonymisation consist in a transitory de-identification, thus enabling a potential future re-identification of a specific identity.

Anonymisation: Anonymisation is a privacy protection method that enables the permanent removal of the link between identity traits and health information, making data no longer identifiable. Once data is correctly anonymised, any possibility of direct or indirect re-identification is impossible. As a result, anonymised data are no longer considered personal data under the GDPR, which applies to all "processing of personal data" (Art 1.1 GDPR).

However, while legally the understanding of anonymisation is simple, the transposition of its conditions in practical and technical terms remains sensitive. The requirement of an irreversible break in re-identification is hampered by the changing nature of technologies, making anonymisation an imperfect process. For this reason, the procedures implemented by the data controller require constant review with the aim of improvements in anonymisation techniques.

To anonymise data, there are several techniques available that were listed in the G29 guidelines in 2014 [5] and remain relevant depending on use cases. These techniques can be categorized into two families: randomization and generalization. Randomization techniques aim to alter the accuracy of data by adding noise, permutation, or differential privacy. Generalization techniques dilute data by generalizing its attribute, and include k-anonymity and i-diversity.

It is important to articulate these techniques together and test them upstream to prevent cumulatively three categories of anonymisation risks: individualization, correlation, and inference. Individualization involves isolating in a dataset at least some identifying part of an individual. Correlation involves extracting two records relating to the same concerned person or group of people. Inference involves deducing with high probability an attribute from a set of attributes.

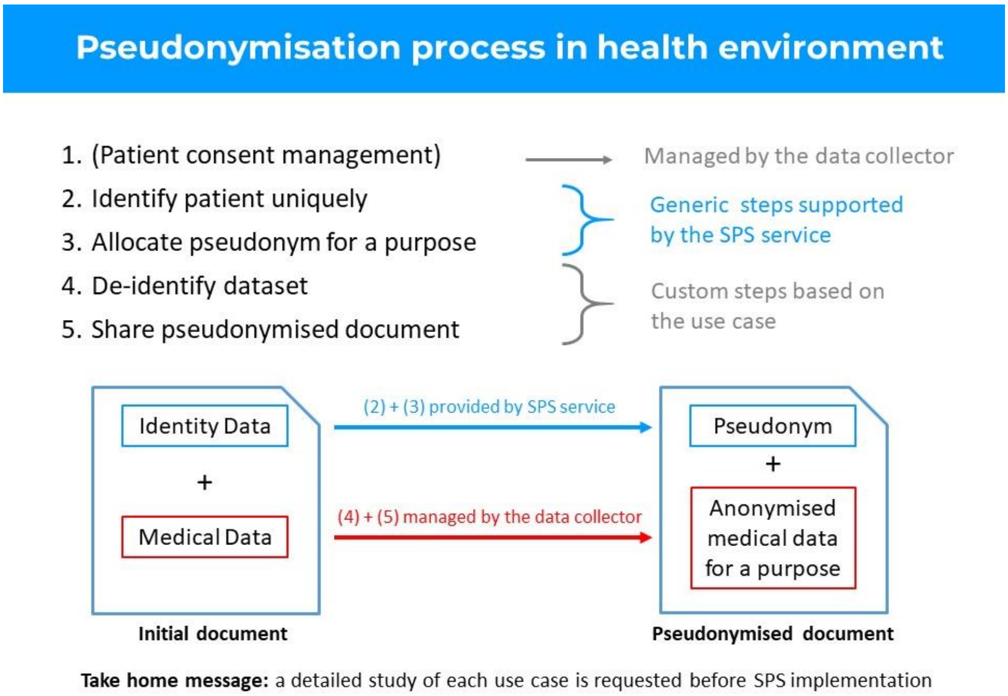
However, when dataset are anonymised in silo, there is no means to make cross-source or cross-study analysis of data as we cannot identify that some data relate to the same patient. Therefore, as aggregating anonymised data from different sources is not possible, and re-identification is not possible in case of interesting discovery that may save life, anonymisation is often discarded in the health sector in favor of pseudonymisation.

Overall, while anonymisation is a powerful privacy protection measure, its successful implementation requires careful consideration of different anonymisation techniques and regular technical reviews.

Pseudonymisation: Pseudonymisation is a technique used to protect personal identity traits by replacing them with pseudonyms that obscure the original identity. In this technique, identity traits are not permanently deleted but stored separately in secure servers by a trusted third party (TTP) [6]. The TTP is responsible for properly storing and protecting the confidential data for the long-term. However, since the link between identity traits and personal data is not permanently altered, re-identification is still possible, making pseudonymised data subject to GDPR regulations. The implementation of a pseudonymisation system is more complex and indirect than an anonymisation system, as it requires more technical and organizational measures, such as database encryption and access control procedures, to ensure GDPR compliance. Despite the challenges, pseudonymisation is highly recommended and required in the healthcare and academic sectors, as it allows for re-identification while protecting personal data.

Here, we aim to describe the implementation in a national eHealth platform of an identity pseudonymisation service, SPS (from French Service de Pseudonymisation en Santé), in order to respond to the growing interest of research as well as medical institutions in regards to the digital exchange of health data in Luxembourg. A service able to perform pseudonymisation within various contexts (Biobanks, national registries, academic research, etc.) has been implemented [2, 3] and reuse the identity management server of the national eHealth platform to firstly link patient identity coming from diverse source and provide after correlated pseudonyms.

