TRANSFoRm

Translational Research and Patient Safety in Europe

Report on regulatory requirements, confidentiality and data privacy issues

Part B: Confidentiality and data privacy framework (Deliverable 3.2)

Heinrich-Heine University Düsseldorf (UDUS), Coordination Centre for Clinical Trials on behalf of the European Clinical Research Infrastructures Network (ECRIN)

Netherlands Institute for Health Services Research (NIVEL) and MedLawconsult

Trinity College Dublin (TCD), The Irish Software Engineering Research Centre

Work Package:	WP3, Deliverable 3.2
Type of document:	Conceptual framework
Version:	Version 1
Date:	31 March 2011
Authors:	C. Ohmann, W. Kuchinke (UDUS), E.-B. van Veen (MedLawconsult on behalf of NIVEL), R. Verheij (NIVEL), S. Farrell (TCD)

TRANSFoRm is partially funded by the European Commission - DG INFSO Under the 7th Framework Programme. (FP7 247787)
http://cordis.europa.eu/fp7/ict/
http://ec.europa.eu/information_society/index_en.htm

European Commission Information Society and Media

eHealth Better Healthcare for Europe

Index

Executive Summary

The objective of WT 3,2 “Confidentiality and privacy” is to develop an extensible privacy and confidentiality framework that supports the different stages of the clinical trials process for finding and recruiting eligible patients while maintaining their privacy. The framework must distinguish between different types of data including anonymised and identifiable clinical data providing approved mechanisms of data access at different levels and at different stages. The overall framework needs to support and preserve different local data sharing policies in different health organization while enabling them to benefit from TRANSFoRm services. The work will take into account different data protection regulations in EU member states. And will inform the choice of networks for the demonstration and validation of TRANSFoRm.

The development on a confidentiality and data privacy framework for TRANSFoRm is not an easy task. The implementation of Directive 95/46/EC has been very divergent in the European member states with regard to the use of data for health. Some frameworks exist already in Europe from other projects. They are, however, dedicated mostly to specific countries (e.g. US), specific diseases (e.g. cancer), specific situations (e.g. only research context with informed consent by the patient) or to specific data sources (e.g. only use of secondary data). Given the divergences between the member states and the problems of consent for the use of data for research, those projects often aim at only using fully anonymous data or explicit consent in all cases. That would make the framework rather easy, but not very nuanced. Sometimes more refined data will be needed for research or – as in some of the TRANSFoRm use cases – it will even be necessary to contact the patient.

The TRANSFoRm approach intends to find a balance between individual privacy interests on the one hand and research with health care data for the public good on the other hand. At the same time TRANSFoRm is very much about using privacy enhancing technologies (PET) to resolve that tension as much as possible. Technology is not considered as neutral, instead It should be seen at the background of ethical and regulatory principles. As a first step we formulated a set of principles which are our normative starting points, based on an assessment of the literature. The next aspect was to create a framework for the research within TRANSFoRm. TRANSFoRm deals with heterogenous data from different data sources (e.g. primary, secondary), related to different context (medical care, research) and associated with different degrees of risk of identification (personal identifiable, pseudonymous, anonymous). In order to arrive at this framework there has been considerable discussion within the TRANSFoRm group about definitions, data sources and the dataflow, leading to the conception of a formal approach for the description of the

framework. So based on the definitions and the principles we made a model to describe the dataflow for research projects. We exemplified this model with schemes of dataflow in projects elsewhere and in the TRANSFoRm use cases.

As a first step clarification of the terminology was needed and a chapter with definitions was created for the most important terms, mainly based on the EU Data Protection Directive, additional documents and the relevant literature (chapter 2). We described the methods applied to develop the framework in chapter 3. In chapter 4, the description of the framework is presented, divided into two parts: the mentioned principles and formal description. In chapter 5 the confidentiality and data privacy framework is applied to TRANSFoRm use cases dealing with research in diabetes and GastroOesophageal Reflux Disease (GORD). Concluding remarks follow (chapter 6), together with references (chapter 7) and a list of abbreviations (chapter 8).

The principles were extracted in a non-formal way from laws and regulations and relevant literature. They are presented as a series of 14 statements, covering: general policy, primary responsibility for treating physician, data chain and following data flow, the use of privacy enhancing technologies, explicit consent, trust, use by third parties, pseudonymisation, data controller, contractual agreements and database statutes, national legislation, publication, data sharing and genetic data.

Based on the principles, a structured representation of the framework was developed based on elements of structured analysis and data flow diagrams. The full range of data sources is divided by zones with the idea to represent context (medical, research). Zones are defined as areas for data sources that are comparable and similar with respect to purpose, rules and regulations for use. The idea is to have personal identifiable data, pseudonymised data and anonymous data in different zones to structure the data sources zones according to risk for confidentiality and data privacy issues, representing areas with low, medium and high risk of identification. In the framework three major zones are distinguished: care zone (e.g. EHR), non care zone (e.g. registers, research databases, cohort studies) and research zone. Within zones subzones are defined, which contain data that are comparable and can be used for the same or a very similar purpose and with similar applicable rules and regulations for their use (e.g. GPs EHRs within one country). Data transfer between zones/subzones is described by data flow diagrams, illustrating the data flow from one data source in one zone to another data source in another zone. To make this data flow possible, specific functions (processes) are applied based on Privacy Enhancing Techniques (PET) in order to ensure optimal privacy protection for the patient (e.g. anonymisation, pseudonymisation, coding and data aggregation). In order to allow combination of data from different data sources data linkers are introduced in the framework. Linking may be performed by one-way coding or by two-way coding. Finally actors involved in the framework together with their roles are defined

(e.g. actor clinician with the role treating physician or researcher). A standardised notation is used for the graphical representation of the framework.

