Postprint Version	1.0
Journal website	http://dx.doi.org/10.1016/j.ijmedinf.2014.08.009
Pubmed link	http://www.ncbi.nlm.nih.gov/pubmed/25241154
DOI	10.1016/j.ijmedinf.2014.08.009

A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model

WOLFGANG KUCHINKE[A,*], CHRISTIAN OHMANN[A], ROBERT A. VERHEIJ[B], EVERT-BEN VAN VEEN[C], THEODOROS N. ARVANITIS[D], ADEL TAWEEL[E], BRENDAN C. DELANEY[E]

[a] Coordination Centre for Clinical Trials, Heinrich-Heine-University, Düsseldorf, Germany
[b] NIVEL, Utrecht, The Netherlands
[c] MedLawConsult, Den Haag, The Netherlands
[d] University of Warwick, Coventry, UK
[e] NIHR Biomedical Research Centre at Guy's and St. Thomas' NHS Foundation Trust and King's College London, London, UK

HIGHLIGHTS

- Easy model to display policies and rules for data privacy.
- Novel concept of privacy zones for research data flow.
- A risk gradient runs from high risk to low risk for patient identification.
- The zone model is presented for important research scenarios.
- Different types of research are considered in each of the three zones.

ABSTRACT

Purpose: To develop a model describing core concepts and principles of data flow, data privacy and confidentiality, in a simple and flexible way, using concise process descriptions and a diagrammatic notation applied to research workflow processes. The model should help to generate robust data privacy frameworks for research done with patient data.

Methods: Based on an exploration of EU legal requirements for data protection and privacy, data access policies, and existing privacy frameworks of research projects, basic concepts and common processes were extracted, described and incorporated into a model with a formal graphical representation and a standardised notation. The Unified Modelling Language (UML) notation was enriched by workflow and own symbols to enable the representation of extended data flow requirements, data privacy and data security requirements, privacy enhancing techniques (PET) and to allow privacy threat analysis for research scenarios.

Results: Our model is built upon the concept of three privacy zones (Care Zone, Non-care Zone and Research Zone) containing databases, data transformation operators, such as data linkers and privacy filters. Using these model components, a risk gradient for moving data from a zone of high risk for patient identification to a zone of low risk can be described. The model was applied to the analysis of data flows in several general clinical research use cases and two research scenarios from the TRANSFoRm project (e.g., finding patients for clinical research and linkage of databases). The model was validated by representing research done with the NIVEL Primary Care Database in the Netherlands.
Conclusions: The model allows analysis of data privacy and confidentiality issues for research with patient data in a structured way and provides a framework to specify a privacy compliant data flow, to communicate privacy requirements and to identify weak points for an adequate implementation of data privacy.

1. INTRODUCTION

Clinical research has led to a growing demand for data from health records and subsequent data sharing. For research into the health status of populations, the aetiology of diseases and the effectiveness of medical treatments, increasingly access to large patient databases and registers is required. Because health research deals with human data, which is subject to special protection, research can take place only within an appropriate regulatory and data privacy framework. For some years primary care databases already exist on a national (e.g., NIVEL Primary Care Database, NIVEL-PCD) or regional scale (e.g., databases in the Maastricht area) in some European countries. These patient databases often act not only as storage site for data, but offer research services [1]. Consequently, the need has arisen to use these services in research projects to answer complex research questions. In addition, different service providers are increasingly responsible for the storage, processing and integration of patient data leading to the problem of sensitive data stored on systems that are not under the control of the entity which submitted the data [2].

Privacy legislation has been slow on reacting to the increasing role of research service provisions and the use of personal health records in research. Indeed, the suitability of conventional privacy requirements has been questioned on this basis [3] and [4]. Any simplistic “global” solution such as banning access to all data without explicit consent can hamper research by excluding “sensitive topics” or biasing results by omitting hard to reach groups such as the poor. More complex solutions for privacy protection involving safe havens or third party linkages of data may be required.

In order to understand complex data, privacy requirements and dependencies, a standardised notation to allow the presentation of privacy needs associated with different research questions could improve understanding and clarity. Our aim was to

create an easy to use graphical method for describing data privacy frameworks and apply to it relevant research data flow scenarios.

2. BACKGROUND

TRANSFoRm (Translational Medicine and Patient Safety in Europe) [5] is a project partially funded by the European Commission developing the digital infrastructure for a "Learning Healthcare System" in Europe consisting of data collection, data mining and decision support that aims to improve both patient safety and the conduct and volume of clinical research [6]. Although IT systems in primary care settings (e.g., general practices) are a large source of electronic clinical data at patient level, data are often located in a multitude of general practice systems, as well as primary care databases derived from those record systems. In TRANSFoRm two clinical use cases comprising of a genotype-phenotype study and a randomised controlled trial with patient-related event-driven outcome measurements necessitate the requirement for the integration of heterogeneous data from different data sources. Each data source may be subject to different privacy and data security requirements. Both use cases represented a useful source to extract privacy and data flow requirements to apply to our privacy model development.

As background research we carried out an analysis of existing data privacy frameworks, data protection infrastructures and applicable regulations for data access, exchange and linkage. In Europe, as well as in the US, data privacy protection frameworks exist that define as general rules, how to protect personal data (e.g., EU Data Protection Directive [7], HIPAA (Health Insurance Portability and Accountability Act) [8] and [9], OECD (Organisation for Economic Co-operation and Development) Privacy Framework [10], APEC (Asia-Pacific Economic Cooperation) [11], Madrid resolution 2009 [12], and US/EU Safe Harbor Agreement [13]). In this context, the EU Data Protection Directive (Directive 95/46/EC) [7] defines the protection of the processing of personal data and health data in Europe; and for this purpose institutes the role of a "data controller". The Data Protection Directive requires data controllers to assure a number of principles (e.g., legitimate purpose, accuracy of data) when they process personal data and to protect personal data against accidental or unlawful destruction, loss, alteration and disclosure. But the implementation of the directive via national legislation has not always been consistent in European member states, because of variability in interpretation in some areas (e.g., definition of anonymisation, informed consent and research exemption) [14]. As a consequence, access to patient data for research is hampered by a fragmented legal European framework, inconsistency in interpretation of regulations, variable guidance and a lack of clarity among investigators, regulators, patients and the public [15] and [16]. In the US a federal data protection standard, known as the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, is applied to protect the privacy of personally identifiable information [17] and [18]. But, like in Europe, the interpretation of the HIPAA rule has not been uniform and reconciliation with other federal regulations is difficult, leading to a patchwork of privacy protections that is neither comprehensive nor easy to apply [19] and [20].

Although data privacy frameworks that build upon legal requirements and rules exist, international research with primary care data is still being hampered by a lack of a universally agreed definition of privacy, and by the fact that privacy is contextual. International privacy frameworks are lacking impact because they are often only non-binding, technology neutral guidances, listing general basic principles for the processing of personal data [21]. Often these privacy frameworks are dedicated only to specific countries (e.g., US [8], [22] and [23]), specific diseases (e.g., cancer [24] and [25]), specific situations (e.g., only research context with informed consent by the patient [26]), research context with anonymous data or to specific data sources (e.g., only use of secondary data [22] and [23]), requiring high aggregation level of data for researchers [26].

An outcome of our analysis of the legal background was that a privacy framework for medical research should consider that Primary Care data routinely stored on computers either within general practices or at national or regional databases can be linked to other healthcare datasets, like hospital admissions records, death certificates, and disease registries [27] and [28]. Any privacy protection framework should consider these activities and should allow as much research as possible and at the same time, protect patient's privacy efficiently. In this paper a generic graphic model that enables the easy representation of core concepts and principles of data privacy in the use of health data for research was developed, able to distinguish the various phases of the research data flow and its risk of identifyability.

3. OBJECTIVES

To support the development of privacy compliant data flow for research, a generic graphic model should enable the easy representation of core concepts and principles of data privacy and confidentiality in the use of health data for research. This graphic model should distinguish the various phases of the research data flow from the primary sources until data reach the researcher for analysis. The secondary objective was that applying the model to common research data flow scenarios should further the understanding of privacy protection requirements, to enable research with primary care data covering the different privacy needs for research with Primary Care data, Electronic Health Records (EHRs), clinical research databases (Case Report Form based data), and data stored in genetic and cancer registries.

4. METHODS

4.1. Model building

4.1.1. Creation of a graphic model

Software engineering knows different models, like data models, information models, process and component models that are used as basis for software development. A model may represent an artefact describing a system through the help of suitable diagrams [29]. We used the model approach to better understand the privacy and data protection requirements of the different research use cases in the TRANSFoRm project and to find a way to depict the context dependency of privacy in a graphical

way. Because in our approach the model is understood as a group of independent elements that act together, a common language and specific symbols was needed to describe privacy protection requirements and to distinguish different contexts of privacy. We used the Unified Modelling Language (UML) [30], which is commonly employed in requirements engineering as basis and enriched it with additional workflow and own symbols to adapt the notation to the extended data flow/data linkage requirements of the research use cases. A model based privacy threat analysis [31] for common research scenarios supported the elicitation of privacy requirements for TRANSFoRm.