Image 1: Pseudonymisation process in health environment



Pseudonymisation process in health environment

1. (Patient consent management)
2. Identify patient uniquely
3. Allocate pseudonym for a purpose
4. De-identify dataset
5. Share pseudonymised document

Managed by the data collector

Generic steps supported by the SPS service

Custom steps based on the use case

Identity Data
+
Medical Data
Initial document
(2) + (3) provided by SPS service
(4) + (5) managed by the data collector
Pseudonym
+
Anonymised medical data for a purpose
Pseudonymised document

Take home message: a detailed study of each use case is requested before SPS implementation

Methods

Identity management through the national master patient (MPI)

The initial version of the pseudonymisation service was developed without incorporating identity management. In this first version, it was discovered that there was a pseudonym duplication rate of 30%, indicating that multiple pseudonyms corresponded to the same patient instead of the expected unique identification. For this reason, we consider that a correct identification of the patient was a prerequisite for a pseudonymisation service. Since 2014, the Luxemburgish eHealth platform relies on a federated identity management model based on a national master patient index (MPI). This MPI enables to link various health related local personal identifiers of a same identity to a unique federal identifier using the IHE Patient Identifier Cross Referencing (PIX) integration profile [7, 8]. For the implementation of the pseudonymisation the national MPI is re-used to identify patients: firstly, to match the identity provided by the source to one federal identifier and secondly to provide to the SPS service one dedicated unique identifier specific to the pseudonymisation service.

Each time the SPS service is called, the MPI's matching algorithm performs an identity analysis. More precisely, the demographic data of patients are compared with the identity traits known by the national MPI to determine if the concerned identity relates to a known patient. In the case the MPI find no match or several matches, a new identity is created. When a patient benefit from the first time of the pseudonymisation service, a local identifier specific to the pseudonymisation service is generated, stored in the MPI and shared to the SPS. When a patient is included into a second use case applying the pseudonymisation service, e this SPS local identifier is shared to the SPS to identify the patient uniquely. The national MPI (Master Patient Index) is only aware that the patient is utilizing the SPS service through the local identifier, and is unaware of their participation in a specific use case. The responsibility of generating one or multiple pseudonyms for the patient's identity lies with the SPS service.

Using the MPI (Master Patient Index) has an additional benefit, as the SPS benefits of the national identity monitoring unit (*Cellule Nationale d'Identitovigilance*) work on identity management. For instance, if a duplicate identity is detected and the identities are merged, the MPI will notify the SPS. The SPS can then merge the pseudonyms and inform the pseudonym consumers about the merge. Similarly, every time the SPS (Secure Pseudonymisation Service) shares a pseudonym, it is linked to a random version control number that corresponds to the de-identification source. This process of versioning the pseudonyms helps to handle any potential collisions of patient identity. In case a data source makes an error in identifying a patient, the SPS may unintentionally provide the pseudonym of another patient. However, when a collision is detected, the version control allows for the separation of the pseudonymised dataset in the research institution, and facilitates the process of correcting the data.

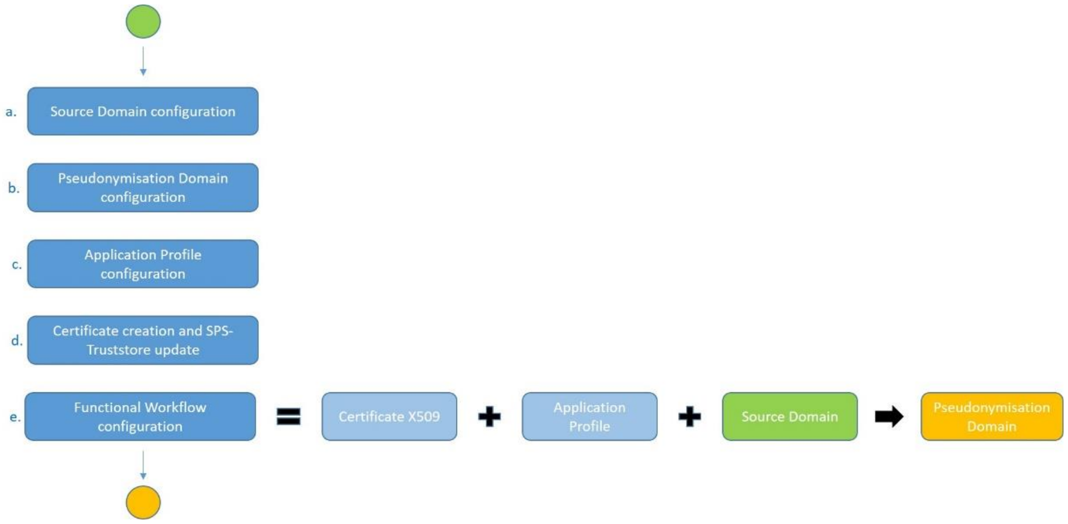
The pseudonymisation service (SPS) enables the workflow set up for the creation of pseudonyms.

The correct patient identification within the MPI is a fundamental step to initiate the pseudonymisation process. However, it is important to underline that the concrete production of a pseudonym does not occur within the MPI but in a separate functional element called SPS. This service create one or several pseudonyms for one identity depending of the need of the use case of the research institutions that are handling medical data. The role of the administrators, is to support the implementation phase. The SPS has a role of Trusted Third Party (TTP), since it is able to stock the link between a unique the SPS local identifier and various pseudonyms.

Three components are fundamental to the functioning of the SPS: the identification domain, the application profile and the functional workflow. First, the identification domain identifies legal entities. The SPS interface allows for the creation of identification domains corresponding to institutions and/or applications that must share data through the use of the pseudonym. Two different identification domains are available in the SPS: a source domain where the patient has been identified, and a pseudonymisation domain where pseudonyms are generated (**Image 2a** and **2b**). The format of the pseudonym must be defined according to the project specifications. In the SPS service, each identification domain is identified by a unique object identifier (OID) that is generated from the interoperability department of the Agence eSanté according to the national standards. Second, the application profile enables the creation of group of rights for different web services (WS) **(Table 1) (Image 2c)**. The application profile are assigned to user, identified by X509 certificates **(Image 2d)**, to authorize the user for specific actions.