The formal description of the framework is applied to five typical research scenarios: merging data from EHRs, linkage and merging of data from different subzones in the care-zone, data transfer from the non-care database to the research zone, linkage and enrichment of data bases in the non-care zone and linkage of a cohort study database with data from the care and non-care zone.

To develop and validate the software and services developed by TRANSFoRm, two representative research use cases (a genomic-phenotype study and an RCT) are being developed within the project. The use cases have been described in WP1-WT1.1 "Development of use cases" (Version 2, 13.9.2010). Research use case 1 deals with diabetes and use case 2 with the gastrooesophageal reflux disease (GORD). In order to support implementation of TRANSFoRm use cases the formal description of the framework is applied to the following use cases (processes): counts of patients with a defined pattern, find patients for clinical research, select patient for research question, extract information of selected patients and linkage of data. For each use case the context, zones, actors and processes involved, data privacy issues and potential problems are displayed together with a formal representation.

Being aware of the divergent implementations and applications of laws and regulations for data protection and confidentiality in Europe, the primary objective within this work package of TRANSFoRm was to analyse and structure the given heterogeneity to come to an approach working out similarities between different scenarios (e.g. between countries, data sources, context), and thus providing practical help in a highly complex environment. The idea behind was to build upon existing means of confidentiality and data privacy measures (e.g. PET, TTP, data transfer agreements) and to incorporate it into the model. The next steps will be to apply the framework to the TRANSFoRm use cases. This will also be a test of the principles and the formal representation of the framework, including the chosen methodological approach. Whether the framework has to be amended or expanded has to be discussed when first experiences with concrete applications are available. Therefore, the resulting document (version 1) should be seen as the first guidance document for all resulting TRANSFoRm activities. It should be mentioned that this is a living instrument. The document will be refined and enlarged with additional principles in the next phases of TRANSFoRm.

1. Introduction

1.1 The question

TRANSFoRm (Translational Research and Patient Safety in Europe) requires in WP 3 "Legal, ethical and security framework" (WT 3.2.2) to develop an extensible privacy and confidentiality framework that supports the different stages of the clinical trials process for finding and recruiting eligible patients while maintaining their privacy. The framework must distinguish between different types of data including anonymised and identifiable clinical data providing approved mechanisms of data access at different levels and at different stages. The overall framework needs to support and preserve different local data sharing policies in different health organization while enabling them to benefit from TRANSFoRm services. The work will take into account different data protection regulations in EU member states. And will inform the choice of networks for the demonstration and validation of TRANSFoRm (see project proposal: FP7-ICT-2009-4, Project no. 247787).

1.2 Preliminary inquiry

Stated as such the question is not an easy task. The implementation of Directive 95/46/EC has been very divergent in the European member states with regard to the use of data for health research (van Veen 2006, Verschuuren 2008, EU Directive 95/46/EC). Some frameworks exist already in Europe from other projects but given the divergences between the member states and the problems of consent for the use of data for research (see hereinafter), those projects aim at only using fully anonymous data or explicit consent in all cases. That would make the framework rather easy, but not very nuanced. Sometimes more refined data will be needed for research or – as in some of the TRANSFoRm use cases – it will even be necessary to contact the patient.

There is an inherent tension between the privacy of the patients and the necessity to use data assembled during the care for this patient (or from other sources, like environmental exposure, workplace hazards) for the public good (Kaira 2006; Verschuuren 2008; van Veen 2008 Academy of Medical Sciences (AMS), 2011). It is a highly contested area. In the Recommendations following from a study group on the implementation of Directive 1995/46/EC and its relation to health research the existing exemptions on the consent

principle were seen as loopholes and an infringement of privacy.[1] On the other hand a Working Group from the Network of Competent Authorities of Health Information and Knowlegde Strand concluded that next to the major differences between national data protection systems regarding possibilities for using identifiable health data for health research, many countries had a too restrictive view of using these data for public health research (Verschuuren, 2008). They pointed at the fact that explicit consent for health research often can lead to a bias in the data and will hamper the use of those data for the public good. To this can be added that if many patients are involved, the consent process can also be very costly. Also van Veen, 2008 and AMS, 2011 pointed at these problems pleaded for leniency in using personal data for health research, the latter restricted to the situation in the UK. From the TRANSFoRm work in WP 3.2.1 it follows that patient organisations are in favour of using personal medical data for health research as well, if certain conditions are met.

Verschuuren, 2008 encouraged the European Commission to improve the legal framework, in the sense of harmonising the present exemption to use personal data for health research. However, other authors recommend not to aim for a new harmonised set of rules on the European level but to formulate only principles to allow for adequate flexibility (van Veen, 2008). Currently, the EU Data Protection Directive is under revision, however with uncertain outcome concerning consequences for performing medical research.

