Mathematical Reconciliation of Medical Privacy Policies

DINESHA RANATHUNGA, MATTHEW ROUGHAN, and HUNG NGUYEN, The University of Adelaide

Healthcare data are arguably the most private of personal data. This very private information in the wrong hands can lead to identity theft, prescription fraud, insurance fraud and an array of other crimes. Electronic-health systems such as My Health Record in Australia holds great promise in sharing medical data and improving healthcare quality. But, a key privacy issue in these systems is the misuse of healthcare data by 'authorities'. The recent General Data Protection Regulation (GDPR) introduced in the EU aims to reduce personal-data misuse. But, there are no tools currently available to accurately reconcile a domestic E-health policy against the GDPR to identify discrepancies. Reconciling privacy policies is also non-trivial because policies are often written in free text, making them subject to human interpretation.

In this paper, we propose a tool which allows the description of E-health privacy policies, represent them using formal constructs making the policies precise and explicit. Using this formal framework, our tool can automatically reconcile a domestic E-health policy against the GDPR to identify violations and omissions. We use of our prototype to illustrate several critical flaws in Australia's My Health Record policy, including a non-compliance with GDPR that allows healthcare providers to access medical records by default.

ACM Reference Format:
Dinesha Ranathunga, Matthew Roughan, and Hung Nguyen. 2020. Mathematical Reconciliation of Medical Privacy Policies. 1, 1 (March 2020), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Healthcare data are arguably the most sensitive of personal data belonging to a subject. Sensitive information such as a person's family history, demographic data, medical conditions, and current medications are often embedded in healthcare data. Such very private information in the wrong hands can lead to identity theft, prescription fraud, financial fraud, insurance fraud and a wide array of other crimes [15]. Electronic-health (E-health) systems such as 'My Health Record' in Australia hold the promise that in sharing medical data, we can improve healthcare quality and reduce costs. E-health removes the need for doctors to work in silos without access to the full range of clinical, prescription and health information about a patient. It enables each healthcare provider to observe patient information online to make healthcare management easier and safer. It can avoid common problems such as drug interactions and be life-saving in emergencies [3]. But, the potential benefits also bring the serious issue of data privacy. The disclosure of medical records is the most frequent of all reported privacy breaches [1].

A key privacy issue in the context of healthcare data is their misuse by 'authorities' [15]. For instance, a health insurance provider accessing the medical records to verify the legitimacy of a claim is acceptable. But, using the data to genetically discriminate against a subject when they obtain health cover, is not. Likewise, a government's ability to access the details of a person's

Authors' address: Dinesha Ranathunga, Dinesha.Ranathunga@adelaide.edu.au; Matthew Roughan, Matthew.Roughan@adelaide.edu.au; Hung Nguyen, Hung.Nguyen@adelaide.edu.au, The University of Adelaide.

XXXX-XXXX/2020/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

sexual behavior or abortion history might allow use of that knowledge for purposes not originally intended; *e.g.,* for a politician to gain advantage over their opponent by disclosing the information during an election campaign. Similarly, access of a child's healthcare data by a sexually abusive father or access to patient data by doctors who malpractice could cause serious harm to the patient.

Data misuse can be reduced through granular control of the aspects of a person's healthcare data that can be accessed by who and for what purpose. The recent General Data Protection Regulation (GDPR) [8] introduced in the EU provides a framework for granular control of personal data. The GDPR offers a good baseline privacy policy for personal data and the robust enforcement of this baseline applies to any organization that collects or processes EU residents' personal data. Reconciling E-health privacy policies against the GDPR for compatibility hence becomes a priority for any such organization [13, 21]. But, reconciling privacy policies against the GDPR is non-trivial because the regulation is written in free text making it largely qualitative and subject to human interpretation [23]. Natural-language based descriptions can also potentially be imprecise, incomplete and/or inconsistent. Reconciling against such a description could lead to well-intentioned mistakes at best and legal battles at worst. A good example is how the placement of commas in the US constitution describing its citizens' right to bear arms caused the Supreme Court to abolish a ban on handguns [19].

The problem of imprecise privacy-policy description can be overcome by representing policies using formal constructs [16, 23]. Mathematical constructs allow policy properties to be captured precisely and explicitly without ambiguity. More importantly, formal underpinnings allow one to accurately reason about policies. So, one could for instance, rigorously reconcile, a domestic E-health privacy policy such as Australia's My Health Record policy against the GDPR. However, policy makers are not mathematicians. They need tools which abstract away complex formal semantics and allow them to intuitively specify and reconcile their policies. The then Australian prime minister, Malcolm Turnbull, stated "The laws of mathematics are very commendable, but the only law that applies in Australia is the law of Australia" [14]. Mathematical laws, in reality, are not negotiable, however there is no reason that the laws of the land and mathematics need be in conflict, if the former are designed well. Healthcare policy is a case in point. The challenge is to find the right abstract model that could be used to describe policies at a human understandable level and could also be used for mathematical reasoning.

In this work we bridge the gap between imprecise privacy-policy descriptions and formal models by developing a rigorous mathematical model based on metagraphs that can be used for expressing human-understandable description of policies and for rigorously analyzing them. We build prototype software of the model - all code and data used in this paper are publicly available[1]. Our software tool allows users to (1) naturally describe E-health privacy policies at a high level; (2) model and analyze policy properties mathematically; and (3) formally reconcile an E-health policy against another policy such as the GDPR.

We take Australia's My Health Record [3] as a prime example to highlight the potential applications of our tool in this domain. Upon reconciling My Health Record policy against the GDPR, we find that the former violates the latter on several counts. For instance, healthcare providers can access My Health Record data by default, without the record owner's explicit consent. Also, the controller's obligation in the GDPR to notify data recipients of subject data rectification, erasure or processing-restriction is not upheld in My Health Record. No doubt clever people could ascertain these facts, but providing water tight mathematical arguments can aid convincing policy makers (other than the aforesaid ex prime minister) of the validity of the arguments and when a policy is

[1]at https://github.com/dinesharanathunga/MyHealthRecord

right, it is valuable to be able to show this. The problem is also much wider than just that of health records as we shall see in the following section.

The rest of the paper is organized as follows. In Section 2, we describe the related work and background materials. In Section 3, we discuss the granularity dimensions of privacy policies. We explain in Section 4 how our model could be used to encode My Health Record and GDPR policies. The two policies are reconciled in Section 5. Finally, we conclude the paper with discussions of its limitations in Section 6.