A pseudonymisation workflow is the association of a user certificate with one application profile and two identification domains (one source domain and one pseudonymisation domain) (**Image 2d**). The functional workflow outlines the actions that specific users are authorized to carry out within the identification domain. For example, in most cases, there are separate workflows for the data source that requests de-identification of a patient and the data collector who retrieves the pseudonyms. The configuration of the SPS (Secure Pseudonymisation Service) is not accessible to service users, and only administrators have access to configure workflows according to the specific use case requirements.

Image 2: Diagram of the SPS channel configuration. The diagram illustrates each step required for the setting up of a pseudonymisation channel. (a) A source domain has to be created and configured. The source domain provides the patients' demographics. (b) Creation of the pseudonymisation domain that generates the pseudonym. (c) Creation and configuration of the application profiles required for the different identification domains, which are integrated in the channel. (d) A certificate is generated by the user and added in the truststore specific to the SPS, which enables the identification and authorization of each user. (e) Set up of the functional channel linking the different identification domains to the specific application profiles.


a.
Source Domain configuration
b.
Pseudonymisation Domain configuration
c.
Application Profile configuration
d.
Certificate creation and SPS-Truststore update
e.
Functional Workflow configuration
=
Certificate X509
+
Application Profile
+
Source Domain
Pseudonymisation Domain

Securisation of the SPS work flow

The fact that the service structure allows for the distribution of data between separate functional elements (MPI and SPS), leads to a better security in the event of a potential intrusion on one of the servers.

On one hand, the MPI (Master Patient Index) stores the demographic information along with the associated SPS (Secure Pseudonymisation Service) local identifier. On the other hand, the SPS stores the SPS local identifier and pseudonyms. Unlike some TTP (Third-Party Providers) that store a correspondence table containing patient demographics and pseudonyms in one place, an intrusion into one system does not provide sufficient information to re-identify a patient.Securisation of the whole SPS work flow happens through X509 certificates responding to the Secure Sockets Layer (SSL) standards. A certificate is generated by the certification authority (*Agence eSanté*) and added in the truststore specific to the SPS. For each certificate, an identification and authorisation of each identification domain is performed.

The de-identification service starts with the traceability, authentication and authorization functions. In order to enable the traceability, the recording of the trace is the first action carried out in the service. All accesses to the SPS services are therefore traced whether authorized or not. Each service accessed is subject to a user authentication as well as authorization. This authentication and authorization management is dedicated to an ad hoc authorization server.

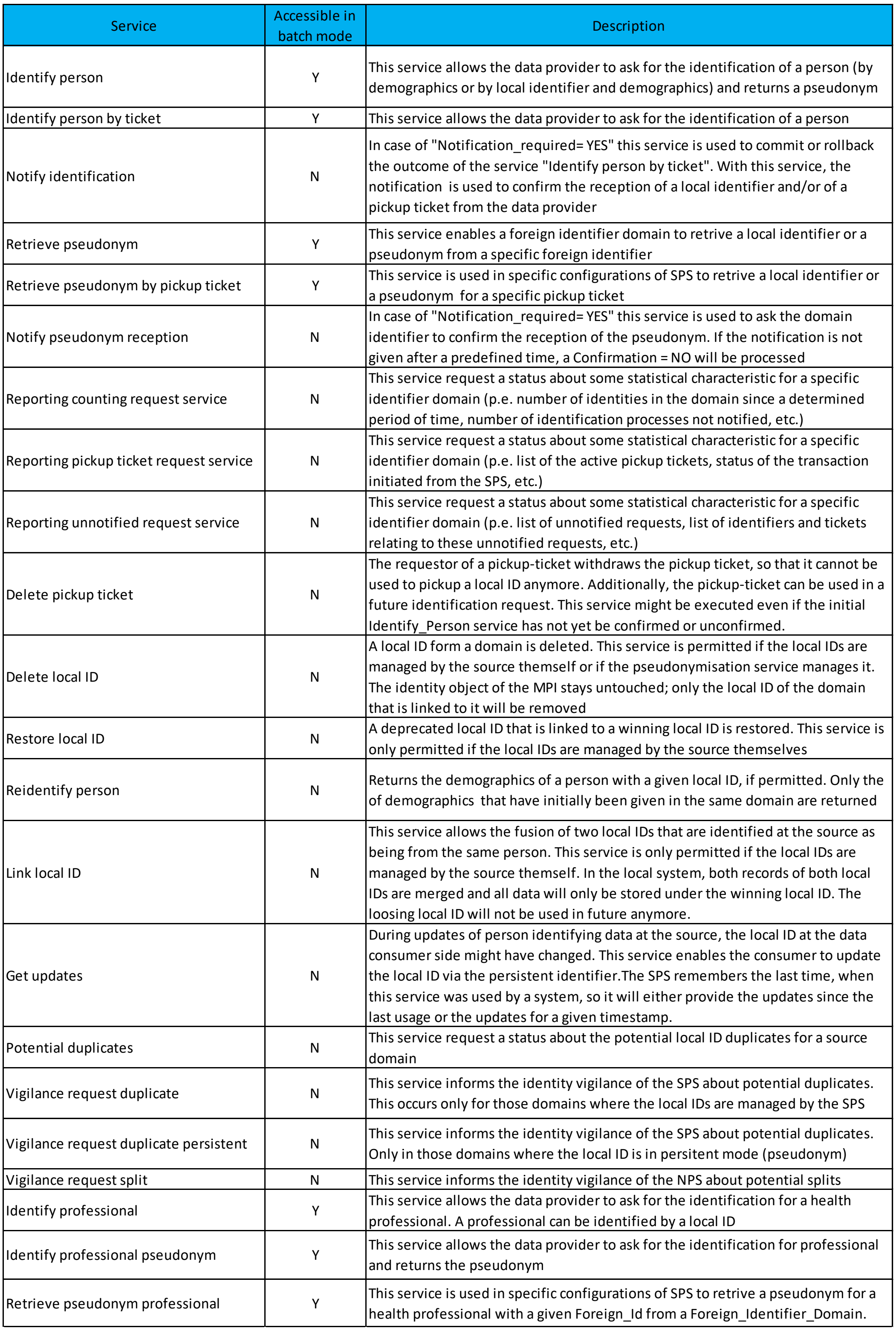
Access to pseudonymisation features

All functions of the pseudonymisation server are accessible via a web service (WS) connector by SOAP (Simple Object Access Protocol) message and/or by a batch mode.