International organizations have issued recommendations like the World Health Organization (WHO, 20011), the Universal Declaration on Bioethics and Human Rights (UNESCO, 2005), Council for International Organizations of Medical Sciences (CIOMS, 2009) and the Organisation for Economic Co-operation and Development (OECD, 2009). The Declararation of Helsinki (World Medical Association) has brought research with identifiable tissue and data within its remit as well (Declaration of Helsinki, 2008). Despite the value of such guidance, these also have disadvantages. They are usually very generally formulated and often the principles forwarded in them can be seen as rather one sided. This issuing of non legally binding recommendations by international organisations has even been described once as a potential (rhetorical) 'race to the top' (Follesdal 2008, see also Van Veen, 2006).

1.3 The TRANSFoRm approach

A balance must be found between individual privacy interests on the one hand and research with health care data for the public good on the other hand. We tend to the view of

[1] Privireal, which was ended in 2005. See www.privireal.org. Two books followed from the study namely Beyleveld 2004 and Beyleveld 2005.

Verschuuren 2008, Van Veen 2008, AMS 2011, and the consulted patient organisations that sometimes that balance is lost in some of present regulations or approaches by Data Protection Authorities.

At the same time TRANSFoRm is very about using privacy enhancing technologies (PET) to resolve that tension as much as possible. However, no technology is neutral. It should be seen at the background of ethical and regulatory principles.

Hence we formulated a set of principles which are our normative starting points, based on aforementioned assessment of the literature. As we first of all must have a common understanding of the terms used in those principles and later in the text, an extensive list of definitions came first.

The next aspect was to indeed to create a 'Framework' for the research of the other partners in TRANSFoRm. TRANSFoRm deals with heterogenous data from different data sources (e.g. primary, secondary), related to different context (medical care, research) and associated with different degrees of risk of identification (personal identifiable, pseudonymous, anonymous). In order to arrive at this framework there has been considerable discussion within the TRANSFoRm group about definitions, data sources and the dataflow, leading to the conception of a formal approach for the description of the framework. So based on the definitions and the principles we made a model to describe the dataflow for research projects. We exemplified this model with schemes of dataflow in projects elsewhere and in the TRANSFoRm use cases. Our idea is that this model can be applied elsewhere as well.

1.4 Set up and character of this framework

As a first step clarification of the terminology was needed and a chapter with definitions was created for the most important terms, mainly based on the EU Data Protection Directive, additional documents and the relevant literature (chapter 2). We described the methods applied to develop the framework in chapter 3. In chapter 4, the description of the framework is presented, divided into two parts: the mentioned principles and formal description. The formal description is based upon structural elements (zones, privacy filters and data linkers) combined in various ways to represent different research scenarios. In chapter 5 the confidentiality and data privacy framework is applied to TRANSFoRm use cases dealing with research in diabetes and GastroOesophageal Reflux Disease (GORD). Concluding remarks follow (chapter 6), together with references (chapter 7) and a list of abbreviations (chapter 8).

The resulting document (version 1) should be seen as the first guidance document for all resulting TRANSFoRm activities. It should be mentioned that this is a living instrument. The

document will be refined and enlarged with additional principles in the next phases of TRANSFoRm.

2. Definitions

Primary use

Use for the purpose, explicit or implied, for which the data were originally collected. In health care the care provider collects data primarily for diagnostic and treatment purposes. The care provider will use these data also for quality assurance and improvement of the practice. Usually this is still considered primary use as in most systems the care provider has a duty to give optimal care and continuously work on quality improvement. To use those data for medical research would generally be considered as secondary use.

Secondary use

Secondary use is the use of data for another purpose than for which they were originally collected. If data have been collected for one specific research question, like the relation between nutrition and cancer, and then would be used for another research question, like the relation between nutrition and osteoporosis, the latter use of aforementioned data would be secondary use. Or, in the example of the doctor having medical data of his patients, if he or she would use those data for medical research.

It should be stressed at this point that 'primary' and 'secondary use' are used here as definitions to clarify the discussion without – as yet – a normative content. Is some cases explicit consent is necessary to collect data for primary use, like – in most countries - if the doctor would ask for data from a hospital were the patient has been treated previously. To use these data by the treating doctor, data which the doctor has already, for research related to conditions for which the patient consulted him (or her) is secondary use but – again in most countries – no consent would be needed for such secondary use as no privacy issue is involved.

However, the fact that the use is secondary should raise special concern about the legitimacy of this use. If the personal data are used by a 'third party' then special consent mechanisms are required or use of research exemptions according to national law.

In the TRANSFoRm zone model (which will be described later in this document), the research zone is always secondary use. Thus, data are not primarily generated in the research zone but are provided from databases (e.g. EHR, clinical trial database, register) for specific research questions. However, not all secondary use takes place in the research zone as defined by TRANSFoRm.

Third party

According to the EU Directive third party shall mean any natural or legal person, public authority, agency or any other body other than the data subject, the controller, the processor and the persons who, under the direct authority of the controller or the processor, are authorized to process the data. Thus, the third party has not collected the data for its original use and the data are transferred to the third party in the data chain. The third party will be the receiver of the data.

Informed consent

a) to treatment

The decision, based on sufficient information, to undergo a diagnostic test or to undergo treatment in medical care. This consent should be based on sufficient information and specified to the circumstances of the case of the possible harms and the possible advantages. The information should be based on what the 'average' or the 'reasonable' patient (health law is somewhat divided on this point) needs to now to make an informed choice about the proposed procedure under the given circumstances.

b) to participate in a medical trial

A process by which a subject voluntarily confirms his or her willingness to participate in a particular trial, after having been informed of all aspects of the trial that are relevant to the subject's decision to participate. Informed consent is documented by means of a written, signed and dated informed consent form. National regulations may provide specific rules for persons not capable of giving consent (e.g. in case of emergency situations) and persons unable to write.