2 BACKGROUND AND RELATED WORK

We review related work and background knowledge in this section. The list of abbreviations is provided at the end of the paper.

2.1 Policy Models, Languages and Tools

Organizations often have a privacy policy which describes how they collect, store and use personal information. These policies are intended to help consumers make informed decisions when interacting and sharing their personal information with the organization. For instance, the popular social media site Facebook's privacy policy describes that a user's profile can only be disclosed to third parties for research purposes if the user has explicitly consented [9]; and the online shopping site eBay's privacy policy states that a user's email addresses can only be disclosed to members involved in a completed eBay transaction [7].

However, there is no standardized vocabulary to describe privacy policies. Moreover, these are often long and involve statements that are at best difficult for an average user to comprehend and at worst impossible to avoid because they are shrink-wrapped with a product or service that is already purchased. Privacy policies are meant to inform users of how an organization collects and shares their user data. But the lack of policy comprehensibility prevents such transparency.

The EU's GDPR is a recent example of how privacy policies are still being articulated in user-unfriendly language. The regulation outlines for instance, conditions for accessing a user's personal data by recipients. For most access purposes (*e.g.,* to research on personal data), explicit user consent is required. However, there are exceptions where user consent is not sought (*e.g.,* when accessing the data is in the public's interest). Such exceptions must be clearly conveyed to stakeholders.

Many policy languages [2, 6, 10, 22] have been proposed to assist with the specification, analysis and enforcing of privacy policies. Most of these languages were designed for specific purposes. For instance, IBM designed the Enterprise Privacy Authorization Language (EPAL) to formalize internal enterprise privacy policies [2]. But, EPAL lacks automated support to reconcile two policies – a key goal of our work.

The Platform for Privacy Preferences (P3P) language was introduced [22] by the World Wide Web Consortium (W3C) to enable expression of website privacy policies in a machine-interpretable format. Likewise, the Role Based Access Control (RBAC) language XACML was developed for expressing both privacy and security policies so that they can be interpreted by computers to perform the actions stated in the policies [10]. Both P3P and XACML are light-weight XML markup languages and are not formally rigorous; *i.e.,* we cannot precisely reason about policies using them.

Solid [5] is a decentralized platform recently proposed for managing users' data in social Web applications. The approach gives a user control over the privacy of their data. Using Solid a user can manage who can access what elements of their data, independent of the applications that create and consume this data. This flexibility enables a user to easily switch between similar applications enabling the reuse of their data.

There have been significant development in the last three years in modelling of privacy policies related to GDPR. Closest to our work is the work of Wang et al. [23] where the authors use knowledge

graphs to compare Chinese and European Internet companies' privacy policy. A privacy knowledge graph contains as nodes the set of all companies and the privacy information that being collected and as edge the links between each company and the privacy data collected by that company. The authors first construct one knowledge graph for all companies in Europe and one knowledge graph for all companies in China. The two graphs are then compared. A different model using ontology design pattern for privacy policies was presented in [16]. PrivacyGuide [20] is a privacy policy summarization tool that uses machine learning and natural language processing techniques to summarize privacy policies using General Data Protection Regulation (GDPR) framework as a guideline for interpretation. PrivacyBot [21] uses the same techniques to detect privacy sensitive information that users post on-line, again using GDPR as a guideline for interpretation of privacy data. An longitudinal assessment of the impact of GDPR on privacy policies online was provided in [13].

Each of the above languages and tools provides formalized policy descriptions with the intention of automating policy implementation but they lack tools to mathematically reason about policies. We develop in this work a generic mathematical tool that rigorously models a broad ranges of privacy policies. These models can then be cross-compared and reconciled using precise mathematical procedures. We present here one example application of our approach in reconciling domestic health policies with GDPR. Our formal policy reconciliation approach can potentially be extended in multiple dimensions. For example, we could use our model to check if privacy preferences in Solid conform with a website's privacy policy. We could also apply the machine learning algorithms in [20, 21] on our metagraph model for privacy compliance of Internet companies and users.

2.2 Mathematical Modeling of Policy with Metagraphs

We provide here, a very brief background on the formal modeling and reconciliation framework used in this paper. We use the concept of a *metagraph*, which is a generalized graph-theoretic structure that offers rigorous formal foundations for modeling and analyzing communication-network policies in general [11, 18] and business management policies [4].

A metagraph is a directed graph between a collection of sets of "atomic" elements [4]. Each set is a metagraph node and each directed edge describes a relationship between two sets. Fig. 1 shows an example where a set of users (U_1) are related to sets of network resources (R_1, R_2, R_3) by the edges e_1, e_2 and e_3 allowing a user u_i to access a resource r_j.

A metagraph is more useful than a graph because graphs associate individual elements not sets of elements (which can have some overlap). As per the access-control example in Figure 1, real-world policies often associate sets of elements and metagraphs allow capture and visualization of such policies naturally and parsimoniously. Graph-based representations require the user to explicitly track node overlaps when analyzing properties such as reachability. Metagraphs reduce this complexity by handling node overlaps automatically. Note that the metagraphs we use in this paper is more general than the meta-graphs used for analyzing heterogenous information networks (HIN) in [24]. In our metagraphs each vertex contains multiple elements whereas the the meta-graphs in [24] use single element for each vertex. This generalization is needed to model privacy policies where each rule can apply to multiple entities.

Metagraphs also support edge attributes. An example is a *conditional metagraph* which has propositions – statements that may be true or false – assigned to its edges as qualitative attributes [4]. A conditional metagraph is formally defined as

Definition 1 (Conditional Metagraph [4]). *A conditional metagraph is a metagraph* $S=\langle X_p \cup X_v, E\rangle$ *in which* X_p *is a set of propositions and* X_v *is a set of variables, and:*

1. at least one vertex is not null, i.e., $\forall e' \in E, V_{e'} \cup W_{e'} \neq \phi$

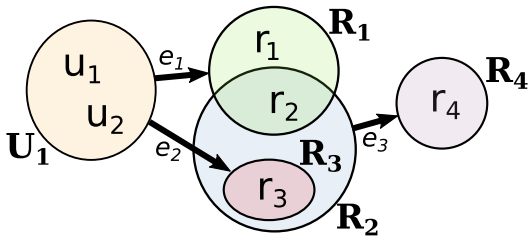

Fig. 1. *A metagraph consisting of six variables* $\{u_1, u_2, r_1, r_2, r_3, r_4\}$*, five sets* U_1, R_1, R_2, R_3, R_4 *and three edges* e_1, e_2 *and* e_3*. One could imagine* u_i *to be users,* r_i *to be resources and* e_i *to express valid relationships.*

2. the invertex (source vertex) and outvertex (destination vertex) of each edge must be disjoint, i.e., $X = X_v \cup X_p$ *with* $X_v \cap X_p = \phi$

3. an outvertex containing propositions cannot contain other elements, i.e., $\forall p \in X_p, \forall e' \in E$*, if* $p \in W_{e'}$*, then* $W_{e'} = p$.