Two authentication methods are available: authentication via Token X509 or authentication via SAML assertion.

Two modes of exchange are offered to access the pseudonymisation services: Web Services (WS) are generally used for basic calls (or low volume identity calls), alternatively, batch mode by file sharing allows for mass pseudonymisation. The exchanges via WS are performed in SOAP 1.2. Differently, in the case of batch mode, the client application deposits a request file in an *ad hoc* directory. The possible exchange protocols are SSH protocol (also referred to as Secure SHell) and secure file transfer protocol (SFTP).

Table 1 indicates the list of WS available for the SPS application. WS are described in web services description language (WSDL).



Service	Accessible in batch mode	Description
Identify person	Y	This service allows the data provider to ask for the identification of a person (by demographics or by local identifier and demographics) and returns a pseudonym
Identify person by ticket	Y	This service allows the data provider to ask for the identification of a person
Notify identification	N	In case of "Notification_required= YES" this service is used to commit or rollback the outcome of the service "Identify person by ticket". With this service, the notification is used to confirm the reception of a local identifier and/or of a pickup ticket from the data provider
Retrieve pseudonym	Y	This service enables a foreign identifier domain to retrive a local identifier or a pseudonym from a specific foreign identifier
Retrieve pseudonym by pickup ticket	Y	This service is used in specific configurations of SPS to retrive a local identifier or a pseudonym for a specific pickup ticket
Notify pseudonym reception	N	In case of "Notification_required= YES" this service is used to ask the domain identifier to confirm the reception of the pseudonym. If the notification is not given after a predefined time, a Confirmation = NO will be processed
Reporting counting request service	N	This service request a status about some statistical characteristic for a specific identifier domain (p.e. number of identities in the domain since a determined period of time, number of identification processes not notified, etc.)
Reporting pickup ticket request service	N	This service request a status about some statistical characteristic for a specific identifier domain (p.e. list of the active pickup tickets, status of the transaction initiated from the SPS, etc.)
Reporting unnotified request service	N	This service request a status about some statistical characteristic for a specific identifier domain (p.e. list of unnotified requests, list of identifiers and tickets relating to these unnotified requests, etc.)
Delete pickup ticket	N	The requestor of a pickup-ticket withdraws the pickup ticket, so that it cannot be used to pickup a local ID anymore. Additionally, the pickup-ticket can be used in a future identification request. This service might be executed even if the initial Identify_Person service has not yet be confirmed or unconfirmed.
Delete local ID	N	A local ID form a domain is deleted. This service is permitted if the local IDs are managed by the source themself or if the pseudonymisation service manages it. The identity object of the MPI stays untouched; only the local ID of the domain that is linked to it will be removed
Restore local ID	N	A deprecated local ID that is linked to a winning local ID is restored. This service is only permitted if the local IDs are managed by the source themselves
Reidentify person	N	Returns the demographics of a person with a given local ID, if permitted. Only the of demographics that have initially been given in the same domain are returned
Link local ID	N	This service allows the fusion of two local IDs that are identified at the source as being from the same person. This service is only permitted if the local IDs are managed by the source themself. In the local system, both records of both local IDs are merged and all data will only be stored under the winning local ID. The loosing local ID will not be used in future anymore.
Get updates	N	During updates of person identifying data at the source, the local ID at the data consumer side might have changed. This service enables the consumer to update the local ID via the persistent identifier.The SPS remembers the last time, when this service was used by a system, so it will either provide the updates since the last usage or the updates for a given timestamp.
Potential duplicates	N	This service request a status about the potential local ID duplicates for a source domain
Vigilance request duplicate	N	This service informs the identity vigilance of the SPS about potential duplicates. This occurs only for those domains where the local IDs are managed by the SPS
Vigilance request duplicate persistent	N	This service informs the identity vigilance of the SPS about potential duplicates. Only in those domains where the local ID is in persitent mode (pseudonym)
Vigilance request split	N	This service informs the identity vigilance of the NPS about potential splits
Identify professional	Y	This service allows the data provider to ask for the identification for a health professional. A professional can be identified by a local ID
Identify professional pseudonym	Y	This service allows the data provider to ask for the identification for professional and returns the pseudonym
Retrieve pseudonym professional	Y	This service is used in specific configurations of SPS to retrive a pseudonym for a health professional with a given Foreign_Id from a Foreign_Identifier_Domain.

Table 1: List of available WS. The table lists all the functions of the pseudonymisation server that are accessible via WS. When used in batch mode, the SPS offers a limited list of WS. Column 2, accessible in batch mode, depicts the WS available (Y) and not available (N) when using the batch mode.

Results

Use cases supported by the SPS

As previously mentioned, one main point of the SPS solution is the identity matching. Demographics (name, surname, sex and birth date) are used for the matching of identities new to the SPS. For identity matching of identities that are already known by the SPS, demographics and/or local patient identifier can be used accordingly to the settings of the specific use case. The local person identifier identifies an identity in a specific identifier domain. Thus, in the SPS, pseudonyms are generated based on a set of demographics and/or a local patient identifier. By producing the pseudonyms and relating them to the MPI-, the SPS represents the only functional element where a pseudonym can be linked to the identity of a patient. Thus, the SPS allows the implementation of fluxes where different institutions, having different identifier domains and separate local person identifiers for the same identity, can retrieve and exchange the medical data of this identity through the use of a unique pseudonym.