It follows from the above that the threshold for informed consent to participate in a trial is (much) higher than for a medical procedure in general.

Explicit consent (for the use of data)

This is the decision of the data subject (usually in this context a patient) that his or her data may be used for a specified purpose. This consent should be based on sufficient general information about the purpose of the data processing and how the data will be handled.

Though explicit consent is the standard, it should be stressed that such a decision is not always necessary in order to use data. For example, if a doctor collects data from a patient in the course of the treatment (including diagnostics first), the informed consent of the patient (to the treatment) implies consent that those data can be used by all others who are

involved in the treatment of said patient (like nurses, the pathologist, etc.) insofar as necessary for their role in the treatment and given that it has been made clear to the patient or it is obvious that those others will be involved. This transparency should be part of the informed consent process.

Furthermore most data protection legislation holds an exemption which makes it possible to use of personal data for research without consent, though usually under strict conditions. Those conditions differ from country to country, which makes it difficult to issue general statements about those conditions.

Sometimes the phrase 'informed consent' is used as well for the consent of the subject to use medical data for a secondary purpose. However, Directive 95/46/EC mentions 'explicit consent' (art. 8, section 2a) (3) (EU Directive 95/46/EC). The distinction can be explained as follows. With 'explicit consent', generally, a lower threshold for the information and decision making process is required than with 'informed consent'. This is understandable, as the situations covered by informed consent usually have a much more direct and deeper impact on the life plan of the person concerned than decisions which are covered by the explicit consent principle.

Informed consent relates to decisions about treatment or participating in medical research. If data are to be used for a secondary purpose for which consent is needed, like research, the term explicit consent should be used. However if there is active involvement, i.e. the participation of the subject, such as for answering questionnaires, then informed consent would be required.

Personal data

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 reference to an identification number or to one or more factors relating to is physical, physiological, mental, economic, cultural or social identity. For data to be considered identifiable account must be taken to the means reasonably likely to be used to identify the person (Directive 95/46/EC, art. 2 a and Recital 26) (3) (EU Directive 95/46/EC).

Anonymous data

Anonymous data are data which cannot be used to identify the person to whom the data relates. Data do not need to be 'absolutely' anonymous to count as anonymous data. It is

sufficient that data are not identifiable by the means reasonably likely to be used by the data controller or any other person who will acquire those data.

Though this 'reasonable test' sounds reassuring for researchers at first hand, it should be mentioned that authorities usually adhere to a high threshold before data can be considered as such, not distinguishing between the context in which those data are used, like that of (academic) research on the one hand and commercial parties on the other hand.

Sometimes researchers can make use of research exemptions in national data protection legislation for data which they consider not reasonably identifiable by them (unless they would use inappropriate means) but then these data are still considered personal data by the authorities.

To arrive from personal data to anonymous data in research two steps must be taken:
- The direct identifiers must be removed
- It must be assured that the aggregation level of the data is sufficiently high so as to make indirect identification not reasonably possible

In case of data linkage from different data sources adequate precautions have been undertaken that subjects are not identifiable due to the combination of data. More on this subject in Opinion 4/2007 of the art. 29 Working Party (Article 29 Data Protection Working Party, 2007).

Pseudonymisation

Pseudonymisation is the process in the data chain (see point 4.1.3 of the principles) by which direct identifiers of the data subject are removed and replaced by a unique pseudonym. Pseudonymisation can be done at multiple levels to decrease risk of re-identification. Another expression of pseudonymisation is coding of data. The pseudonym is then called a codenumber. The ISO report on pseudonymisation (ISO, 2008) states the requirements for such pseudonymisation that the resulting data can be considered anonymous data in the sense of Directive 95/46/EC. The ISO specification also contains requirements for privacy risk assessment design. However, not all pseudonymisation will meet those requirements. Hence we have made a distinction.

Pseudonymised (or coded) anonymous data

Those are data which are anonymous because of their aggregation level and because of the fact that the pseudonymisation process is sufficiently secure so that the pseudonym cannot reasonably be traced back to the individual by the recipient of those data.

Medical research with data cannot be done with a 'data soup' in which individuals cannot be unique discerned. Uniquely discerning individuals is completely different from identifying data subjects as understood by the data protection legislation. It simply means that the researcher can link research data to one person and not to another person, for example who did not have that disease at all or who was a completely different age etc.. Medical research with data is about finding patterns, not about finding persons.
Hence research with data always has two tiers:

- The unique number of the individual under which research data relating to said person are headed, and
- The data which are needed for the actual research

Pseudonymisation (or coding) means a way to generate such unique numbers in a sufficiently safe way so that the pseudonym cannot be reasonably traced back to the individual by the recipient of the data. Especially when data from various sources (but possibly relating to the same patients) need to be used such pseudonymisation procedures need to become more sophisticated.

The aforementioned Opinion 4/2007 acknowledges that pseudonymised or coded data can be – if certain conditions are met - considered as anonymous data (Article 29 Data Protection Working Party, 2007). However, the requirements are high, as already mentioned. Especially when data are two-way coded, as in that case someone at the start of the data chain will have the key to retrieve the identity of the subject through the coding mechanism.