The following steps towards the creation of a graphic model were performed: definition of the system in question (health care research domain), identification of relevant elements and features (based on the workflow and data flow of TRANSFoRm clinical use cases), definition of risks to privacy and methods to deal with these risks, and finally building a conceptual model [32] using these elements. Because graphical modelling is known to be an important method to efficiently represent and analyse information in complex systems [33], we applied it to the process of privacy framework generation.

4.2. Exploration of privacy frameworks and access policies to primary care databases

Data privacy frameworks [34], [35], [36], [37], [38] and [39], including data access and data sharing policies of European primary care databases (e.g., NIVEL-PCD in the Netherlands [27], CPRD in the UK [28], etc.) were analysed. For research projects, these privacy frameworks often apply the most stringent approach to control research data flow, requiring combinations of explicit consent, restrictive definitions of anonymisation, data encryption and data access contracts. Such a restrictive approach can result in difficulties with aligning data flow requirements with research process needs. In general, a more flexible approach is needed; our graphic representation method to describe data flows [40] may aid development of suitable structures able to guarantee both, privacy of patient data and research, with as little restrictions as necessary.

5. RESULTS

5.1. Formal description of the model

Our model employs the basic definitions of schemes for data types based on the EU Data Protection Directive (Table 1). To account for these definitions, a set of basic elements, like zones for data sources (zone, subzone), operators for transforming data (data linker, privacy filter) and actors/roles (General Physician (GP), researcher) (Fig. 1), was created. The zone plays a central role in our model; it ensures that context sensitivity of privacy protection is always considered. A single rule applies in our model: the zones should always be oriented in such a way that data flows from a zone with a high risk of patient identification to a zone with a lower risk.

[TABLE 1] [FIGURE 1]

Zones are areas containing data sources controlled by similar policies and applicable regulations (Table 2). Three main zones, with a decreasing risk of patient identification are introduced: Care Zone, Non-care Zone and Research Zone. The **Care Zone** is the area of patient diagnosis and treatment. In general, patient identifyability constitutes the basis for any medical treatment and for the special trust relationship between patient and GP (family doctor). In the Care Zone personal identifiable medical data is stored and used within the care context by the treating physician. **Personal data** are data which relate to an identified or identifiable natural person. An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors relating to its physical, physiological, mental, economic, cultural or social identity.

[TABLE 2]

Once data is outside the Care Zone and not anymore protected by medical confidentiality, patient data is in need for a different kind of protection. For example, EHR data can be transferred to primary care databases. The **Non-care Zone** contains such research databases (e.g., clinical trials, cohort studies, etc.), and secondary use databases that have been derived from primary medical care data. Data providers maintaining data in this Non-care Zone often offer data services and usually make use of policies that employ strict access control allowing access to their data only in pseudonymous or anonymous form. In addition, access may be based on explicit consent (by presenting the patient with an option to agree or disagree with the collection, processing, or disclosure of personal information) or on country-specific or local regulations (e.g., exemptions to consent for research) usually allowing for an opt-out regime. For example, in case of NIVEL-PCD, patients can object to the collection of their data in the Care Zone for research purposes. This may result in a problem, because once patient data is transferred to the Non-care Zone, it is often no longer possible for database owners to identify patients that have objected to the use of their data. These preconditions have to be addressed by suitable consent forms allowing the patient to opt-out for “future use of their data”, with the restriction that the data already used in a study cannot be retrospectively removed (e.g., from a publication, or a running analysis). Recently in some cases general opt-out regime has been criticised (e.g., in the UK). More and more physicians demand that doctors have a duty to maintain patient confidentiality and that therefore the ‘opt-in’ system, as opposed to ‘opt-out’ system would be the method of choice. However, opt-in regimes usually result in fewer patients agreeing to have their information made available and questions around the quality of information and consent may be raised [41].

In the **Research Zone** the researcher receives data suitable for processing and analysis in specific research projects, addressing specific research questions with approved protocols. For this purpose, the Research Zone may receive data in anonymous or at least “coded anonymous” form. In this context, “de-identified” data refers to data that have been stripped of all direct subject identifiers, like name and birth date; but in this case, each record may have its own ID (e.g., a pseudonym),

which is the link to the identifiable information stored elsewhere, such as name or medical record number. Coding can be done “one-way” or “two-way”. With one-way coding, it is always possible to translate the identifier (ID) into a code number (CN), but not the other way around. Thus, one way coding can be considered as anonymous and irreversible, because it is not possible to go back from the CN to the ID. With two-way coding, the latter is still possible [26]. Although the Research Zone offers the least restrictions to data users, it is not an unrestricted area; researchers are bound by codes of conduct, the scrutiny of their peers [26] and by agreements under which data were transferred to them, and the privacy protection policies of their institutions [42].

Any international privacy framework must consider that databases in different countries may operate under different rules and regulations concerning confidentiality and data privacy. Even in one country differences need to be considered (e.g., between England and Scotland). To represent this heterogeneity, **subzones** within the main zones (i.e. Care Zone, Non-care Zone and Research Zone) were introduced. Each subzone is homogenous in terms of rules and regulations and the extent to which individuals in each subzone are identifiable. As components of the Zones/Subzones, our model considers the roles of **informed consent**, **contractual agreements** and **database statutes** that regulate the transfer of data within and between zones. The significance of these components is rather generic and open to interpretation on the risk to privacy.

There exist several symbols that may stand for privacy (mask, lock, etc.); but symbols for privacy functions are absent from the common notations for data flow, UML, DIN ISO 1219 (Fluid power systems and components – Graphic symbols and circuit diagrams) and business processes. Therefore, we developed several functional symbols that represent privacy, ensuring functions, like de-identification, coding, aggregation. Two types of new symbols play an important role for the research data flow: **privacy filters** that operate on data and **data linkers** that allow the connection of databases within or between zones/subzones. We do not define the functions of each privacy filter and linker in detail, and leave an elaboration of the functions and algorithms to an update of the model. Linking can only be performed if one set of data relating to an individual has the same pseudonym as that same individual in another dataset. The use of irreversible pseudonyms allows the **linkage** of records for the same individual and simultaneously anonymising these records (e.g., by “fuzzy matching” methods).

Pseudonymisation is the process by which direct identifiers of the data subject, like the name and birth date, are removed and replaced by a unique number, the pseudonym. Another expression of pseudonymisation is coding of data. In contrast, **anonymous data** cannot be used to identify the person to whom the data relates. In case of data linkage from different data sources, adequate precautions must ensure that subjects are not identifiable due to any combination of data available. Even a set of laboratory values may become identifiable in a specific context. In this case, but especially in the case of linking a data set to **genomic data,** the need for pseudonymisation of indirectly identifiable data must be considered. **Actors** involved in the research data flow include: patient, clinician/GP (which may have the double

role of a treating physician and/or researcher), the EHR, databases, and data controller.

The following additional definitions apply. Data controller is the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of the data concerned in the research. It should be noted that "controller" is used in a broader meaning than in article sub d of Directive 95/46/EC [7]. In the Directive, the term "controller" applies only to personal data. As the Directive is not applicable to anonymous data, it lacks a term for the entity responsible for a database with anonymous data. Data linking is merging data from different databases that can be done on the record level by using common identifiers or pseudonyms. Privacy filters are software tools to render data less identifiable or even anonymous to a subsequent controller (as defined in this paper) of these data.

5.2. CHARACTERISATION OF SUBZONES

Because in TRANSFoRm use cases (Table 3), different databases located in different regions and countries are accessed, the subdivision in only three zones (Care Zone, Non-care Zone and Research Zone) turned out to be insufficient to consider local and national requirements. Therefore, a new element was introduced: the Subzone. Two examples for research involving databases in different subzones were worked out in detail, indicating how research is possible in the Non-care Zone and the Research Zone (Fig. 2).

5.2.1. Example 1: Subzones defined by differences in regulations: German federated state-specific legislation applied to the use of clinical care data for research (Fig. 2A)

For demonstrating the subzone concept, the highly fragmented health care system of Germany lends itself to be modelled. Here, federated state-specific legislations have to be taken into consideration when using clinical care data for research purposes within the institution where these data have been collected. Altogether, six states allow the use of personal data for research within the hospital where data have been collected (research exemption) (**Subzone A**). In the city-state of Berlin the use of clinical care data for research is only allowed after anonymisation (**Subzone B**). In ten states the use of clinical care data for research is possible without consent, after a risk assessment results in favour of research and anonymisation of data is performed (**Subzone C**). Using our notation, it becomes clear that the interpretation of research purposes is context dependent and relates to the subzones involved in the data flow; that is to say on the local rules and regulations. Thus, in some cases, a privacy filter (with or without additional rules) may be needed (Fig. 2A) to enable research with patient data. In the care area of a hospital, the situation may be different. Here a privacy filter protects against data access by any person, with the exception of the treating physician; though the definition of the "treating physician" may include others, such as nurses or even administrative staff.