The SPS can generate pseudonyms according to four of the different use cases described in the Roth manuscript [2].

Use case A: Health data are shared between two sites belonging to the same source identification domain using a unique pseudonym. As the source identification domain is the same, the local personal identifier of the patient as well as his pseudonym is known to all the different sites belonging to this identification domain **(Image 3)**.

Use case B: The only site that interacts with the SPS service is the site that harbors the true identity of the patient (Site A, "From"). The pseudonym is shared between two different identification domains belonging to different sites (From and To). The identification domain "From-Domain" indicates the domain where the local person identifier "From-Domain" is identified. The identification domain "To-Domain" indicates the domain where the pseudonym is generated. In this specific case, to retrieve a pseudonym, the identification domain "From-Domain" calls the SPS using the local person identifier specific to the site "From-Domain". The pseudonym will be used to share health data between a source domain (identification domain "From-Domain") and the destination domain (identification domain "To-Domain"). The pseudonym is unique for both identification domains (From and To) and both sites are aware about the pseudonym **(Image 4)**.

Use case C: The site interacting with the SPS service is the one that wishes to use pseudonymised data (Site B, "To"). The two sites with different identification domains (From and To) are sharing health data through the use of the local person identifier from the source site "From-Domain". The destination site "To-Domain" then uses this local person identifier "From-Domain" to retrieve a pseudonym. In this way, the destination site "To-Domain" does not have any information about the person's identity and is the only domain that knows about the pseudonym **(Image 5)**.

Use case D: Two different identification domains belonging to different sites (From and To) are sharing health information through the use of a ticket. In this case, the identification domain "From-Domain" calls the SPS using its own local person identifier to retrieve the concerned ticket. Following this, the ticket is coupled to health data that are shared with the identification domain "To-Domain". The pseudonym is exclusively retrieved from the identification domain "To-Domain" by calling the SPS through the use of the ticket. In this use case, the destination site "To-Domain" does not have any information about the local person identifier nor about his identity and is the only domain that is aware about the pseudonym **(Image 6)**.

The challenge of implementing a pseudonymisation use case is due to the fact that a unique workflow cannot be used as a golden standard but has to be customized according to the users' needs and the sequence/nature of information that need to be shared between different parties. Customization occurs by combining different use cases as well as WS

(Table1) that are implemented accordingly to the specificity of the use cases. In addition to the described use cases (A, B, C and D), the SPS allows for the de-identification of a batch of identities. This option is often required when pseudonymisation is applied to a study that does not start from scratch but that has already a history of identity to de-identify. The batch mode will supply all the parameters and options of the services through file exchange. This is done through file sharing between a user application and the SPS. Each client is associated to a secured repository directory via SFTP. Pseudonymisation through batch mode will have access to a limited list of WS **(Table 1)**.

The system supports re-identification, this is not performed by users of the service SPS but by the administrators of the SPS, for instance following a request where a manual mistake was made during the entry process. A process for re-identification following normal or incidental findings has still to be defined. Moreover, to date, a national law dealing with re-identification following pseudonymisation is not available.

Image 3: Use case A. Health data are shared from source sites to destination sites that are part of the same unique identification domain. The local person identifier of the patient as well as his pseudonym is known to all the different sites belonging to this identification domain.

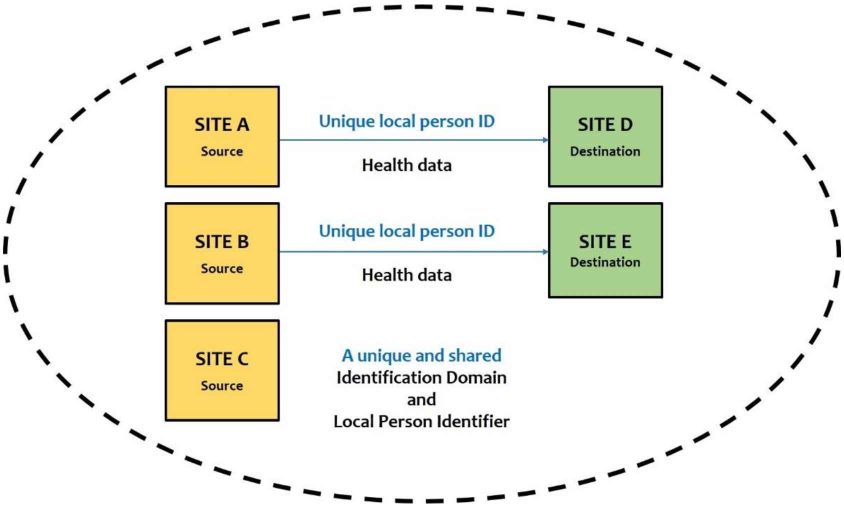


Image 4: Use case B. Health data are shared between two sites belonging to different identification domains (From and To). The identification domain "From-Domain" represents the source domain where the patient is identified through a local person identifier. The identification domain "From-Domain" represents the domain where the pseudonym is generated and used to share the health data with the "To-Domain".