One way or two way pseudonymised or coded data

With one-way coded data the direct identifiers of the subject are transformed into a unique pseudonym without the possibility that the entity who performs this transformation can do the reverse, from the pseudonym back to the direct identifiers, which means that no reference list back to the identifiers is available. With two-way coded the latter is still possible. Via the coding mechanism the identity of the subject can be retrieved from the pseudonym.

Coding or pseudonymisation can be more or less sophisticated depending on the intended data use. If at the source a simple list of patients were kept and every new entry would get a new sequence number then sent to the research zone, that would already be coding.

However, if many patients are entered in the study or data from various sources are to be combined, the coding needs to more sophisticated. Also to protect privacy and the claim that data are anonymous at the receiver, validated computational mechanisms need to be used to prevent potential identification. Such computational mechanisms are not defined in this document. In terms of the mentioned ISO report: one way coding would irreversible pseudonymisation and two way coding reversible pseudonymisation (ISO, 2008).

Pseudonymised or coded indirectly identifiable data

These are data which have been pseudonymised but nevertheless are considered personal data because they are still indirectly identifiable.
They might be considered indirectly identifiable either because of the fact that:

- the pseudonymisation procedure is considered to be insufficiently safe (the receiver might retrieve the identifiers either by contacting the sender or by 'cracking' the mechanism), or
- The aggregation level of the research data is insufficiently high

It should be remarked that research data can lead to re-identification as well. Gender, age (or age category) are always necessary research data, next to data about the disease, treatment etc. In many cases the researcher also needs to know at least the region where the subject lives. If the research question is about the consequences of environmental exposure a much more detailed location will usually be necessary. In combination these data might lead to re-identification, at least starting working on the presumption that researchers would be inclined to do so. Though that presumption can be challenged, that is beyond the scope of this explanation. See also hereinafter under the 'Trust' principle.

Aggregation level

As follows from the above (not a data soup), data on which research is based will pertain to specific subjects which must be distinguished from each other, even if they are anonymous.. Age group, gender, onset of disease, details about the disease, sometimes profession and location are relevant research data but can under circumstances be considered indirectly identifiable. The aggregation level refers to the degree of specification of these data. Certain classes of similar data can be grouped together. A higher aggregation level means that more data of that class have been grouped together, so instead of age by year, '5 year age groups'. Or instead of the full 'zip code, only 3 digits (for example the Dutch zip code has 6).

Fully anonymous data

These are data which are anonymous and are not pseudonymised, or if, the were, where the pseudonym has been replaced by random number or has been completely deleted.

Double coding

This is not the same as two way or reversible coding. It simply means that after one layer of coding another is applied. Like when data are transferred from the non-care zone to the research zone (see hereinafter). It should be mentioned that the EMEA document on definitions for genomic biomarkers etc.[2] uses somewhat different definitions than used in this document. Coding in the EMEA document meant two way or reversible coding as described here. Only the term 'double coding' is the same here as in that document. The EMEA document was issued before the ISO document on pseudonymisation and documents of the art. 29 Working Party of the concept of patient data (Article 29 Data Protection Working Party, 2007). We have followed the line of thought of those latter documents.

PET

PET are privacy enhancing technologies. Pseudonymisation is a PET. In the schemes we use the term 'data filter' for the application of PET. Sometimes PET result in anonymous data (whether coded or not) sometimes only in indirectly identifiable data.

TTP

TTP means Trusted Third Party. *The* TTP does not exist. In financial transactions the TTP has a different function than in dealing with medical data. In that context TTP refers to a reliable entity which is independent of the source of the data and the receiver and assures that the receiver will only receive coded-anonymous data. A TTP is most of all employed when data from various sources have to be uniquely combined at the receiver. As only the TTP holds the key to the coding mechanism, the TTP is equipped to assure that the identifiable data of the data subject at the various sources are converted into the same codenumber or pseudonym at the receiver.

Controller

[2] Note for guidance on definitions for genomic biomarkers, pharmogenomics, pharmocogenetics, genomic data and sample coding catagories EMEA/CHMP/ICH/437986/2006

A 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 processing personal data. (Directive 95/46/ EC, art. 2 under c).

Processor

The processor is the natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller (Directive 95/46 EC art. 2 under d). In the Dutch data protection Act there is added: and without being under the direct authority of the processor, meaning that an employee cannot be considered to be a processor. An example of a processor is when the data of the controller are stored at an applied service provider. The latter is then processor. The Directive requires a formal contract between controller and processor which contains clauses about the data security.

It must be noted that the terms 'processor' and 'controller' formally only apply to the processing of personal data. If the data are anonymous (whether pseudonymised or not), these terms do not apply. In that case the term 'holder' of the data will be used.

<table>
<tr><th colspan="2">Anonymous data</th><th colspan="3">Personal data</th></tr>
<tr><td rowspan="2">Fully anonymous data</td><td rowspan="2">Coded anonymous (pseudonymised) data</td><td colspan="2">Indirectly identifiable data</td><td rowspan="2">Directly identifiable data</td></tr>
<tr><td>Coded but either coding insufficiently secure

Or

Aggregation level too low</td><td>Not coded but aggregation level too low</td></tr>
</table>

Table 1: **Scheme of data**

3. Methods

To come to a formal description of the framework, a two step methodology was applied. In a first step, principles of the framework were formulated based on existing rules and regulations for confidentiality and data privacy on the European level. For this first step no formal methodology was used. Techniques that systematically extract and manage requirements from laws and regulations in order to support requirements compliance to such laws and regulations is still a matter of research (Islam, 2010). Nevertheless, we extracted the basic principles in a non-formal way to provide input for a formal description of the framework. Therefore, the principles can be seen as an intermediate step between rules and regulations and a formal description of the framework. The principles are presented as a series of statements in section 4 of this document.