These metagraphs are useful to model stateful policies. For instance, in My Health Record, third parties can access a subject's medical records for research purposes, only if the subject (or their representative) provides explicit consent. This conditional access can easily be modeled by propositions. Metagraphs also support several useful operators which allow one to analyse policy properties like consistency. One such operator is a *metapath* [4] which describes connectivity between sets of elements in a metagraph, but is somewhat different from a simple path in a graph. A metapath is represented by a set of edges and can inform of redundancies (edge or element wise) in a metagraph. Hence it is a very useful operator when reconciling metagraphs because only metapaths with no redundancies (*i.e.,* dominant metapaths [4]) need to be considered. An implementation of metagraphs with the associated mathematical operators described in this paper is provided in [17].

We want to emphasize here, the strong mathematical foundations of this work, but a key aim is also to provide access to these ideas and tools to non-mathematical users. Metagraphs have the advantage of providing both mathematical rigour but also an intuitive, visual description of policies.

3 PRIVACY POLICY DESCRIPTION

The goals of privacy policies vary, but they can be roughly categorized as Transparency, Intervenability and Unlinkability [12]. The protection goal of Transparency is defined as the ability to understand and reconstruct all personal data processing enabled at any given point of time. Intervenability is the ability to enforce changes and corrective measures to ongoing or planned personal data processing. Unlinkability ensures that personal data cannot be linked across multiple data-collection platforms [12].

An important consideration when achieving a privacy goal is the granularity at which the goal needs to be specified. Granularity refers to the finest level of discrimination we can make. For instance, in an image, resolution or granularity is the pixel size (we can't separate objects in an image that are smaller). The first concern when compressing an image would be how many pixels do we need, and so this is an initial step in considering how a privacy goal needs to be achieved. Finer granularity allows more control but creates extra complexity in the policy. For instance, understanding what data processing is enabled at a given point of time (*i.e.,* Transparency) requires us to consider aspects such as what data is processed, by who and for what purpose. We refer to these multiple aspects as *granularity dimensions*. Granularity dimensions requirements

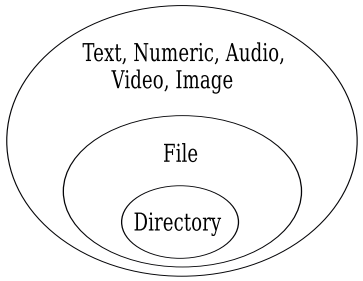

Fig. 2. *Data class granularity hierarchy: coarse granularities are subsets of the finer.*

are determined by the use case of the data. We describe below the use cases and the granularity dimensions required for modeling My Health Record and GDPR in metagraphs.

3.1 Data Use Case

In My Health Record, the various uses of its data are
Healthcare provision: Sharing of medical records allows healthcare providers to better diagnose patients and improve their treatments;
Representation: a legally-authorized or nominated representative may be required to manage a patient's medical records on their behalf;
Safeguard public interest: healthcare providers can also use a patient's medical records to prevent serious threats to public health or public safety;
Authority: Patient medical records can also be used in the exercise of official authority. For instance, the data may need to be disclosed to courts for coroner's investigations or for law enforcement purposes;
For research and evaluation: Patient medical records can be of value to medical and public health researchers.

Managing medical records on a subject's behalf (*i.e.,* representation) would have coarser Transparency requirements in comparison to providing healthcare to the subject (for instance, due to higher auditing requirements). We can systematically construct privacy policies per use case by identifying privacy-goal granularity requirements, discussed in the next section.

3.2 Granularity Dimensions

We identified the granularity dimensions applicable to each of the three privacy goals discussed for the GDPR in this section. For each dimension, we propose granularity levels applicable to the universal concept of E-health. These dimensions and levels enable systematic construction of privacy policies for each E-health use case. We provide a high-level explanation of the granularity dimensions here and details can be found in the Appendix.

Data class: describes the type of healthcare data available for processing. An example of the data class granuality is shown in Fig. 2.

Data category: Although a data class informs the type of data processed, it provides little contextual information. Thus, we define data category to additionally describe the context of the data processed. Each category can comprise of multiple data classes.

Purpose limitation: The GDPR also specifies the data-protection principle of purpose limitation [8], by restricting data processing to legitimate purposes only.

Processing purpose: We define processing purpose to describe legitimate objectives of processing medical records. The purposes applicable to E-health are linked to the use cases described earlier. A processing purpose differs from a use case in that the latter may require data processing for multiple purposes; *e.g.,* providing healthcare to a subject requires data processing for *view, subject interest, public Interest* and *official authority*.

Data recipient: As per the GDPR, a data recipient describes an entity who is allowed to process personal data.

Data-recipient category: Likewise, data-recipient category describes the type of entity allowed to process the data.

Consent provider: We define consent provider to describe the entity providing consent for purposes which require explicit consent. Where data processing is based on a subject's consent, privacy legislations specify a minimum age for a subject to give consent. For instance, the GDPR recognizes individuals who are 16 years or over as subjects being able to give consent [8]. This minimum age is increased to 18 years in the My Health Record policy [3]. Thus, an important consideration in determining the consent provider is the ***subject's age***.

Data breach nature: describes the type of data breach occurred. E-health privacy policies also often cover how data breaches within a system are notified to authorities, affected subjects and the public. This notification obligation is, enforced upon E-health system operators by privacy legislations (*e.g.,* in My Health Record the operator (ADHA) is bound by the data breach notification obligations in the My Health Records Act [3]). Thus, the ability to understand the data-breach notification mechanisms enabled in a system, is an important aspect of Transparency.