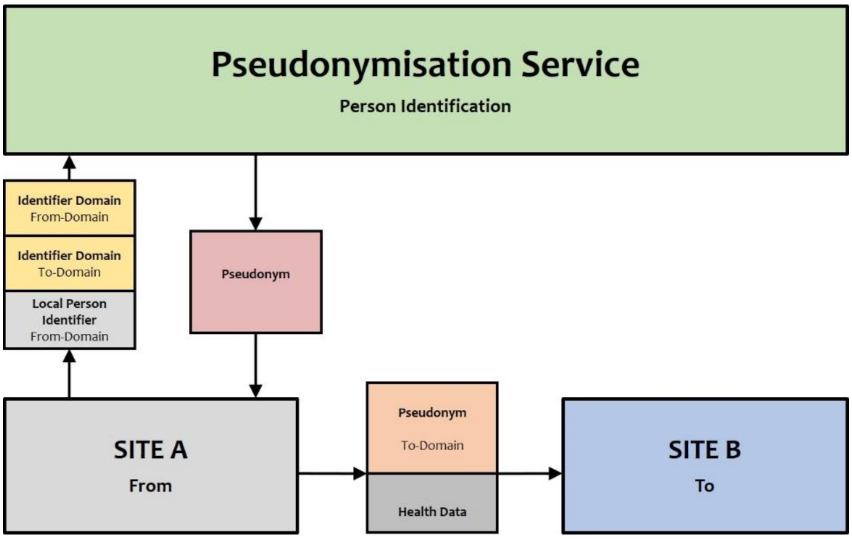


Image 5: Use case C. Two sites belonging to different identification domains (From and To) are exchanging health data. Health data are exchanged through the use of the local person identifier from the source site "From-Domain". Through this local person identifier, the identification domain "To-Domain" retrieves the pseudonymisation. The identification domain "From-Domain" represents the domain where the pseudonym is generated.

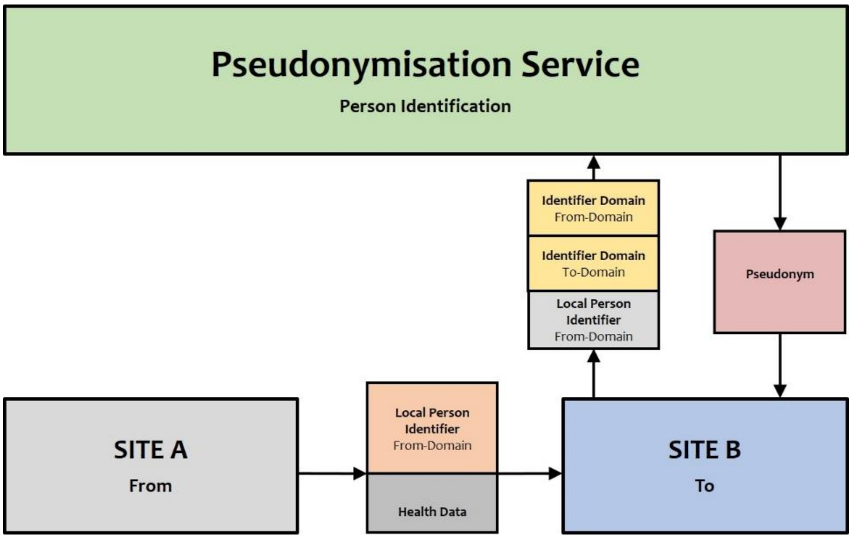
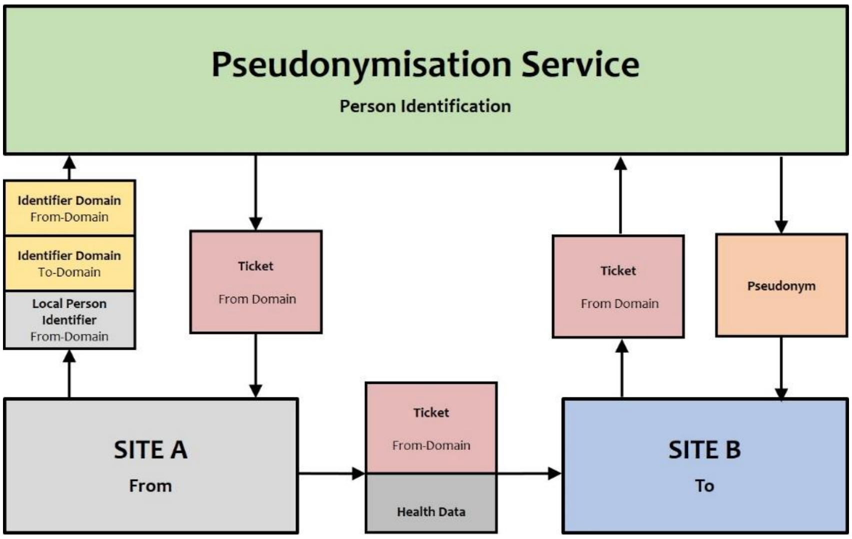


Image 6: Use case D. Health data are shared using a ticket between two different identification domains belonging to different sites (From and To). Patient identification is performed in the identification domain "From-Domain" that retrieves the ticket. This ticket is used by the "To-Domain" to call the SPS and retrieve the pseudonym.



Discussion

The healthcare and modern biomedical research sectors are gathering personal data at an ever-increasing pace, causing concerns over privacy and driving the need for data minimization techniques. [9]. On the regulatory level for instance, the EU has first elaborated directives regarding the processing of personal data and on the free movement of such data [10]; then, the GDPR was adopted in 2018 [1]. GDPR mirrors, through the article 25 "Data protection by design and by default" [11] and the article 32 "Security of processing" [12], the European will to preserve confidentiality while handling health data of patients. In addition to that, through article 4(5), GDPR has provided a clear definition of pseudonymisation as an act of processing personal data. With GDPR, the EU strongly encourages the use of pseudonymisation as minimization technique for personal data [13]. Furthermore, updated recommendations for best-practices on pseudonymisation are released yearly from the European Union Agency for Cybersecurity (ENISA) [14-16]. Similarly to the case in the EU, the California Consumer Privacy Act (CCPA) of 2018, establishes one of the most comprehensive data privacy regulations in the USA [17].