Based on the principles formulated, a structured representation of the framework was developed as a next step. The formal description of the framework was based on elements of structured analysis and data flow diagrams. Structured analysis views a system from the perspective of the data flowing through it (Yourdon, 1989). The function of the system is described by processes that transform the data flows. A data flow diagram is a graphical representation of the flow of the data through an information system.

In order to come to a formal description of the framework, adequate representation of context is of utmost importance. With respect to data flow and data use it makes a major difference whether data have been collected in the context of medical care, have been derived from primary medical care data or have been collected within clinical trials or cohort studies. Instead of representing context by context diagrams, a technique used within structured analysis, with the system in the middle (e.g. EHR) and actors outside that system (e.g. researcher), it was decided to use the full range of possible data sources and to divide it by zones with the idea to represent each different context by a different zone. Data transfer between zones is described by data flow diagrams, illustrating the data flow from one data source in one zone to another data source in another zone. To make this data flow possible, specific functions (processes) are applied based on Privacy Enhancing Techniques (PET) in order to ensure optimal privacy protection for the patient (e.g. anonymisation, pseudonymisation, coding and data aggregation). The different structural elements of the formal specification of the framework together with its notation are described in the next sections.

3.1 Zones

Zones are defined as areas for data sources that are comparable and similar with respect to purpose, rules and regulations for use. The idea is to have personal identifiable data, pseudonymised data and anonymous data in different zones to structure the data sources zones according to risk for confidentiality and data privacy issues, representing areas with **low, medium and high risk of identification** (cancer Biomedical Informatics Grid (caBIG,2011). In the framework two major zones are defined:

- data source zone
- research zone

The data source zone contains data sources available and needed for research. These data are transferred to the research zone after adequate transformation.

Examples of data sources in the data source zone are:

- primary not aggregated data in electronic health care records from the original care provider
- databases which have been derived from such records and been structured in a certain way. They will be derived from one type of source though, like GP electronic health care records (EHCR's)
- databases which combine data from various types of sources, like GP EHCR's, hospital records, etc.

Next to these there are databases based on:

- answers given by participants of cohort studies
- databases based on analyses of human tissue (whether or not combined with electronic health care data, cohort studies, combinations of databases)
- databases, where answers of various cohort studies and databases on analyses of human tissue are combined
- databases which are based on Case Report Forms (CRF's) from participants in a clinical trial
- etc.

The data source zone can be split into two zones:

a) care zone

b) non-care zone

The **care zone** is dedicated to data for patient diagnosis and treatment. The data have been collected in the medical care context by the care provider. In the care zone there are personal identifiable medical data, which can be used within the care context by the treating physician. It needs explicit consent by the patient and/or authorisation for use by researchers outside the care context.

The **non-care zone** contains research databases, registers, etc. This zone covers databases that have been derived from primary medical care data (e.g. General Practice Research Database (GPRD), Netherlands Institute for Health Services Research (NIVEL), data bases from cohort studies, databases on analysis of human tissue or databases based on CRFs from clinical trials. Data sources in this zone are normally pseudonymous. They are either based on explicit consent or the research use will be based on country-specific regulations (e.g. exemptions to consent for research) usually allowing for an opt-out mechanism. The use of these databases is authorized and controlled by the data controllers of these sources based on a defined policy. Data sources in the non-care zone may contain primary data (e.g. clinical trial database, cohort study database) or secondary data (e.g. GPRD, cancer register).

The defining characteristic of this non-care zone is that the data are in some way aggregated. The zone holds data of many patients or persons and these data can be searched for certain characteristics pertaining to certain classes of patients.

To the **research zone** data are provided that are needed for a specific research project. Based on a research protocol and specific research questions only those data arrive at the researcher, which are necessary for the research at hand. The research zone normally contains data that are truly anonymous or at least coded anonymous for the researcher.

The boundary between the data source zone, which also contains research databases, and the research zone shifts according to the viewpoint. The research zone is only opened, if there is a research project with a research question. So, for example, NIVEL can be as well in the research zone (if a research question is posed by NIVEL itself) or in the data source zone (if data from NIVEL are extracted for a researcher for a new research question).

The three zones are illustrated in Figure 1. Zones are represented by rectangles with different shading throughout this document for the care, non-care and research zone. For data sources the database symbol is used as notation from data flow diagrams.