The subzone model considers the fact that researchers can do research in the Research Zone, but may also do research in the Non-care Zone. The second case pertains to all research done for example at a service provider associated with a primary care database. Here, the researcher is bound by special contractual agreements.

5.2.2. Example 2: Subzones defined by differences in the type of data sources: care databases in The Netherlands as resource for health care research (Fig. 2B)

Health care data is the source for the collection of patient data for several non-care databases (Fig. 2B). For example, a health insurance database may contain data about medical consumption, whether statutory or contractual necessary for the reimbursement scheme. Thus, **Subzone Y** covers directly identifiable data by the insurer (D). On the other hand, NIVEL's PCD contains EHR data from about 300 general practices (**Subzone A**). The data at NIVEL-PCD (E) is coded pseudonymised, but may be indirectly identifiable, thus belonging to **Subzone XY** for indirectly identifiable data. In the same way, the national cancer registry (B), which covers about 99% of all cancer patients in The Netherlands contains pseudonymised patient data (**Subzone Y**: indirectly identifiable data). The death registry at Statistic Netherlands (E) is based on the statutory obligation to report the cause of death by physicians (**Subzone Y**: directly identifiable by Statistic Netherlands). In case the database provider conducts research, the research is done in the Non-care Zone. Although different data users were covered, only two subzones are sufficient, one for identifiable data and one for indirectly identifiable data, each allowing a different kind of research according to the rules of the corresponding subzone.

5.3. Representation of research scenarios

Our model can be used to describe, compare and analyse typical research scenarios for research questions with primary care data, such as the merging of data from different EHRs, linkage and merging of data from different subzones in the Care Zone, data transfer from a database in the Non-care Zone to the Research Zone, data enrichment by linkage of databases in the Non-care Zone and linkage of a cohort study database with data from the Care and Non-care Zone. Here we discuss two of these scenarios, in more detail.

5.3.1. Scenario 1: Merging data from different EHRs

In this scenario (Fig. 3A), data from different EHRs of general practitioners (GPs) belonging to the same subzone A (e.g., GPs in the same state/country) within the Care Zone are merged in a research database after passing privacy filters. No linkage on a patient level has to be performed, but there has been a merging of cases. In this case, only anonymous data can be used for research (e.g., the research query may result in counts).

5.3.2. Scenario 2: Linkage and merging of data from different subzones in the Non-care Zone

This scenario (Fig. 3B) represents a situation where two databases in the same subzone of the Non-care Zone, such as a cancer registry and a civic registry, are linked after passing a privacy filter. The resulting database is linked by two-way coding with another secondary database, which itself has to pass a privacy filter. The aim of this scenario is the case, commonly found in research, to enrich a linked database with additional data from a secondary database. The graphic representation of this scenario (Fig. 3) helps in the identification of constraints and limitations on the data flow in this scenario and indicates two locations for research in the Non-care Zone and the Research Zone.

5.4. Application of the model to clinical research use cases

The TRANSFoRm project investigates two different clinical research scenarios: (1) a genotype-phenotype association study of response to oral type 2 diabetic medication using primary care and genomic databases; (2) a randomised controlled trial of on demand vs. continuous proton pump inhibitor treatment in gastro-oesophageal reflux disease (GORD). Our model was applied to both of these clinical research scenarios, containing two use cases: 1. Find patients for clinical trials (by the treating physician, or by using a non-care database); 2. Linkage of databases (linking of databases in the Non-care Zone, or linking of databases in the Care Zone and the Non-care Zone). For these use cases a process description was developed (Table 3) that is used for the application of our model to the use cases. The graphical representations display zones and data flows for all four sub-use cases: “find patients for research by the treating physician” (Fig. 4A), “find patients for research using non-care databases” (Fig. 4B), “linkage of databases within the Non-care Zone” (Fig. 5) and “linkage of databases between Care Zone and Non-care Zone” (Fig. 6). In each case the context, zones, actors, processes, and data privacy issues were identified, analysed and a formal graphical representation was developed.

5.4.1. Use case: Find patients for clinical research

The identification of potential participants for clinical trials may be done in the care context by the treating physician (e.g., during an appointment or a patient visit) or by searching databases in the Non-care Zone (e.g., research database, register). Preconditions for a successful search are the availability of inclusion and exclusion criteria, and of suitable data in the GP's EHR or in primary care databases in the Non-care Zone. Using our model, it is possible to display the different conditions and rules that apply to searches by a physician in their database (Care Zone), as well as to a researcher using a primary care database (Non-care Zone). A privacy filter in the Non-care Zone ensures that neither the controller of the database nor the researcher knows about the identity of a potentially eligible patient discovered within the dataset. The case of finding an eligible patient to recruit for a trial by the treating physician is presented in Fig. 4A. Due to the special physician/patient relationship, only the treating physician is allowed to assess data of his/her patient and to invite this patient for trial participation.

Finding eligible patients for trial participation in Non-care databases (e.g., research database, register) is more complicated than the above case and consists of different

steps, Fig. 4B. The permission for a researcher to obtain access to research databases depends on country regulations (e.g., research exemption, opt-out; pseudonymous data treated as anonymous) and local policies. After the data controller of the database allows the search, potential participants can be identified as pseudonyms, which are retrieved. These pseudonyms have to be sent to the corresponding physician to identify and contact the corresponding patient. In general, only the treating physician is allowed to access identifiable data of his/her patients and to invite the identified eligible patients for trial participation.

5.4.2. Use cases: Linkage of databases

In this scenario, a researcher wants to extract information about selected cases from linked databases. In this context data linkage may be done within a zone or subzone or between different zones or subzones. Involved databases may be the EHR in the Care Zone and research databases or registers in the Non-care Zone. Linking is initiated by the researcher (Fig. 5) and done at the pseudonym level. Transferred data sets are pseudonymised (or coded) anonymous or fully anonymised data. The result is a linked database transferred to the Research Zone, with pseudonymised (or coded) anonymous data (if re-identification is necessary and allowed) or fully anonymised data. Because in this case already pseudonymised data are linked and combined in database C, research on these data is possible in the Non-care Zone, although researchers in the Non-care Zone are bound by additional rules and obligations, in contrast to researchers in the Research Zone.

In the next case a link is generated between the databases in the Care Zone (e.g., EHR database, hospital information system, etc.) with a database in the Non-care Zone (e.g., research database, register) (Fig. 6). In general, coding done with the purpose to link data may be a one-way coding or a persistent recoding, to ensure longitudinal consistency for linkage. Once more, because pseudonymised data are linked and combined in database C, research is possible in the Non-care Zone on database C data, as well as by researchers in the Research Zone, after data pass an additional privacy filter.

5.4.3. Application example: data processing in the Dutch primary care database NIVEL

NIVEL's primary care database (NIVEL-PCD) [43] is one of the larger primary care databases in Europe, holding GP data of 1.6 million individuals. In TRANSFoRm part of this database is used as a data source in the Diabetes use case. We applied our model and its notation to describe the dataflow in the Care Zone, the Non-care Zone and the Research Zone associated with research done in association with NIVEL-PCD. The aim was to obtain a correct graphical representation of the privacy model and to examine whether our zone model and its available filters would be able to describe the data flow and the data protection restrictions and rules in an adequate way.

NIVEL-PCD services are based on the fact that under certain conditions, the use of extracts of Electronic Health Records for research purposes is allowed by Dutch law.

These conditions are described in the code of conduct for health research, issued by the Dutch Federation of Biomedical Scientific Societies [44] and approved by the Dutch Data Protection Agency. Legally this form of data collection is based on research exemption with implied consent by the patient, but with the option to opt-out. This regulation forms the foundation that data recorded by participating GP practices are sent regularly to NIVEL to be processed and entered into the database. Patient data are pseudonymised before leaving the Care Zone using a Trusted Third Party (TTP) (Fig. 7), so that identifiable information does not leave the Care Zone. In this way EHR data is sent from the practices to NIVEL only after having passed a privacy filter. Using these pseudonyms, it is technically possible to link data from different health care and other disciplines, without having to resort to identifiable information.

In detail, extracts from EHR data are generated in the practices, sent to NIVEL and stored in a repository in the Non-care Zone (database A), protected by a pseudonym (Fig. 7). From this master database, data extracts can be made again for specific research projects B1, B2 and B3, following approval by a steering committee. These research extracts are subject to a number of quality checks and a second pseudonymisation step. To allow the linkage of data and at the same time, to prevent unauthorised linking of data in the Research Zone, different pseudonyms are generated for every research project.