Data breach scope: describes the extent of a data breach.
Data breach consequence: describes the impact of the data breach in terms of what subject rights were violated.

3.3 Policy Model with Data Use Case and Granualtiy Dimensions

A summary of privacy-goal granularity requirements for three My Health Record use cases is provided in Table 1, based on the framework described above and the Appendix.

For instance, data processing is required at a *file* level granularity when providing healthcare to a subject, but is sufficient at a coarser *directory* level when representing a subject. Using these granularity requirements, we can systematically construct privacy-policy descriptions for each E-health use case. The policy description for representing a My Health Record owner can be constructed as shown in Figure 3.

We describe next, Australia's domestic E-health policy using the granularity framework above and formally compare it to the GDPR. We use My Health Record to demonstrate a concrete comparison, but the approach can be used with any domestic E-health policy.

4 POLICY METAGRAPHS

4.1 My Health Record Policy Metagraph

Figure 4 shows a use case of a conditional metagraph model of the privacy policy for representing a My Health Record owner. The nodes of this metagraph comprise of the data recipient (*i.e.,* the set of individuals comprising the subject and his/her representatives) and the My Health Record data set. We use the propositions *purpose*, *consent_provider* and *subject_age* to capture the access conditions. They describe that an individual is allowed to manage a subject's medical records on their behalf (a) if the subject is 18 years or over and the subject or their representative has consented for viewing

Table 1. *Summary of privacy-goal granularity requirements for three My Health Record use cases: (a) representing a My Health Record owner; (b) research and evaluation; and (c) providing healthcare to the owner. rep.=representative, cons.=consequences, conn.=connections, doc. = documents.*

Goal	Dimension	Required granularity for use case		
		Representation	Research	Provide care
Transparency	Data class	directory	directory	file
	Data category	all categories	documents	profile, doc.
	Data recipients	individuals	third parties	care providers
	Recipient category	level-1 conn.	indirect conn.	level-1, indirect
	Purpose	manage	research	subject interest
	Consent provider	subject, rep.	subject, rep.	subject, rep.
	Breach nature	disclosed	corrupted	corrupted
	Breach scope	single subject	subject group	subject group
	Breach cons.	freedom affected	rights affected	rights affected
Intervenability	Access data	send my data copy	send data copy	send data copy
	Rectify data	data category	data category	data category
	Erase data	data category	data category	data category
	Restrict processing	single purpose	single purpose	single purpose

```
individuals −−> my_health_record {
    use_case = representation;
    granularity.transparency={
      purpose={manage};
      consent_provider={subject,representative}
      subject_age={>=18}
      data_class={directory};
      data_category={profile,documents,history,
                     access_control,personal_notes};
      data_recipient={individuals};
      recipient_category={level_1_connections}}}

individuals −−> my_health_record {
    use_case = representation;
    granularity.transparency={
      purpose={manage};
      consent_provider={representative}
      subject_age={<18}
      data_class={directory};
      data_category={profile,documents,history,
                     access_control,personal_notes};
      data_recipient={individuals};
      recipient_category={level_1_connections}}}
```

Fig. 3. *Privacy policy description (partial) for representing a My Health Record Owner. The description is based on the granularity requirements in Table 1.*

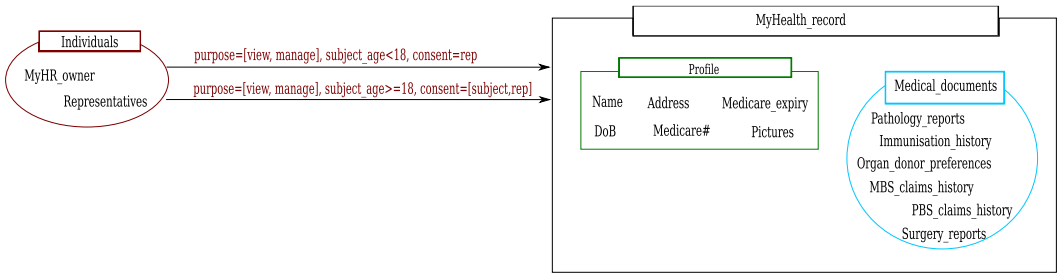

Fig. 4. *A conditional metagraph describing conditions for accessing medical records of individuals as defined in My Health Record policies. The conditional edges of metagraphs provide a powerful tool to encode Event-Condition-Action rules.*

or managing the records; or (b) if the subject is less than 18 years and the subject's representative has consented for viewing or managing the records.

Likewise, we modeled the privacy policies of all My Health Record use cases using conditional metagraphs. The resultant composite metagraph is shown in Figure 5.

As the figure shows, the complexity of the My Health Record policy metagraph is manageable; we can clearly describe the entire policy using a metagraph that fits a single A4 page. It only takes 141 LoC in our metagraph specification language (which we describe in Section V) to specify this policy metagraph model. This is in contrast to the privacy policy description in [3], which has more than 5000 words of hard-to-read text.

The edge propositions on this metagraph describe for instance, how access to a subject's medical records by healthcare providers is conditional on (a) the subject or their representative providing consent; or (b) access being in the interest of the subject; or (c) access being in the interest of public health and safety; or (d) access being the default system behavior.

The propositions on Figure 5 also show that third parties (such as portal operators and law enforcement agencies) are allowed read-only access to the medical records if the subject or their representative consents or if access is required for duties of official authority. Health researchers are also allowed access to the medical records if the subject or their representative provides consent.

The controller – the organization which determines the means and purpose of data processing – can also maintain healthcare data belonging to a subject (*e.g.,* by synchronizing Medicare data with the My Health Record) once the subject (or their representative) has given consent. In My Health Record, a subject can also intervene in the processing of their medical records by requesting the controller to (a) rectify incorrect or incomplete data; or (b) restrict data processing for a particular purpose; or (c) erase the data; or (d) withdraw the subject's consent. The controller is also obliged to promptly report details of any data breaches including the *nature*, *scope* and *consequences* of the breach to the Supervisory Authority (the COAG Health Council) and affected subjects. We use the propositions shown in Figure 5 to model these conditions accordingly.

4.2 GDPR Metagraph

Likewise, we modeled the privacy policies outlined in the GDPR. The resulting conditional metagraph is depicted in Figure 6. According to the GDPR, access to personal data by a recipient is conditional on (a) the subject or their representative providing explicit consent for purposes such as to view, manage or for research; or (b) access being in the interest of the public; or (c) access being required for duties of official authority; or (d) access is by the subject's employer and is in the

subject's interest. We use the propositions *purpose*, *consent* and *subject_age* to model these access conditions.