As previously mentioned throughout this paper, we believe that a prerequisite for an efficient pseudonymisation service is based on a trustworthy identity management infrastructure. In line with our approach, literature presents other examples of pseudonymisation models that depend on an IT infrastructure for federated identity management [18, 19].

In a previous publication, we explained the implementation of a national master patient index (MPI) based on a hierarchical federated identity management model that enables cross-system patient identification using a unique identifier [7]. In the same paper, we discussed about the importance of this unique identifier as a key precondition not only for a univocal match between the patient and the correct medical records, but also for data minimization purposes. The relevance of a unique global patient identifier, allowing for the integration of data coming from different information systems (ISs), is a concept that is highly corroborated in literature [20-22]. In this paper, we describe the implementation of a pseudonymisation service for identity traits, the SPS, dedicated to the health and research sector. Our model is composed of two functional components: the first one is the MPI that offers identity management. The second component is the SPS, the pseudonymisation service that is employed for patient de-identification and for the eventual re-identification.

Importantly, the increasing amount of digital data collected during patient care, could help clinical and academic research institutions [23-25]. Nevertheless, the manipulation of large quantities of health information (national registries, biobanks and research studies) creates challenges for these organizations in regards to protecting the privacy of patients and research subjects. For this reason, consequently to the GDPR adoption, de-identification via basic anonymisation techniques has been the strategy of choice for health care providers and research institutions [26, 27]. This is linked to the accessibility of this method in its basic version, where identification traits are erased definitively, as well as to the misbelief that once anonymised, data can never be re-identified.

However, due to advances in data science as well as fast evolving methods for data storage, new requirements have arisen leading to an increased demand for the use of pseudonymisation to de-identified large data sets for any application that interconnects medical, research and public health needs. While implementing a pseudonymisation service can pose significant challenges for an organization, its ability to provide de-identified data for secondary uses makes it a highly desirable option for patient-targeted innovation or research purposes. By using pseudonymisation, organizations can ensure the privacy and confidentiality of patient data while still being able to develop and implement innovative solutions that improve patient outcomes and experiences. Indeed, the de-identification techniques are required to be able to handle data from patients in an environment characterized by an ever increasing complexity in terms of the number of interfaces and subsystems, which need to be implemented to guarantee system interoperability [28, 29]. In our role of independent trusted third party (TTP), we offer these two functional components as part of a unique pseudonymisation service integrated to the eHealth national platform and that can address the specific requirement in transitional research as well as in the health care sector. By performing pseudonymisation with our service, health and research institutions mutualize our identity management infrastructure and do not have to implement an additional service or to hire specialized personal to handle these tasks.

It is undeniable that the increase of digital clinical data relates significantly with a higher risk of the re-identification of individuals [17]. For instance, Rocher et al. present in their paper different cases of supposedly anonymous datasets that have recently been released and re-identified [30]. An increased risk of re-identification exists as well when combining data sets from different sources. For example, Sweeney in one of his study shows that the combination of a medical database combined with a voters lists was enough for re-identification [31]. Furthermore, Rocher et al., proposes a method to correctly re-identify heavily incomplete datasets [30]. These studies underline the need for the future challenge of the scientific community to satisfy the data minimization standards imposed by GDPR.

We argue that a peculiar attention must be applied by data protection authorities and further efforts must be implemented in evaluating and reporting risks as well as elaborating security countermeasures and increase the implementation of privacy-enhancing systems to preserve people's privacy. The standard ISO/TS 25237, for instance, proposes technical specifications detailing the principles and requirements for privacy protection using pseudonymisation services in regards to the protection of personal health information [32].

Conclusions

With the here described method, we can effectively propose a service that enables the pseudonymisation of identities for the health sector as well as for the research sector. The services we provide enable the management of the unique link present between demographic traits of a patient and his different pseudonyms that can result from different studies and/or applications. Being integrated in the architecture of the national eHealth platform, this system offers a high level of system interoperability with ISs that belong to the national health ecosystem of Luxembourg. We argue that, identity and data protection issues are upcoming topics for the translational research sector, as such a specifically dedicated identity management IT infrastructure is still lacking in this context. Of note, when employing this service, users from the research sector can benefit from the infrastructure of the national eHealth platform. As such, they dispose of multiple test environments as well as of the identity vigilance requirements that are necessary in such an architecture. Nevertheless, we believe that future work should be aimed at guarantying the mutualization of existing specialized infrastructures in order to continue to insure the same level of system interoperability for the health sector and the research sector at the national level as well as at the international level.

Overall, we consider that this approach represents a valuable tool that could be applied to facilitate data aggregation of various data sources for the health care sector, translational research applications as well as for data warehousing and data lake approaches.

Declarations

Consent for Publication Availability of data and material
No dataset has been used and/or analyzed during the current study. Documentation relating to the implementation of the pseudonymisation services available from the corresponding author upon reasonable request.

Authors' contributions
Wrote the paper: RV, FW, PD
Contributed to the paper with review/technical expertise/images: FW, CP, FM, HZ, HB
Conceived the paper: RV, HB, FW
All authors have read and approved the manuscript.

Acknowledgements
We gratefully acknowledge the members of the *Agence eSanté* as well as the external collaborators who contributed directly or indirectly to the here described work. A special thanks goes to Jérôme Gauthier for his review of the manuscript and to Luis Batista for his expertise in images editing.