The re-use of EHR data for patient recruitment for clinical studies proceeds differently. Although a researcher in the Research Zone can select patients for research, individual patients can only be identified and invited to participate in a study by their corresponding physician in the Care Zone. For recruitment purposes, the physician needs a patient identifier that will enable him to identify patients that are eligible for the study. This identifier is extracted from the EHR, together with yet another pseudonym and stored in a separate database (containing only the identifier and the pseudonym, not shown in Fig. 7). This pseudonym can be linked to the master database through the TTP allowing the data controller to ask GPs in the Care Zone to invite patients to participate in a study or in additional research for which again informed consent is needed. In summary, the pseudonym in database A is linked to a patient identifier that enables the physician to ask a patient to participate in a study. With respect to aggregate data, such as counts, other rules apply [45].

6. Discussion

Under EU law personal data can only be collected under strict legal conditions and for a legitimate purpose. From the researcher's point of view, it seems that the technical development and sophistication of anonymisation and de-anonymisation techniques is outrunning the legal/policy developments of data privacy protection. It helps to take a step back and review how research with primary care data is done, and analyse the corresponding privacy requirements by using a standardised graphic representational method. The main conclusion of our model is that an overview over the privacy conditions in common research workflows can be a valuable contribution for discussions with researchers, physicians, database owners and software developers. To work with three functional privacy zones is a new approach; until

now different concepts were used to model differences in privacy environments: (1) zone of privacy: an area or aspect of life that is protected from intrusion by specific laws (e.g., the right to be secure in one's house) [46]; (2) privacy domains: domains of trust where all the members of a domain have to comply with a defined privacy policy enforced in this domain [47]; (3) privacy dimensions: three privacy dimensions covering respondent privacy (preventing re-identification), owner privacy (autonomous entities), and user privacy (queries to interactive databases) were defined [48], and (4) collection zone, primary use zone, and secondary use zone [49]. Concepts of privacy zones are also known in architecture (privacy spaces), and in eHealth (presentation zone, application zone, and data zone) [50]. The strategic plan of MITA (Medicaid Information Technology Architecture) recommends boundaries and zones [51] and [52], with each zone protected by firewalls and intrusion detection devices. The security services for healthcare applications project of Purdue University developed a framework for the interoperation of security services [53] consisting of zone 1: patient and doctor, zone 2: covered entities (Medicaid/Medicare CMMS, insurances, local public health agencies, disease registries, FDA), zone 3: business associates (legal services, hospital staff employees, data processing firms), and zone 4: Gramm Leach Billey financial services act (banks, financial institutions, financial service providers). In contrast to zones, pervasive health infrastructures developed the model of digital bubbles, representing personal spaces [54]. Inside a bubble, systems have common privacy regulations and rules.

In contrast to these approaches, but similar to the privacy framework of cancer Biomedical Grid (caBIG), our model defines three categories of data: first, directly identifiable data, second, pseudonymised (one-way and two-way coded data) data and third, anonymous data [55] and creates for the flow of this data a decreasing gradient of the risk of patient identification. The problem with most privacy frameworks is that they rely solely on technical solutions (e.g., PET, Privacy Enhancing Technology) and the use of fully anonymised data for research [26]. But truly anonymised data are often too general a requirement for detailed analysis for certain research questions (e.g., rare diseases).With the proliferation of personal information on the internet anonymous data may become identifiable by linking of datasets with publicly available data. At the same time the volume of data of patients or participants in cohort studies in non-care databases is increasing and data sharing has become a requirement by many funders of research or even an ethical duty [56]. Any privacy model has to consider these aspects and in our model this is done by different zones that allow for research friendly policies. The identifyability of data subjects presents a considerable problem for research. In case of linking a patient clinical data set to genomic data, the impact of indirectly identifiable data must be considered. Indirectly identifiable data are data which have been coded but nevertheless have to be considered as personal data because they are still identifiable. The model, however, does not provide a specific solution for this problem.

EHR data is rich in clinical information, but it is available in formats that can be both structured and unstructured. Although our model focuses on structured medical data and their use for a large variety of pharmacoepidemiologic and population studies, it is easily extensible to other forms of health care investigations, clinical studies and

research (e.g., biobank research). For example, by linking biological material to EHR data, multi-institutional biobanks are advancing in the field of pharmacogenomics. Despite these benefits, the reuse of EHR data for research has been limited by a number of factors, including difficulties in ensuring transparency and data protection as described here, and concerns about the quality of the data and their suitability for research [57].

An alternative approach for the graphic-based representational analysis of privacy frameworks would have been to use conventional UML diagrams or the Business Process Modelling language. UML possesses a security extension, called UMLsec, which contains security principles, expressed as UML stereotypes. Data security analysis applying UMLsec can be used to develop security critical software, where security requirements such as integrity, user authenticity and data security can be specified within UML [58] and [59]. The Business Process Modelling and Notation (BPMN) [60] does not explicitly consider mechanisms to represent data security requirements. However, artefacts able to express security requirements can be designed to extend the notation. BPMN can be extended with a security language, called SecureBPM, which allows for specifying role-based access control. Access policies can be represented as annotations using a security vocabulary, attached to business process model elements [61] and [62]. In contrast to our model, these modelling approaches seem to be only focused on data security policies (e.g., authentication, certification, identity, integrity, and encryption), which are specific for single frameworks and lack the simplicity necessary to model, at a higher level, privacy requirements that are relevant for research processes.

However, our model has some limitations. Firstly, it does not propose a single privacy solution for all research processes. We are not offering a single technical solution but a generic toolkit and associated validated research scenarios that represent different pattern of research. In our model, all pseudonymisation/anonymisation methods are subsumed under the concept of “privacy filter”. Although our model is not concerned with how anonymisation is technically achieved, it can help the analysis of privacy requirements by structuring and displaying privacy frameworks for important research scenarios. Once a privacy compliant research data flow has been achieved, state-of-the-art anonymisation techniques can be applied at the points in the data flow identified by our model. Here any of the constantly emerging new anonymisation methods, like k-anonymity, data obfuscation, synthetic microdata, i-site diversity, may be applied for a privacy filter. Nonetheless, the problem that data can be linked with data in the public domain (e.g., recreational genetic data) still exists and the so-called “failed anonymisation” (i.e. when it becomes possible to “re-identify” or “deanonymise” individuals hidden in anonymised data) can become an issue [63] and [64]. In this context, our model suggests the use of additional filters or strict rules to the data flow in the non-care zone. Because our model's notation is built on UML, our model can be part of the capture of data protection requirements, be used for data security engineering and may be extended by data security specifications, like UMLsec [65], to inform software developers about privacy and security requirements.

Secondly, our model not only concentrates on a very limited subset of symbols from UML, but its notation is simple (e.g., instead of a number of use case swim lanes, only three zones exist). This simplicity of its design may counteract requirements for more sophisticated privacy protection mechanisms: only two data transformation operators (privacy filters, linkers) are provided and this limitation may be not sufficient for all cases of research data flows. For example, operations that partly eliminate information from data or employ data obfuscation are different from coding operations performed to achieve anonymity. Our model focuses more on the state of the data (e.g., personal data, pseudonymised data, etc.) and not so much on how these conditions are created. However, the use of more technically sophisticated symbols may be a next step in the further development of our model.

Thirdly, our model still relies heavily on PET to create anonymised data in the Research Zone; the assumption that data should always be anonymous for research is too restricting, but additional use of PET between data in the Non-care zone and the final research use can be helpful. The further development of our model may regard more specific privacy legislations and policies and may add these policies as amendments to the objects of the model. Another approach using legal interoperability to bring together technology and regulatory frameworks, is the use of privacy ontologies to semantically model privacy obligations [66] and [67].

Although our model does not offer something substantially new to many researchers involved in the privacy protection discussion [4], our model combines and presents these ideas in a novel, structured and organised way, together with a simple graphic notation, thus increasing the comprehension of privacy requirements during discussions. In research, cases exist where a researcher queries for specific data but may have only access to different kind of data and only under special new conditions. Any privacy model must be able to easily adapt to such a change of the data and research workflow.

What does our model offer to vulnerable user populations whose privacy requires special protection? Our approach is rather to apply generic privacy filters and policies, because our model raises awareness of the complexity and context sensitivity of privacy protection, it is easy to add, for example, the protection of children's personal data as additional policy into the workflow. Effective privacy protections must be implemented in a way that does not hinder health research or inhibit medical advances. Reflecting this need, new approaches for the regulation and use of data in health research were propagated from the Academy of Medical Science (AMS) [15] and the enormous variation in policy requirements that hampers health care information exchange has led to realizing the need for common data access policies [68]. Furthermore, researchers must be willing to assess the perceived and actual impact of the data they collect and generate, shaping their attitudes and conduct according to experience [69]. We believe that such “thoughtful research” will build the trust needed for the positive application the enormous amount of patient data gathered in a society that is increasingly unsecure and inconsiderate what data protection is concerned.

7. Conclusion

The model allows analysis of data privacy and confidentiality issues for research in a structured way, using standardised graphic-based notational representations of data sources, data flow and privacy functions within a flexible zone model. It does not suggest a privacy protection framework on the technical level suitable for all research projects, but provides a framework with its components and a privacy compliant data flow. Applying our model, weak points in the definition of a privacy framework can be identified and communicated to developers, thereby providing requirements for privacy framework interoperability. Impairments in legal and ethical regulations are highlighted, for example the still existing unsatisfactory legal safeguarding of the double role of physician as carer and researcher, and the legal definition of identifyability vs. anonymity.