The GDPR also aims to preserve several rights of the data subject. Namely, the right to (a) rectify a subject's incomplete or incorrect data; (b) erase all subject's data; (c) restrict processing of subject's data for a particular purpose; (d) withdraw consent; and (e) access a subject's own data. We describe conditions (a) to (d) using the propositions *rectify_data*, *erase_data*, *restrict_processing* and *withdraw_consent* respectively. Similarly, the subject right (e) is described using the proposition *data_access*. The proposition indicates that the subject can request from the controller, if their data is processed, if so for what purpose and request a copy of the subject's data for inspection.

The GDPR, also bestows several obligations on the controller. For one, the controller must report alterations to a subject's data (*e.g.,* due to rectification or erasure) to all affected data recipients. This obligation is captured by the proposition *subject_data_altered*. Also, the controller must inform a subject of any data breach which affects their rights or freedom *without delay*. The controller must also report all data breaches (regardless of consequence) to the Supervisory Authority within 72 hours [8]. We use the propositions *nature*, *scope* and *consequences* to describe these obligations respectively.

5 RECONCILIATION OF THE TWO POLICIES

The GDPR and My Health Record metagraphs appear similar but there are differences. An important question is "are the differences important?" We show in the rest of this paper through mathematical analyses of the metagraphs that these differences have significant implication on the privacy of personal health data.

A first step to reconciling a domestic E-health policy against the GDPR is to identify the privacy-goal granularity requirements per E-health use case. We described in Section 3 the use cases applicable to My Health Record. Once the policy descriptions are constructed, we model them using *conditional metagraphs* as per Section 4.

5.1 Policy Definition and Formal Reconciliation

We use *MGtoolkit* [17] – a package we wrote for implementing metagraphs – to instantiate our policy metagraph models. *MGtoolkit* is implemented in Python 2.7. The API allows users to instantiate metagraphs, apply metagraph operations and evaluate results.

The toolkit provides a `ConditionalMetagraph` class which extends a `Metagraph` and supports proposition attributes in addition to variables. A `ConditionalMetagraph` inherits the base properties and methods of a `Metagraph` and additionally supports methods to check reachability properties and redundancy properties. We use the `ConditionalMetagraph` class to instantiate our My Health Record policy models. We then invoke the API methods to reconcile the metagraphs.

We use the property *dominance* [4] to reconcile policy metagraphs. Dominance can be introduced constructively as follows:

Definition 2 (Edge-dominant Metapath). *Given a metagraph $S{=}\langle X, E\rangle$ for any two sets of elements B and C in X, a metapath $M(B, C)$ is said to be edge-dominant if no proper subset of $M(B, C)$ is also a metapath from B to C.*

Definition 3 (Input-dominant Metapath). *Given a metagraph $S{=}\langle X, E\rangle$ for any two sets of elements B and C in X, a metapath $M(B, C)$ is said to be input-dominant if there is no metapath $M'(B', C)$ such that $B' \subset B$.*

In other words, edge-dominance (input-dominance) ensures that none of the edges (elements) in the metapath are redundant. Based on these concepts, a dominant metapath can be defined as follows:

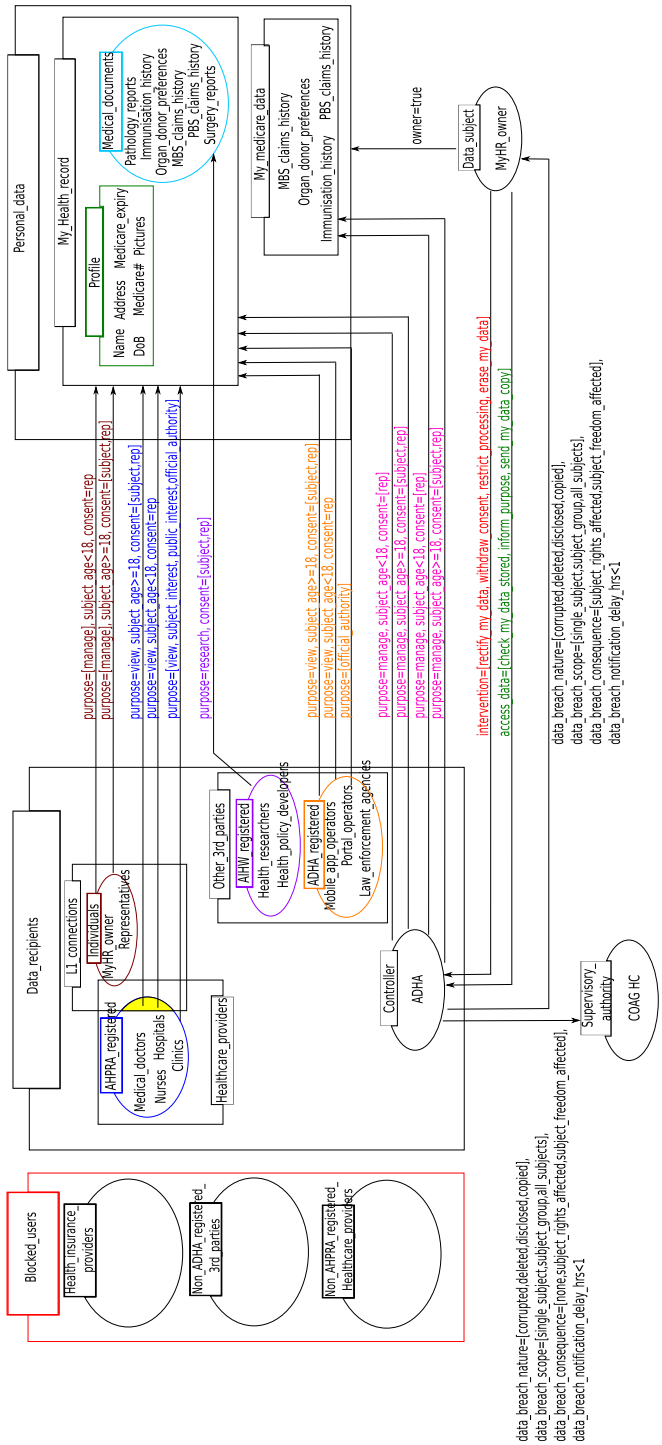

Fig. 5. *My Health Record policy metagraph describing conditions for accessing medical-record data by all recipients.*