References

1. *EU GDPR, REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL*. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj.
2. Roth, U., *A generalized view on pseudonyms and domain specific local identifiers - Lessons learned from various use cases.* International Journal on Advances in Security, 2014. 7: p. pp. 76-92.
3. Roth, U., *Protecting the Privacy with Human-Readable Pseudonyms : One-Way Pseudonym Calculation on Base of Primitive Roots. Sixt International Conference on eHealth, Telemedicine, and Social Medicine - eTelemed 2014, Barcelona, Spain, 24-27.03.2014, pp 111-115, IARIA 2014.*
4. Roth, U., *A Generalized View on Pseudonyms and Domain Specific Local Identifiers - Lessons Learned from Various Use Cases.* International Journal on Advances in Security, December 2014. vol. 7: p. pp. 76-92.
5. PARTY, A.D.P.W. *Lignes directrices ex G29 p.9* 2014; Available from: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.
6. Adams, C., *Trusted Third Party*, in *Encyclopedia of Cryptography and Security*, H.C.A. van Tilborg and S. Jajodia, Editors. 2011, Springer US: Boston, MA. p. 1335-1335.
7. Vaccaroli, R., et al., *Grand Duchy of Luxembourg: a case study of a national master patient index in production since five years.* BMC Med Inform Decis Mak, 2020. 20(1): p. 163.
8. *IHE Patient Identifier Cross Referencing (IHE-PIX)*. Available from: https://wiki.ihe.net/index.php/Patient_Identifier_Cross-Referencing#PIX_.28html.29_specification.
9. Kohlmayer, F., R. Lautenschlager, and F. Prasser, *Pseudonymization for research data collection: is the juice worth the squeeze?* BMC Med Inform Decis Mak, 2019. 19(1): p. 178.
10. *European Parliament and Council of the European Union. Regulation (EU) 2016/679 European Parliament and council directive 95/46/EC of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. .* 2016.
11. *Art. 25 EU GDPR*. Available from: https://www.privacy-regulation.eu/en/article-25-data-protection-by-design-and-by-default-GDPR.htm.
12. *Art. 32 EU GDPR*. Available from: https://www.privacy-regulation.eu/en/article-32-security-of-processing-GDPR.htm.
13. Mourby, M., et al., *Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK.* Computer Law & Security Review, 2018. 34(2): p. 222-233.
14. ENISA, *Pseudonymisation techniques and best practices : https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices/@@download/fullReport*. December 03, 2019.
15. ENISA, *Recommendations on shaping technology according to GDPR provisions - An overview on data pseudonymisation: https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions/@@download/fullReport*. January 28, 2019.
16. ENISA, *Data Pseudonymisation: Advanced Techniques and Use Cases: https://www.enisa.europa.eu/publications/data-pseudonymisation-advanced-techniques-and-use-cases/@@download/fullReport*. January 28, 2021
17. Ulf Mattsson, M., *Practical Data Security and Privacy for GDPR and CCPA: https://www.isaca.org/resources/isaca-journal/issues/2020/volume-3/practical-data-security-and-privacy-for-gdpr-and-ccpa*. ISACA JOURNAL 2020.
18. Fischer, H., R. Röhrig, and V.S. Thiemann, *A Generic IT Infrastructure for Identity Management and Pseudonymization in Small Research Projects with Heterogeneous and Distributed Data Sources Under Consideration of the GDPR.* Stud Health Technol Inform, 2019. 264: p. 1837-1838.

19. Nitzlnader, M. and G. Schreier, *Patient identity management for secondary use of biomedical research data in a distributed computing environment.* Stud Health Technol Inform, 2014. 198: p. 211-8.
20. Agrawal, R. and S. Prabakaran, *Big data in digital healthcare: lessons learnt and recommendations for general practice.* Heredity, 2020. 124(4): p. 525-534.
21. Gliklich RE, D.N., Leavy MB, editors. . *Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition. Rockville (MD): Agency for Healthcare Research and Quality (US); 2014 Apr. 17, Managing Patient Identity Across Data Sources. Available from:* https://www.ncbi.nlm.nih.gov/books/NBK208618/.
22. Riplinger, L., J. Piera-Jiménez, and J.P. Dooling, *Patient Identification Techniques - Approaches, Implications, and Findings.* Yearbook of medical informatics, 2020. 29(1): p. 81-86.
23. de Lusignan, S., *Effective pseudonymisation and explicit statements of public interest to ensure the benefits of sharing health data for research, quality improvement and health service management outweigh the risks.* Inform Prim Care, 2014. 21(2): p. 61-3.
24. De Meyer, F., G. De Moor, and L. Reed-Fourquet, *Privacy Protection through pseudonymisation in eHealth.* Stud Health Technol Inform, 2008. 141: p. 111-8.
25. Heurix, J., et al., *Recognition and pseudonymisation of medical records for secondary use.* Med Biol Eng Comput, 2016. 54(2-3): p. 371-83.
26. Kushida, C., et al., *Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies.* Medical care, 2012. 50 Suppl: p. S82-101.
27. Chevrier, R., et al., *Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review.* Journal of medical Internet research, 2019. 21(5): p. e13484-e13484.
28. Rinty, M.R., U.K. Prodhan, and M.M. Rahman, *A prospective interoperable distributed e-Health system with loose coupling in improving healthcare services for developing countries.* Array, 2022. 13: p. 100114.
29. Shull, J.G., *Digital Health and the State of Interoperable Electronic Health Records.* JMIR Med Inform, 2019. 7(4): p. e12712.
30. Rocher, L., J.M. Hendrickx, and Y.-A. de Montjoye, *Estimating the success of re-identifications in incomplete datasets using generative models.* Nature Communications, 2019. 10(1): p. 3069.
31. Sweeney, L., *<i>k</i>-anonymity: a model for protecting privacy.* Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 2002. 10(5): p. 557–570.
32. *ISO/TS 25237:2008 Health informatics — Pseudonymization:* https://www.iso.org/standard/42807.html 2008.

Agence eSanté G.I.E / 2-4, rue Eugène Ruppert, L-2453 Luxembourg / B.P. 2511, L-1025 Luxembourg

www.esante.lu