Authors' contribution

Wolfgang Kuchinke was the main author; all other authors contributed to the paper, with Christian Ohmann and Evert-Ben van Veen in graphical modelling, Adel Taweel and Brendan C Delaney in the zone creation and validation, Theodoros N. Arvanitis in the validation and Robert Verheij contributing the NIVEL example.

Conflict of interest

No conflicting interests exist.

Summary points

What was already known before this study:

• Privacy is recognised as an important challenge for health care and health research.

• Many different privacy protection frameworks exist that are project specific and for a single type of research with patient data.

• Several Privacy Enhancing Techniques (PET) have been developed and new anonymisation methods are emerging (e.g., k-anonymity, data obfuscation, synthetic microdata, i-site-diversity).

• Privacy regulations differ between nations and even regions, making a harmonization of privacy protection policies difficult.

What this study has added to our knowledge:

• A generic model is presented able to display, understand and characterise existing policies and rules for data privacy.

• Our model introduces the novel concept of privacy zones with associated operators to transform data (data linker, privacy filter) and with a risk gradient from a zone of high risk to a zone of low-risk for patient identification.

• The model can help to represent data privacy frameworks for complex environments, for international data exchange and for interoperability between different database types.

• It considers the relations of different types of research in each of the three zones and the special roles databases play in providing research services.

ACKNOWLEDGEMENTS

TRANSFoRm is partially funded by the European Commission – DG INFSO (FP7 247787). DG INFSO is now DG Connect.

The research was partly supported by the UK National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St. Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

The initial concept of the privacy zones was developed in a joint TRANSFoRm meeting in Antwerp (Belgium) with contributions by the authors and by Anna Nixon Andreasson, Vasa Curcin, Peter Leysen, Mark McGilchrist, Jean Karl Soler, Hilde Bastiaens, and Paul van Royen. Thanks also to Bram Elffers and ZorgTTP for providing information.

REFERENCES

[1] NHS (UK) observational data and interventional research service: www.cprd.com (accessed 17.07.13).
[2] S. Foresti Preserving Privacy in Data Outsourcing Springer, New York (2011)
[3] O. Tene Privacy – the next generations Int. Data Privacy Law, 1 (1) (2011), pp. 15–27
[4] B.A. Malin, K. El Emam, C.M. O'Keefe Biomedical data privacy: problems, perspectives, and recent advances J. Am. Med. Inform. Assoc., 20 (2013), pp. 2–6
[5] TRANSFoRm project: www.transformproject.eu (accessed 22.08.14).
[6] B.C. Delaney, K.A. Peterson, S. Speedie, A. Taweel, T.N. Arvanitis, F.D.R. Hobbs Envisioning a learning health care system: the electronic primary care research network: a case study Ann. Fam. Med., 10 (1) (2012), pp. 54–59
[7] EU Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Off. J. Eur. Communities, 1999; No. L281/31-281/39.
[8] Health Insurance Portability and Accountability Act of 1996, Public Law 104-191, Report 104-726, 104th Congress (1996).
[9] Department of Health and Human Services, Office of the Secretary: 45 CFR Parts 160 and 164. Standards for Privacy of Individually Identifiable Health Information; Final Rule. Federal Register vol. 67, No. 157 (2002).
[10] The OECD Privacy Framework, OECD Paris, France (2013). Online available http://www.oecd.org.proxy.library.uu.nl/sti/ieconomy/oecd_privacy_framework.pdf (accessed 22.08.14).
[11] APEC Privacy Framework, APEC Secretariat, Singapore (2005), ISBN 981-05-4471-5.
[12] International Standards on the Protection of Personal Data and Privacy, The Madrid Resolution, Spanish Data Protection Agency (2009).
[13] U.S.-EU Safe Harbor Framework Documents, US Federal Register, July 24, 2000. Online available: http://export.gov/safeharbor/eu/eg_main_018493.asp (accessed 23.08.14).
[14] M. Verschuuren, G. Badeyan, J. Carnicero, M. Gissler, R.P. Asciak, L. Sakkeus, M. Sternbeck, W. Devillé Working group on Confidentiality and Data Protection of the Network of competent Authorities of the Health Information and Knowledge strand of the EU Public Health Programme 2003-08 Eur. J. Public Health, 18 (2008), pp. 550–551

[15] Academy of Medical Sciences (AMS). A new pathway for the regulation and Non-care of health research. January 2011, available at: www.acmedsci.ac.uk/p47prid88.html (accessed 23.08.14).
[16] E.B. van Veen Obstacles to European research projects with data and tissue: solutions and further challenges Eur. J. Cancer, 44 (2008), pp. 1438–1450
[17] Office for Civil Rights, Department of Health and Human Services. HIPAA Privacy Rule. Title 45 of the Code of Federal Regulations Parts 160 and 164. Washington, D.C, Feb 2009, available at: www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf (accessed 23.08.13).
[18] D. McGraw Paving the regulatory road to the “learning health care system” Sanford Law Rev. Online, 64 (2012), p. 75 (February 8, 2012), available at: www.stanfordlawreview.org/online/privacy-paradox/learning-health-care-system (accessed 23.08.13).
[19] S.J. Nass, L.A. Levit, L.O. Gostin (Eds.), Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research, Institute of Medicine, The National Academies Press, Washington, DC (2009)
[20] Office of the National Coordinator for Health information Technology, US Department of Health and Human Services: Nationwide Privacy and Security Framework for Electronic Exchange of Individually Identifiable Health Information, December 15, 2008, available at: www.healthit.gov/policy-researchers-implementers/standards-interoperability-si-framework (accessed 23.08.14).
[21] C. Runnegar International privacy frameworks: an overview 2011 International Cloud Symposium (OASIS); 2011 October 10–12, The Internet Society, Heathrow, UK (2011)
[22] C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, D.E. Detmer Toward a national framework for the secondary use of health data J. Am. Med. Inform. Assoc., 14 (2007), pp. 1–9
[23] M. Bloomrosen, D. Detmer Advancing the framework: Use of health data – a report of a working conference of the American Medical Informatics Association JAMIA, 15 (2008), pp. 715–722
[24] D. Kalra, P. Singleton, D.J. Milan, D. Detmer, A. Rector, D. Ingram Security and confidentiality approach for the Clinical E-Science Framework (CLEF) Methods Inf. Med., 44 (2005), pp. 193–197
[25] N. Forgó (Ed.), The ACGT ethical and legal requirements. ACGT deliverable 10.2. 13.03.2007, available at: http://acgt.ercim.eu/uploads/media/ACGT_D10.2_IRI_Final_01.pdf (accessed 22.08.14).
[26] E.-B. van Veen, Patient data for health research. October 2011. Available at: www.medlaw.nl/wp-content/uploads/patient-data-for-health-research.pdf (accessed 23.08.14).
[27] R.A. Verheij, C.E. Van Dijk, I. Stirbu-Wagner, S.A. Dorsman, S. Visscher, H. Abrahamse, R. Davids, J. Braspenning, T. Van Althuis, J.C. Korevaar. Landelijk Informatienetwerk Huisartsenzorg. Feiten en cijfers over huisartsenzorg in Nederland. LINH: Netherlands Information Network of General Practice. Utrecht/Nijmegen: NIVEL/IQ, 2009.
[28] General Practice Research Database (GPRD), Has been renamed as: Clinical Practice Research Datalink (CPRD), available at: www.cprd.com/intro.asp (accessed 23.08.14).
[29] T. Kuehne What is a model? Language engineering for model-driven software development 2005; 04101 Dagstuhl Seminar Proceedings, Dagstuhl, Germany (2005, March)
[30] Unified Modelling Language (UML), Object Management Group, available at: www.uml.org (accessed 23.08.14).
[31] R.M. Friedenberg Patient–doctor relationships Radiology, 226 (2003, February), pp. 306–308
[32] Institute of Electrical and Electronics Engineers: IEEE standard 1471. IEEE Recommended Practice for Architectural Description of Software-Intensive Systems (2000). Replaced by 42010-2007.
[33] C. Borgelt, J. Gebhardt, R. Kruse Graphical Models (2002)