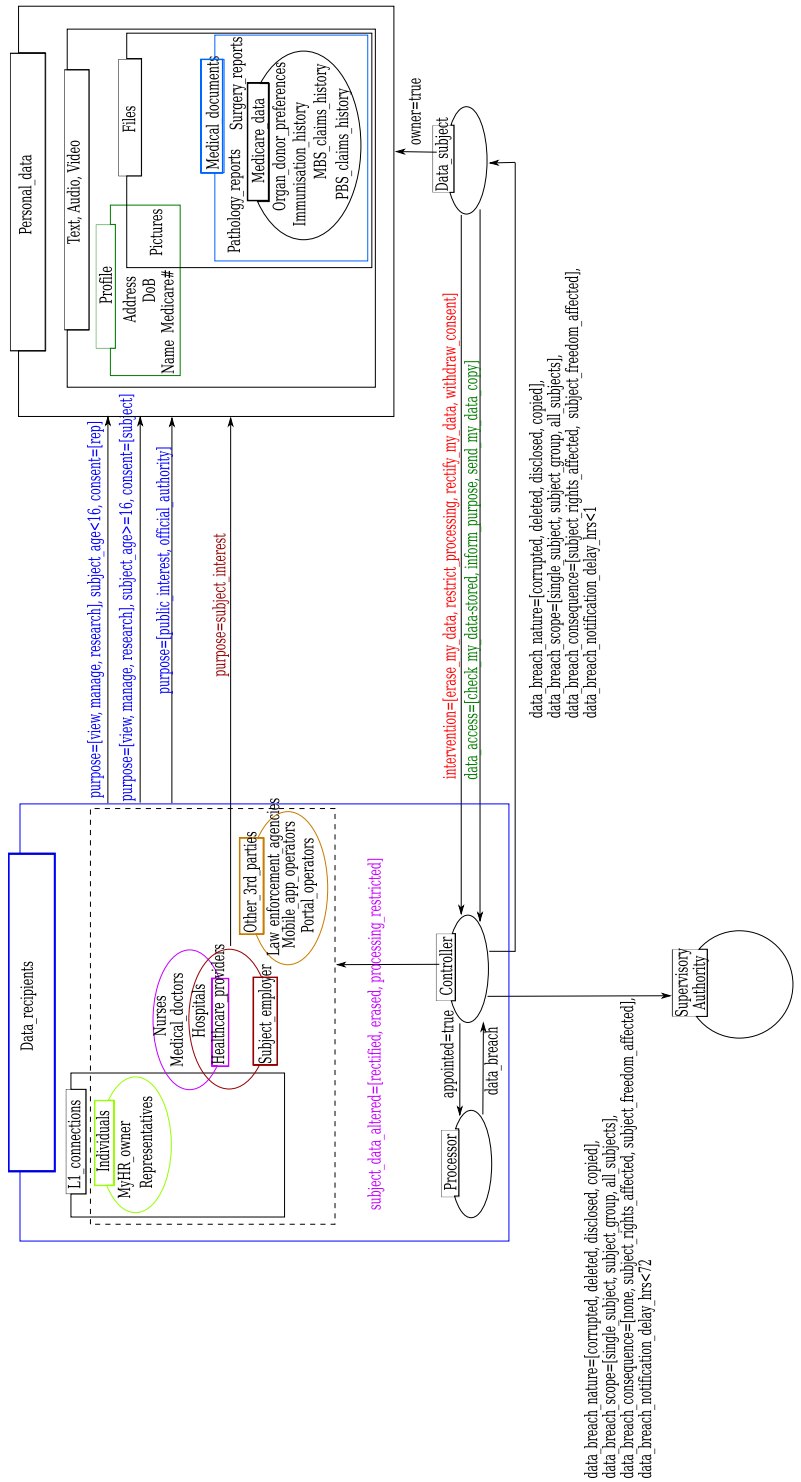

Fig. 6. *European Union's General Data Protection Regulation (GDPR) policy metagraph describing conditions for accessing a subject's personal data.*

Definition 4 (Dominant Metapath). *Given a metagraph $S=\langle X, E\rangle$ for any two sets of elements B and C in X, a metapath $M(B, C)$ is said to be dominant if it is both edge dominant and input-dominant.*

We consider only dominant metapaths when reconciling policy metagraphs because such metapaths provide the least-restrictive access enabled between a data recipient and a data class/category. Least-restrictive access corresponds to the minimal propositions set; where redundant elements exist in a metapath (edge or input element wise) it does not yield this minimal set.

We test if a policy complies with another, by evaluating if it is equally or more restrictive. In that context, we define an *inclusive metapath* as follows

Definition 5 (Inclusive Metapath). *Let $S=\langle X, E\rangle$ and $S'=\langle X', E'\rangle$ be conditional metagraphs and $B, C \subseteq X$ and $B', C' \subseteq X'$. A dominant metapath $M(B, C)$ in S is included by a metapath in S' if and only if there exists at least one dominant metapath $M'(B', C')$ in S' such that $B \subset B'$, $C \subseteq C'$ and $Propositions(M') \subseteq Propositions(M)$.*

Thus, a policy complies with another if each of its dominant metapaths is an inclusive metapath. A non-inclusive metapath indicates a potential policy violation; by finding such metapaths in the My Health Record policy we can detect instances where it violates the GDPR. Likewise, we can detect omissions in the My Health Record policy by applying the process in reverse; *i.e.,* by finding non-inclusive metapaths in the GDPR. For example, the metapath from "Medical Doctors" to "Personal Data" in the My Health Record metagraph is non-inclusive in the GDPR graph. This metapath represents a violation. The metapath from "Controller" to "Healthcare Providers" in the GDPR graph is non-inclusive in the My Health Record graph. This path represents an omission. Complete analysis of the two graphs is provided in the next section.

5.2 Results and Discussion

We ran our policy reconciliation tool on a standard desktop computer (an Intel Core CPU 2.7-GHz computer with 8GB of RAM running Mac OS X). Table 2 shows the number of compliances and violations found in the data processing enabled by the My Health Record policy.