Proceedings of International School for the Synthesis of Expert Knowledge (ISSEK'98), Wiley, Hoboken, NJ (2002), pp. 51–68
[34] OASIS Privacy Management Reference Model, available at: https://www.oasis-open.org (accessed 23.08.14).
[35] Nationwide Privacy and Security Framework for Electronic Exchange of Individually Identifiable Health Information (HHS, US), 15 December 2008, available at: www.healthit.gov/sites/default/files/nationwide-ps-framework-5.pdf (accessed 23.08.14).
[36] Health privacy framework, National Health and Medical Research Council Australia, available at: www.nhmrc.gov.au/health-ethics/human-research-ethics-committees-hrecs/health-research-privacy-framework (accessed 23.08.14).
[37] Clinical E-Science Framework (CLEF), University of Sheffield, UK, available at: http://nlp.shef.ac.uk/clef/ (accessed 23.08.14).
[38] ACGT (Legal and Ethical issues), available at: http://acgt.ercim.eu/documents/legal-and-ethical-issues.html (accessed 23.08.13).
[39] GenoMatch, available at: www.tembit.de (accessed 23.08.13).
[40] V.B. Kaujalgi Structured Systems Analysis and Design: Data Flow Approach Orient Blackswan, Himayatnagar, India (1994)
[41] School for Primary Care Research: Will the National Primary Care Database Bring Big Benefits? University of Oxford, Tweet (2014, February 24)
[42] E.B. van Veen Europe and tissue research: a regulatory patchwork Diagn. Histopathol., 19 (9) (2013), pp. 331–336
[43] Nederlands instituut voor onderzoek van de gezondheidszorg, available at: www.nivel.nl/taxonomy/term/all?gegevensverzameling[]=45 (accessed 23.08.13).
[44] Code of Conduct health research, Commissie Regelgeving en Onderzoek, available at: www.federa.org/sites/default/files/bijlagen/coreon/code_of_conduct_for_medical_research_1.pdf (accessed 23.08.14).
[45] Central Bureau of Statistics Act, Staatsblad 2004, 695, available at: http://www.cbs.nl/NR/rdonlyres/F10515CB-91C6-426C-9BD9-177F743F72C6/0/cbswet15122004.pdf (accessed 23.08.14).
[46] Legal Dictionary (Lawyers.com), available at: http://research.lawyers.com/glossary/zone-of-privacy.html (accessed 23.08.14).
[47] H. Loehr, A.R. Sadeghi, C. Vishik, *et al.* Trusted privacy domains – challenges for trusted computing in privacy-protecting information sharing. Information Security Practice and Experience 5th International Conference, ISPEC 2009: Proceedings, vol. 5451, Xi'an, China, April 13–15 (2009)
[48] J. Domingo-Ferrer, Y. Saygın Recent progress in database privacy Data Knowl. Eng., 68 (11) (2009), pp. 1157–1159
[49] B.A. Malin, K.E. Emam, C.M. O'Keefe Biomedical data privacy: problems, perspectives, and recent advances (Editorial) J. Am. Med. Inform. Assoc., 20 (2013), pp. 2–6
[50] L. Ohno-Machado, V. Bafna, A.A. Boxwala, B.E. Chapman, *et al.* iDASH: integrating data for analysis, anonymization, and sharing J. Am. Med. Inform. Assoc., 19 (2012), pp. 196–201
[51] CMS Centres for Medicare Medicaid Services: Harmonized Security and Privacy Framework – Exchange Reference Architecture Supplement. Version 1.0, August 1, 2012.
[52] MITA Application Architecture, CMS Centres for Medicare Medicaid Services. May 8, 2006, Medicaid Info Technical Archive.
[53] IPS: Security services for healthcare applications, Purdue University, Computer & Information Technology, 2007, available at: www.cs.purdue.edu/homes/bertino/IIS-eHealth/ehealth.shtml (accessed 23.08.14).
[54] P.S. Ruotsalainen, B.G. Blobel, A.V. Seppälä, H.O. Sorvari, P.A. Nykänen A conceptual framework and principles for trusted pervasive health J. Med. Internet Res., 14 (2) (2012), p. e52
[55] Cancer Biomedical Informatics Grid (caBIG): Data Sharing and Security Framework, available at: https://wiki-nci-nih-gov.proxy.library.uu.nl/display/DSIC/Data+Sharing+and+Security+Framework (accessed 23.08.14).

[56] B.M. Knoppers, J.R. Harris, I. Budin-Ljøsne, E.S. Dove A human rights approach to an international code of conduct for genomic and clinical data sharing Hum. Genet., 133 (2014), pp. 895–903
[57] N. Gray Weiskopf, C. Wenig Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research J. Am. Med. Inform. Assoc., 20 (2013), pp. 144–151
[58] Shareeful Islam, Jan Jürjens. Incorporating security requirements from legal regulations into UMLsec model. Available at: www.secse.cs.tu-dortmund.de/jj/publications/papers/modsec08IJ.pdf (accessed 23.08.14).
[59] S.H. Houmb, S. Islam, E. Knauss, J. Jurjens, K. Schneider Eliciting security requirements and tracing them to design: an integration of Common Criteria, heuristics, and UMLsec Requirements Eng., 15 (2010), pp. 63–93
[60] Business Process Modeling Notation Specification. OMG Final Adopted Specification. Object Management Group (February 2006). Online available: http://www.omg.org/bpmn/Documents/OMG_Final_Adopted_BPMN_1-0_Spec_06-02-01.pdf (accessed: 23.08.14).
[61] A. Rodrigez, E. Fernandez-Medina, M. Piattini A BPMN extension for the modeling of security requirements in business processes IEICE Trans. Inf. Syst. (2007) E90-D(4)
[62] A.D. Brucker, I. Hang, G. Lückemeyer, R. Ruparel SecureBPMN: modeling and enforcing access control requirements in business processes Proceedings of the 17th ACM Symposium on Access Control Models and Technologies (SACMAT '12), ACM, New York, USA (2012), pp. 123–126
[63] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, Y. Erlich Identifying personal genomes by surname inference Science, 339 (6117) (2013), pp. 321–324
[64] P. Ohm Broken promises of privacy: Responding to the surprising failure of anonymization UCLA Law Rev., 57 (2010), pp. 1701–1711
[65] J. Jürjens Secure Systems Development with UML Springer-Verlag, Heidelberg (2010)
[66] H.B. Rahmouni, T. Solomonides, M.C. Mont, S. Shiu Privacy compliance and enforcement on European healthgrids: an approach through ontology Philos. Trans. A Math. Phys. Eng. Sci., 368 (1926) (2010), pp. 4057–4072
[67] H.B. Rahmouni, T. Solomonides, M. Casassa Mont, S. Shiu Modelling and enforcing privacy for medical data disclosure across Europe Stud. Health Technol. Inform., 150 (2009), pp. 695–699
[68] C. Wolter, M. Menzel, A. Schaad, P. Miseldine, C. Meinel Model-driven business process security requirement specification J. Syst. Architect., 55 (2009), pp. 211–223
[69] L. Dimitropoulos, S. Rizk A state-based approach to privacy and security for interoperable health information exchange Health Aff. (Millwood), 28 (2) (2009), pp. 428–434

TABLES EN FIGURES

Table 1 – Scheme of data types.

Type of data (according to EU Directive 95/46/EC)	Identifyability	
Anonymous data	Fully anonymous data	
	Coded anonymous (pseudonymised) data	
Personal data	Indirectly identifiable data	Coded but either coding insufficiently secure or aggregation level too low
		Not coded but aggregation level too low
	Directly identifiable data	

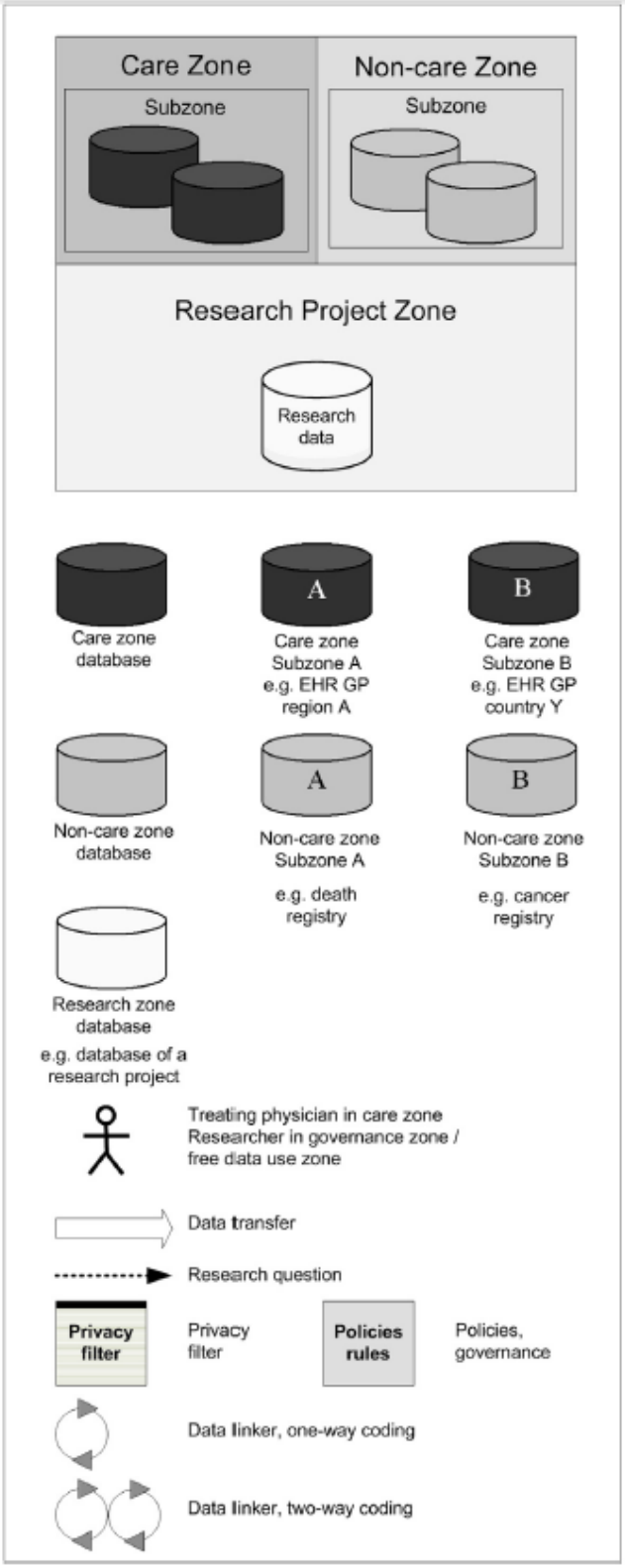

Fig. 1 – The building blocks of the zone model. Notation for zones and subzones, databases in the different zones/subzones and symbols for data transfer, privacy functions and actors.

Table 2 – Zones and corresponding data types.

Zone	Data types in the model
Care	• identifiable clinical data • patient/GP relationship • patient data contained in the EHR • explicit authorisation, explicit consent, or remote searches and flagging where no data is transferred
Non-care	• non-care databases (e.g., primary care databases) with pseudonymised data • tools to enable anonymisation, pseudonymisation and linkage between pseudonymous data sources (privacy filters, linkers) • anonymisation of data before transferred to research zone
Research	• anonymised data (general authorisation); pseudonymous data (linked) • explicit authorisation, but consent model varies according to risk of disclosure • genetic data deserve special attention as these can be potentially identifying

Table 3 – Process description of TRANSFoRm use cases.

Use Case	Process steps
Use case: Find patients for clinical research	
Identification of trial patients by the treating physician within the treatment context	• eligibility search is triggered by the treating physician or triggered within EHR (e.g., data trigger) • search within EHR identifies potential trial patient • patient identity transferred to treating physician • treating physician invites patient for trial participation • patient gives informed consent (or not)
Identification of trial patients in non-care repositories (e.g., research database, register)	• eligibility search is triggered by the researcher • data controller of non-care database allows search (consistent with policy to use research database) • search within non-care database with pseudonymous data identifies potential patient • data controller of non-care database transfers pseudonym of identified patient to treating physician • treating physician is able to identify potential trial participant • treating physician invites patient for trial participation • patient gives informed consent (or not)
Use case: Linkage of databases	
Linking of databases in the Non-care Zone	• trigger of linkage by researcher • checking the permissibility of linkage of the non-care databases by the data controllers of these databases • authorisation of data linkage (e.g., ethical or data protection committee) • performance of linkage (e.g. use new pseudonym) • linked database is re-coded (one-way coding for anonymous data, two-way coding for pseudonymous data) • linked database transferred into Research Zone after privacy filtering • linked database analysed according to research question by researcher
Link between databases in the Care Zone (e.g., EHR database, hospital data warehouse) with database in the Non-care Zone (e.g., research database, register)	• trigger of linkage by researcher • checking the permissibility of linkage of the care database with the non-care database by the data controllers of these databases • authorisation of data linkage (e.g., data protection committee) • preparation of linkage procedure in care and non-care database • performance of linkage in Non-care Zone (e.g., using new pseudonym) • linked database coded two times (the data are re-coded using a different pseudonymisation method, that is a different key) • linked database transferred into Research Zone after privacy filtering • linked database analysed according to research question by researcher

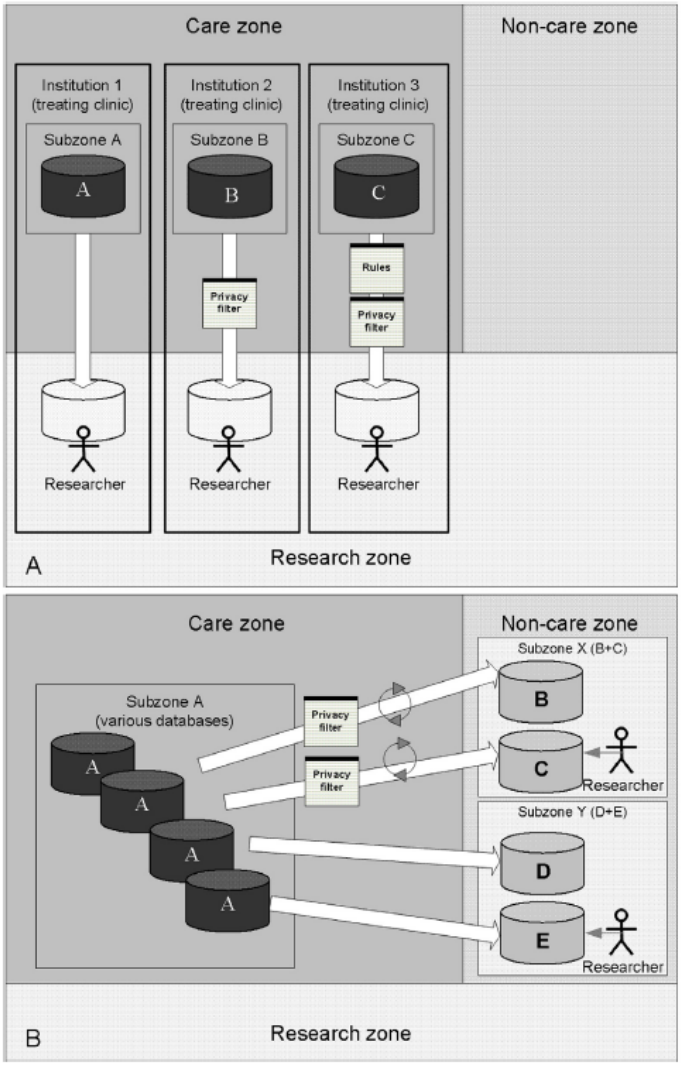

Fig. 2 – Models for the illustration of two scenarios for the research use of medical care data. (A) Use of clinical care data for research within the institution where the data are collected. Subzone A: research use without informed consent is possible (research exemption), Subzone B: use without informed consent only after anonymisation, Subzone C: use without informed consent after anonymisation and application of rules. (B) Collection of data for different data bases in the Non-care Zone. (B) National cancer registry, indirectly identifiable; (C) NIVEL-PCD database, coded-indirectly identifiable. (D) Data about medical consumption at the health care insurer. (E) Death registry at Statistic Netherlands.

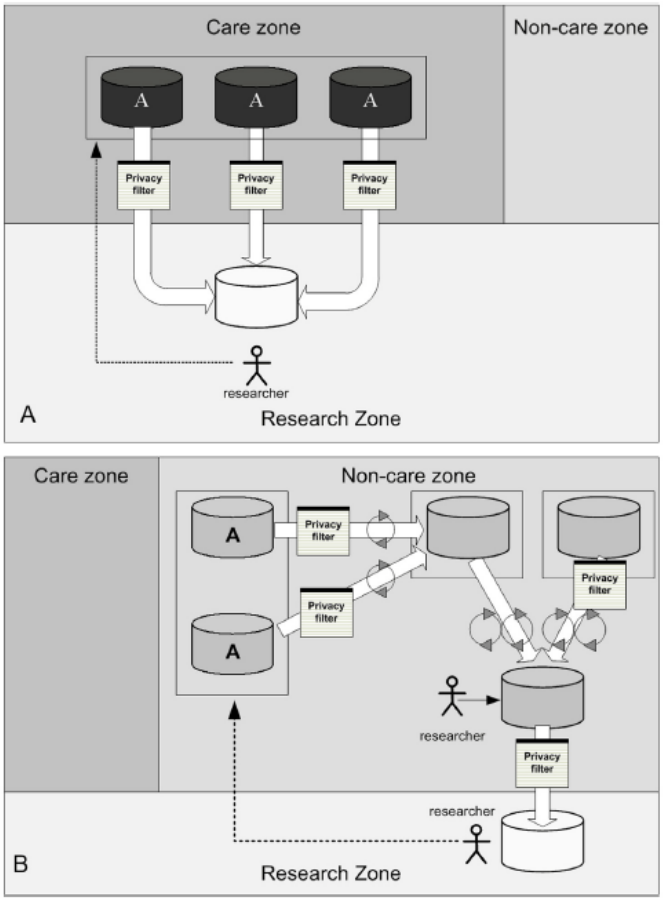

Fig. 3 – Representation of research scenario 1 and 2. (A) Data from EHRs of different GPs belonging to the same Subzone A within the Care Zone (e.g., GPs that are located in the same state/country) are merged in a research database after passing a privacy filter. The data in the Research Zone is anonymous. (B) Two databases in the same subzone of the Non-care Zone (such as cancer registry, civic registry) are linked with other databases after passing a privacy filter. Different linkers are used and the resulting database is linked by two-way coding with another secondary database, which itself has passed a privacy filter (the broken arrow means raising the research question).