We identified GDPR-compliant processing instances by determining inclusive metapaths in the My Health Record policy. Likewise, GDPR violations were identified using non-inclusive metapaths. Our reconciliation algorithm consists of two steps. In the first step, we check all the dominant metapaths in the My Health Record graph to see if they are inclusive in the GDPR graph. Inclusive paths indicate compliance and non-inclusive paths indicate violations. We then run the reverse check to determine if all dominant metapaths in the GDPR graph are inclusive in the My Health Record graph. For each path, inclusive means compliance and non-inclusive indicates omission.

Our analysis found that for instance, the data processing enabling an individual to represent a My Health Record owner is compliant with the GDPR. The privacy policy in My Health Record ensures that a representative can only be assigned with the subject's explicit consent.

In contrast however, data processing enabling healthcare providers to care for a subject in My Health Record, is not GDPR compliant. The violation stems from healthcare providers being allowed access to a subject's medical records by default. This is a serious flaw; personal data is analogous to 'toxic waste' and healthcare data has highest toxicity. The collection, storing and processing of such 'waste' must be minimized to reduce contamination. So, giving healthcare providers who are not treating a subject access to all their medical records is unnecessary, reckless and can only promote data misuse! The policy violates the GDPR's protection goals of privacy by design and data minimization, and should be rectified.

The last column in Table 2 describes the types of data processing enabled in the GDPR which are ignored in the My Health Record policy. These omissions are an important consideration in

Table 2. *Metagraph-based reconciliation summary of the My Health Record policy against the GDPR. Each metapath source and target identifies nodes in the My Health Record policy metagraph in Figure 5.* # Compliances *indicate occurrences of GDPR-compliant data processing enabled between the source and the target.* # Violations *indicate the number of GDPR breaches (data processing allowed by My Health Record but not by GDPR) and* # Omissions *indicate the occurrences of data-processing enabled in the GDPR, but absent in My Health Record.*

Metapath source	Metapath target	#Compliances	#Violations	#Omissions
Individuals	My Heath Record	2	0	0
Healthcare providers	My Health Record	4	2	0
ADHA registered 3rd parties	My Health Record	3	0	0
AIHW registered 3rd parties	Medical documents	2	0	0
Data subject	Controller	7	0	0
Controller	My Health Record	2	0	0
Controller	My Medicare data	2	0	0
Controller	Data subject	4	0	0
Controller	Supervisory Authority	1	0	0
Controller	Healthcare providers	0	0	3

privacy-policy reconciliation and we identify them using non-inclusive metapaths in the GDPR metagraph model. We find that alarmingly, the controller's obligation in the GDPR, to notify all affected data recipients when a subject's data has been rectified, erased or processing restricted, is not upheld in the My Health Record policy. This omission means that a healthcare provider for instance, may not be able to stop the administering of a medication or treatment prescribed based on an erroneous blood report of a subject, if this has been updated but not checked.

6 CONCLUSION AND LIMITATIONS

Reconciling privacy policies helps to reduce misuse of personal data, but the task is non-trivial because privacy policies are often written in free text, making them subject to human interpretation. In this paper, we propose a tool which allows to describe E-health privacy policies granularly, represent them using formal constructs, making policies precise and explicit. Our tool can automatically reconcile an E-health policy against the GDPR to identify violations and omissions. We use our prototype to illustrate several critical flaws in Australia's My Health Record policy.

The research has several limitations. First, we assume that the My Health Record policy needs to comply with the GDPR. GDPR is a European legislation and as a sovereign nation, Australia's My Health Record is not required to be GDPR compliant. There may be political reasons behind the different requirements that we did not explore in this research. Second, at the moment, the metagraph model for each policy has to be developed manually. This process could be time consuming. An automatic tool that allows auto-parsing of the policies would make the tool more scalable and applicable to dynamic policies. Third, we have shown that metagraphs have great potentials in modeling and analyzing policies. In this work, we have concentrated mostly on dominant metapath properties. Other mathematical properties associated with metagraphs should be explored in future work for more advanced analyses.

LIST OF ACRONYMS

GDPR European General Data Protection Regulation

EPAL Enterprise Privacy Authorization Language
RBAC Role-Based Access Control
XACML eXtensible Access Control Markup Language
ADHA Australian Digital Health Agency
MBS Medicare Benefits Schedule
PBS Pharmaceutical Benefits Scheme
COAG Council of Australian Governments
AHPRA Australian Health Practitioner Regulation Agency
AIHW Australian Institute of Health and Welfare

REFERENCES

[1] ABC News. 2018. Health service providers suffer the most data breaches, as overall numbers jump. [Online]. Available: www.abc.net.au/news/science/2018-07-31/information-commissioner-health-sector-leads-data-breaches.

[2] Paul Ashley, Satoshi Hada, Günter Karjoth, Calvin Powers, and Matthias Schunter. 2003. Enterprise privacy authorization language (EPAL 1.2). *Submission to W3C* (2003).

[3] Australian Digital Health Agency. 2018. My Health Record - Keep track of your important health information all in one place. [Online]. Available: https://www.myhealthrecord.gov.au/.

[4] Amit Basu and Robert W Blanning. 2007. *Metagraphs and their applications*. Vol. 15. Springer Science & Business Media.

[5] Tim Berners-Lee. 2018. You own your data, and choose apps to manage it. [Online]. Available: https://solid.inrupt.com/how-it-works.

[6] Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. 2002. The platform for privacy preferences (P3P1. 0) specification. *W3C* (2002).

[7] Ebay. 2018. User Privacy Notice. [Online]. Available: www.ebay.co.uk/pages/help/policies/privacy-policy.html.

[8] European Commission. 2016. General Data Protection Regulation (GDPR). [Online]. Available: https://gdpr-info.eu/.

[9] Facebook. 2018. Data Policy. [Online]. Available: www.facebook.com/policy.php.

[10] Simon Godik and Tim Moses. 2002. Oasis eXtensible Access Control Markup Language (XACML). *OASIS Committee Specification cs-xacml-specification-1.0* (2002).

[11] Ayyoob Hamza, Dinesha Ranathunga, Hassan Habibi Gharakheili, Matthew Roughan, and Vijay Sivaraman. 2018. Clear as MUD: Generating, Validating and Applying IoT Behavioral Profiles. In *Proceedings of the 2018 Workshop on IoT Security and Privacy*. ACM, 8–14.

[12] Marit Hansen, Meiko Jensen, and Martin Rost. 2015. Protection goals for privacy engineering. In *Security and Privacy Workshops (SPW)*. IEEE, 159–166.

[13] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 47–64.

[14] NewScientist. 2017. Laws of mathematics don't apply here, says Australian PM. [Online]. Available: www.newscientist.com/article/2140747-laws-of-mathematics-dont-apply-here-says-australian-pm/.

[15] Jillian Oderkirk, Elettra Ronchi, and Niek Klazinga. 2013. International comparisons of health system performance among OECD countries: opportunities and data privacy protection challenges. *Health Policy* 112, 1-2 (2013), 9–18.

[16] Harshvardhan J Pandit, Declan O'Sullivan, and Dave Lewis. 2018. An Ontology Design Pattern for Describing Personal Data in Privacy Policies.. In *WOP@ ISWC*. 29–39.

[17] D Ranathunga, H Nguyen, and M Roughan. 2017. MGtoolkit: A python package for implementing metagraphs. *SoftwareX* 6 (2017), 91–93.

[18] D Ranathunga, H Nguyen, and M Roughan. 2020. Verifiable Policy-Defined Networking using Metagraphs. *IEEE Transactions on Dependable and Secure Computing*. Preprint (2020), 1–15.

[19] Reuters. 2017. U.S. appeals court blocks D.C. law restricting gun rights. [Online]. Available: www.reuters.com/article/us-usa-guns-washingtondc-idUSKBN1AA27U.

[20] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics (IWSPA '18)*. Association for Computing Machinery, New York, NY, USA, 15–21. https://doi.org/10.1145/3180445.3180447

[21] W. B. Tesfay, J. Serna, and K. Rannenberg. 2019. PrivacyBot: Detecting Privacy Sensitive Information in Unstructured Texts. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. 53–60. https://doi.org/10.1109/SNAMS.2019.8931855

[22] W3C. 2006. The Platform for Privacy Preferences 1.1 (P3P1.1) Specification. [Online]. Available: www.w3.org/TR/P3P11/.
[23] Yajing Wang. 2019. *A comparative study of Chinese and European Internet companies' privacy policy based on knowledge graph.* Master's thesis. University of Turku, Finland.
[24] Carl Yang, Yichen Feng, Pan Li, Yu Shi, and Jiawei Han. 2018. Meta-Graph Based HIN Spectral Embedding: Methods, Analyses, and Insights. In *2018 IEEE International Conference on Data Mining (ICDM)*, Vol. 2018-. IEEE, 657–666.

APPENDIX - GRANULARITY DIMENSIONS

We provide details of the granularity dimensions in our model in this section. High-level description was provided in Section 3.

Data class describes the type of healthcare data available for processing. We define these types in decreasing granularity:

- **Text, Numeric, Audio, Video, Image** are fundamental data types used to describe healthcare data.
- **File** is a collection of fundamental data identified by a filename. A file can be a document, picture, audio or video data etc. Enabling processing of this type of data may be required when providing healthcare to the subject.
- **Directory** is a collection of fundamental data and files. Enabling processing of this type of data is required when representing a subject.

Data category describes the context of the data processed. Each category can comprise of multiple data classes. Categories applicable to E-health are

- **Profile** which describes an individual's name, DoB, address, picture and other identification information.
- **Documents** which describe a subject's medical reports such as pathology reports, surgery reports, current medications, allergies and immunization history etc.
- **History** which describes E-health platform activity history and subject's (or representative's) interaction history with the controller (*i.e.,* queries and resolutions).
- **Access control** which describes access codes enabling medical record access by individuals, healthcare providers etc.
- **Personal notes** which describe health-diary entries which are only accessible by the subject and his/her representatives.

Processing purpose differs from a use case in that the latter may require data processing for multiple purposes; *e.g.,* providing healthcare to a subject requires data processing for *view, subject interest, public Interest* and *official authority.*

Data recipient As per the GDPR, a ***data recipient*** describes an entity who is allowed to process personal data. In E-health, recipients who can process medical records can be

- **Individuals** are the subject and his/her representatives.
- **Healthcare providers** are individuals and organizations who provide healthcare services in general. In My Health Record, only providers registered with the Australian Health Practitioner Regulation Agency (AHPRA) are allowed to access a subject's data. This prerequisite restricts overseas healthcare providers from accessing this data.
- **Subject employer** is the organization which employs the medical-records owner (if employed).
- **Third parties** include other individuals and organizations such as mobile application operators and law enforcement agencies. In My Health Record, only those third parties registered with the Australian Digital Health Agency (ADHA) are permitted to access medical records.

Data-recipient category describes the type of entity allowed to process the data. Each recipient category (listed below) can comprise of multiple data recipients.

- **Level-1 connections** are a subject's representatives and their immediate healthcare providers.
- **Indirect connections** are healthcare providers and third parties who do not provide direct services to the subject.
- **Blocked users** are individuals and organizations who are prohibited from accessing a subject's medical records.
- **Controller** is an organization which determines the purpose and means of data processing. In My Health Record, this entity is the Australian Digital Health Agency.
- **Processor** is an organization appointed and directed by the controller to collect and/or process medical records.
- **Supervisory authority** is an independent public authority responsible for monitoring the application of E-health privacy policies. In My Health Record, this entity is the Council of Australian Governments (COAG) Health Council.

Consent provider describes the entity providing consent for purposes which require explicit consent. Entities applicable in E-health are

- **Subject** *i.e.,* the medical-records owner.
- **Representative** an individual who manages medical records on behalf of the subject.
- **subject's age**

Data breach nature describes the type of data breach occurred. These types in decreasing granularity are:

- **Disclosed** means the medical records were only viewed by an unauthorized party. A subject or their representative needs to be informed of breaches of this granularity or finer by law [3].
- **Copied** means the data was additionally copied.
- **Corrupted** means the data was additionally modified.
- **Deleted** means the data was additionally deleted.

Data breach scope describes the extent of a data breach. The levels below describe this scope in decreasing granularity.

- **Single** means the breach affected only one subject.
- **Group** means the breach affected multiple subjects.
- **All** means all subjects in the E-health platform were affected.

Data breach consequence describes the impact of the data breach in terms of what subject rights were violated. We describe these impacts below in decreasing severity.

- **Subject freedom affected** indicates that the subject's freedom was violated.
- **Subject rights affected** indicates that the subject's rights were violated.
- **None** indicates that neither the subject's rights nor freedom were violated.