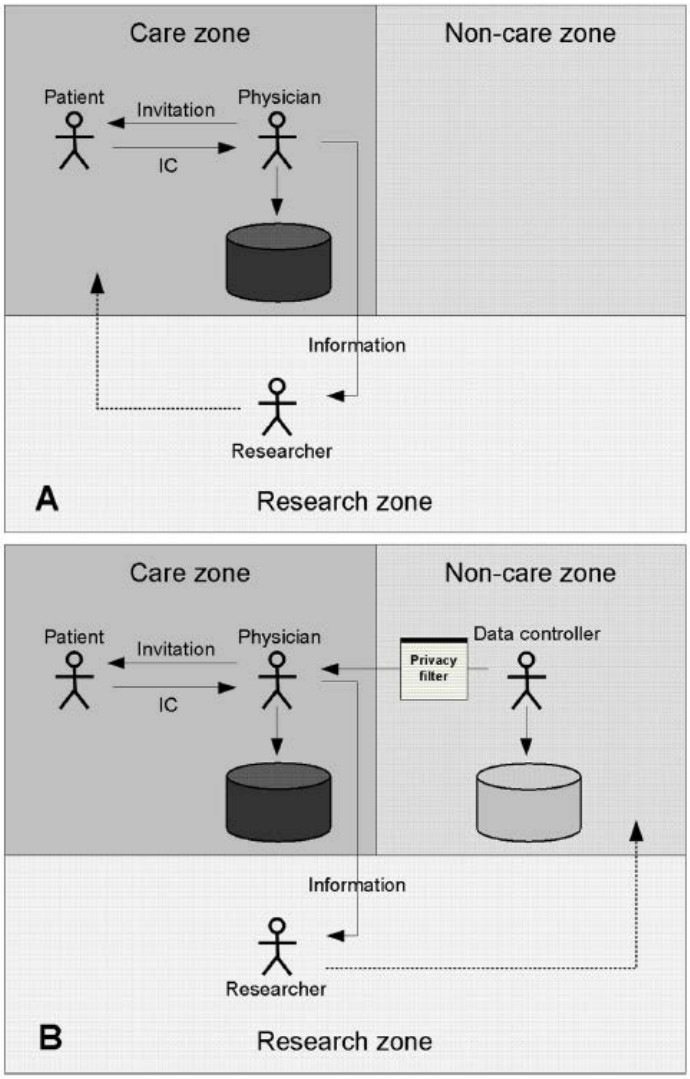

Fig. 4 – Recruitment of patients for clinical studies. (A) Representation of use case "Find patients for clinical research by treating physician"; (B) representation of use case "Find patients for clinical research in Non-care database"; (IC = informed consent).

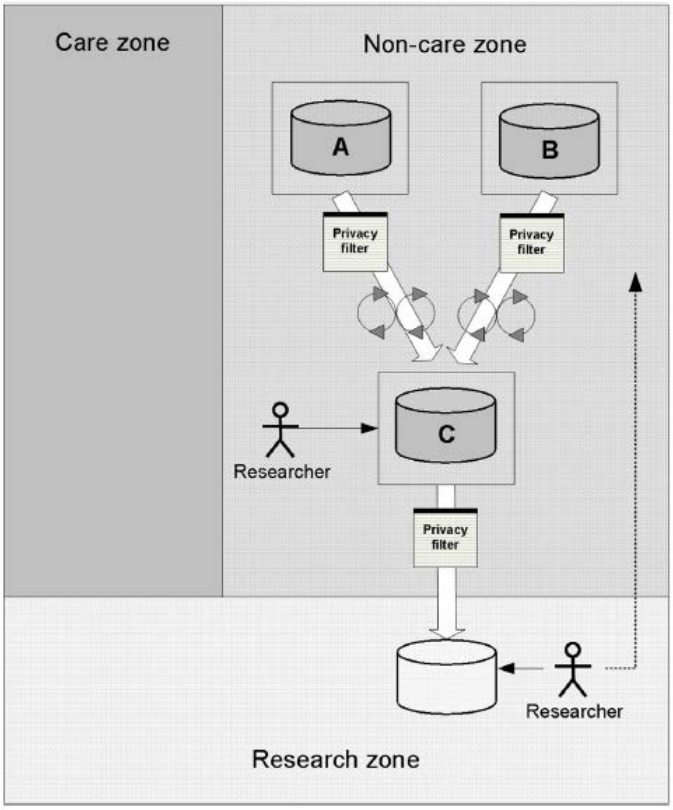

Fig. 5 – Representation of use case "Linkage of databases in the Non-care Zone". The researcher receives pseudonymised, (coded) anonymous data (if re-identification is necessary) or fully anonymised data according to the research question and authorisation obtained.

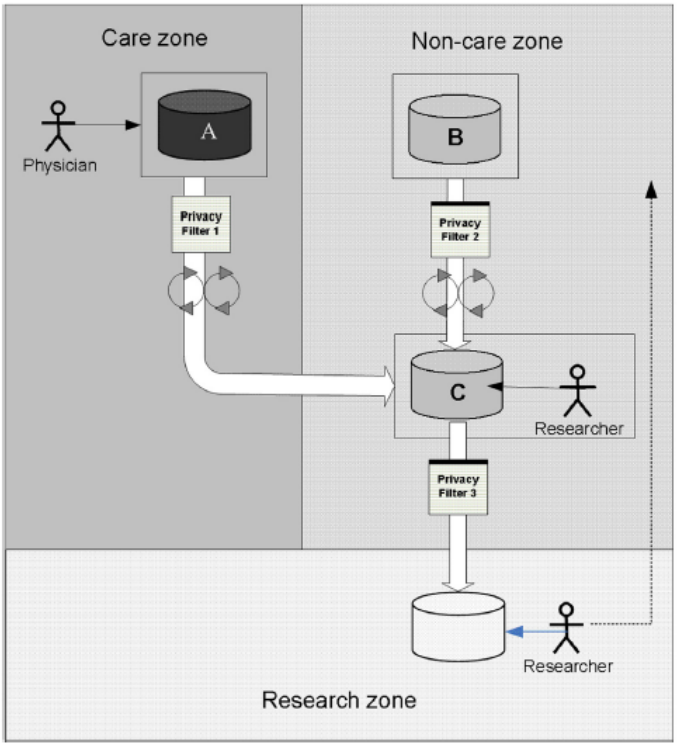

Fig. 6 – Representation of use case "Linkage of database in the Care Zone with database in the Non-care Zone". The researcher is provided with different types of data according to research question and authorizations: pseudonymised or (coded) anonymous data in case a re-identification is necessary or fully anonymised data for analysis. Therefore, depending on the privacy filter applied, linked data that reach the Research Zone is fully anonymous or is pseudonymous in such a way that data appears anonymous to the researcher. The researcher analyses the linked datasets in their database.

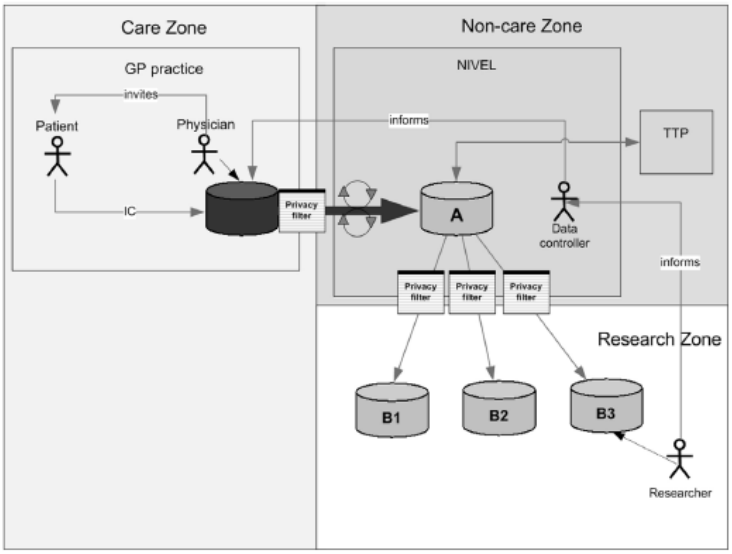

Fig. 7 – Representation of the example depicting data exchange and research done with NIVEL-PCD. Regularly patient data is transferred from the GP to the NIVEL database, where data is stored with pseudonyms (database A). Researchers can obtain additionally pseudonymised and aggregated extracted data sets (database B1-B3); only a subset of patient data that is doubly pseudonymised (practically anonymised) can be analysed by the researcher; (IC = informed consent, TTP = Trusted Third Party).