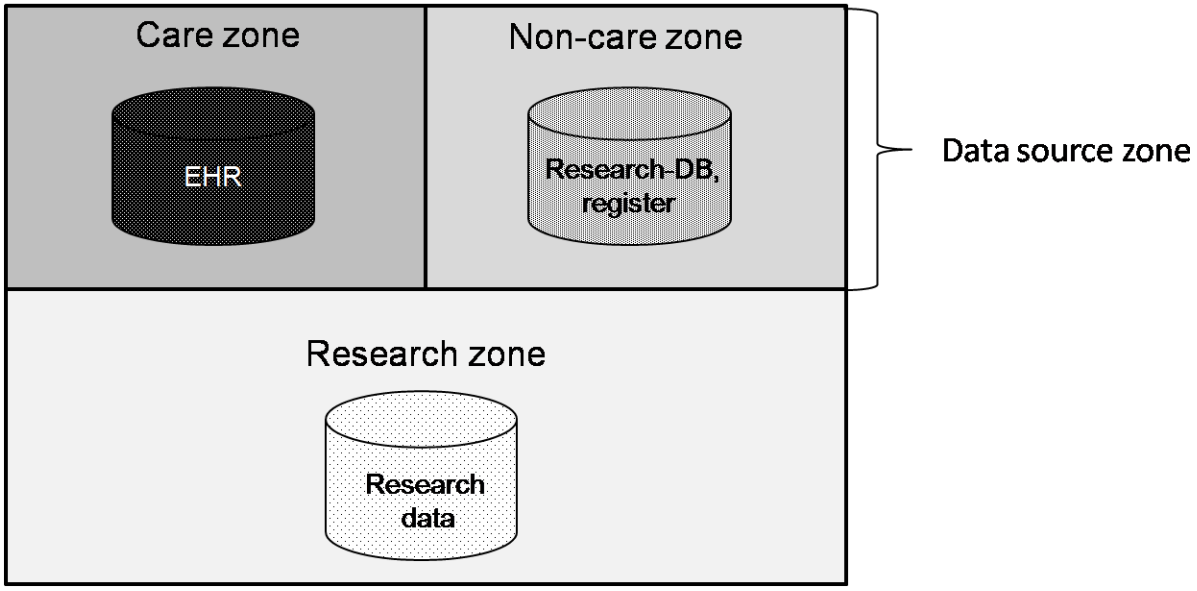
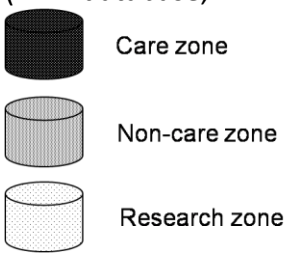


Figure 1: **Notation for the zone model of the framework**
(DB = database, EHR = electronic health record)



Data bases from the care zone always mean directly identifiable data, data bases from the non care-zone usually indirectly identifiable data (whether pseudonymised or not) and data from the research zone anonymous data (whether pseudonymised or not). As already mentioned, a data base in the non-care zone can also be in the research zone. The NIVEL database is a research database in its own right for NIVEL. It can be a secondary database for other research projects and therefore is part of the non-care zone.

3.2 Subzones

Databases in different countries operate under different rules and regulations concerning confidentiality and data privacy. Even in one country, differences between different types of data bases exist that needs to be considered. Therefore the main zones defined (care zone, non-care zone and research zone) might be insufficient to be of practical value for TRANSFoRm. For that reason, **subzones within** the main zones were defined, which contain data that are comparable and can be used for the same or a very similar purpose and with similar applicable rules and regulations for their use. Examples of subzones are:

- GP EHR's within one country is one subzone (similar types of data and similar applicable rules and regulations)
- GP EHR's from another country are a different subzone (often somewhat different data and certainly different applicable rules and regulations)
- For EHR in hospitals the same applies: a subzone different from GP EHR's and different from country to country
- Some registries may subsumed under one subzone and other may not
- Some research databases may be subsumed under one subzone and other may not. Like if data for X and Y can be used for research Z and the rules and regulations for X and Y are both similar, it would be one subzone. But if X data cannot be used for Z while Y data can, it would not be one subzone. And if the rules and regulations would be different though both can be used for Z, X and Y would be different subzones and some type of filter would be needed to merge X and Y for Z

For the research zone no subzones are defined. The reason is that the framework describes the data privacy and confidentiality requirements applied to one specific research protocol and research question at a time. Working on different research questions in parallel is not considered in the model in order to simplify the framework. This makes sense because within the research zone normally anonymous data are considered, which are not personal data and outside the European Data Protection Directive.

The subzones are illustrated in the figure 2:

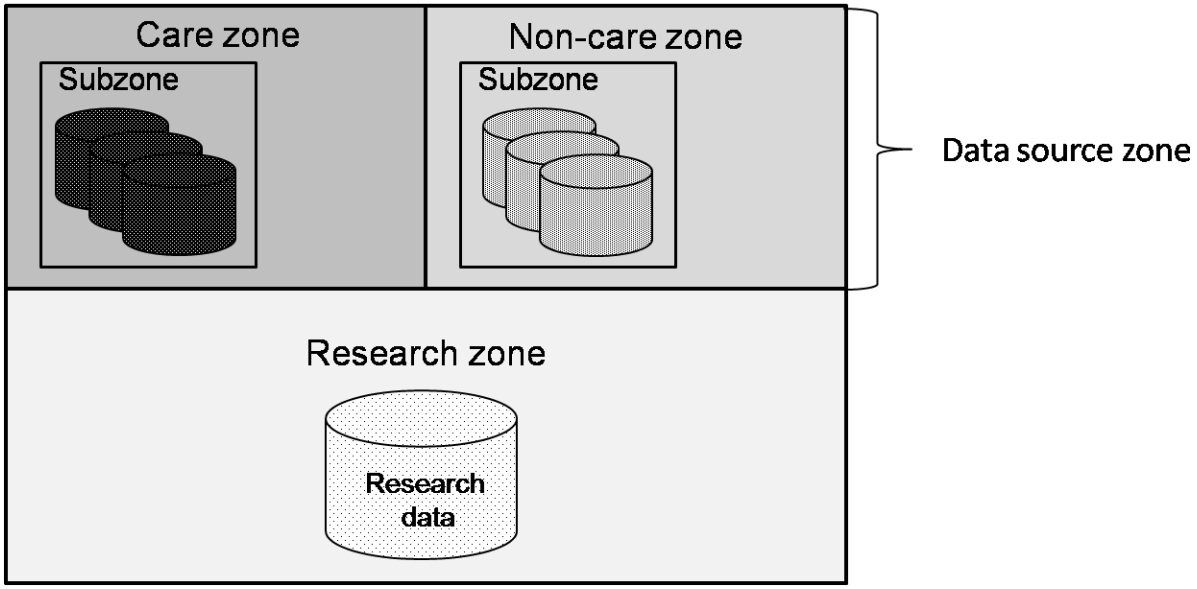


Figure 2: **Notation for zones and subzones in the framework**

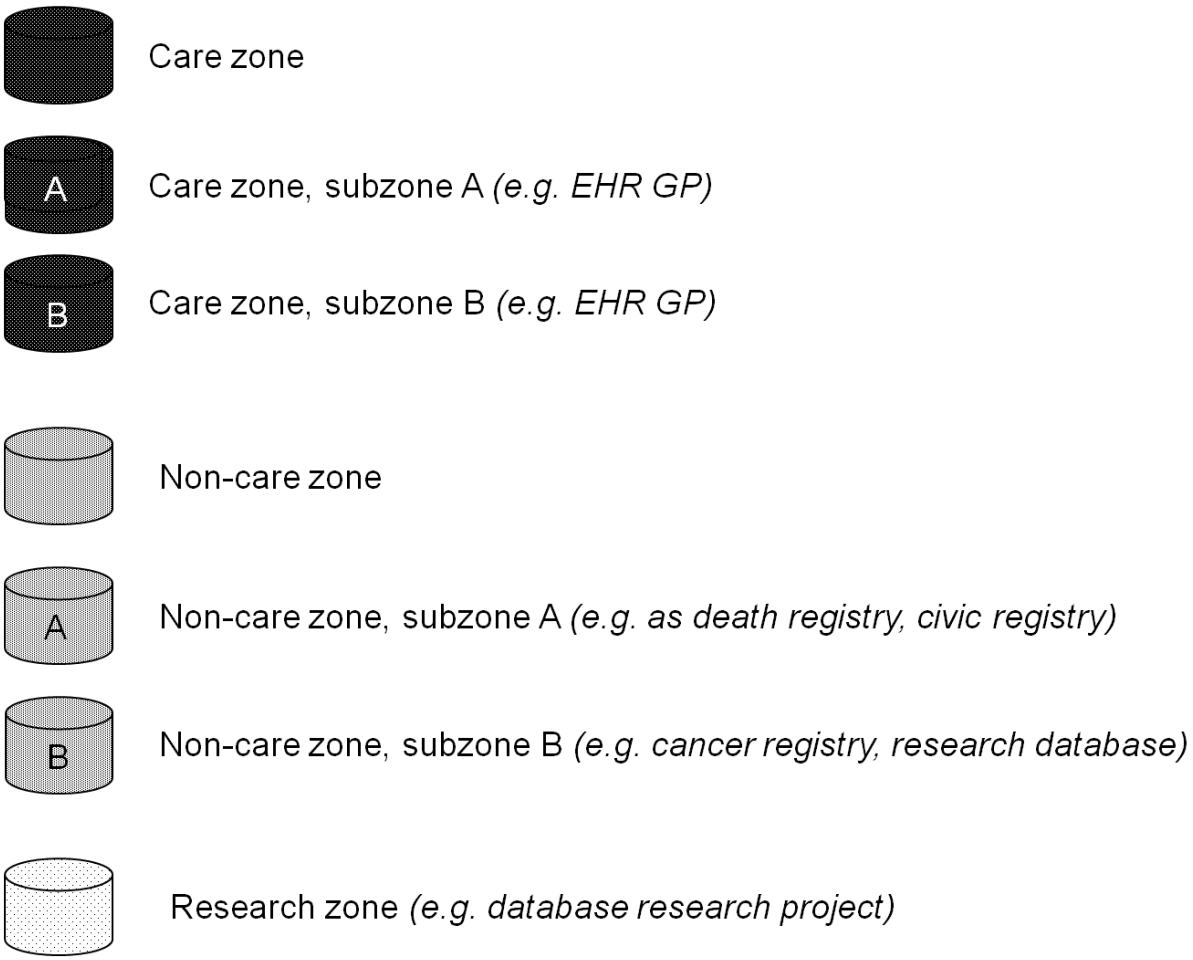


Figure 3a: **Notation for data bases in the different zones/subzones**

3.3 Privacy filter and data linker

The most important aspect is the flow of data from the data source zone to the research zone. If this flow is possible according to the data use policy and the rules and regulations to be applied, data are transferred from a zone with high or medium privacy risk to a zone with low identification risk, enabling the research intended. In order to transfer data between zones and subzones, it is necessary to use as much Privacy Enhancing Techniques (PET) as possible ensuring the optimal privacy protection for the patient. Major functions are anonymisation, pseudonymisation, coding and data aggregation. In our framework Privacy Enhancing Techniques are represented by **privacy filters**. Therefore privacy filters are

neither data nor belong to zones or subzones, however, they are operating on data bases and allow transfer between different zones or subzones.

For some research projects it is necessary to combine patient-specific data from different data bases to answer the research question. The procedure necessary is data linkage. Therefore, in our framework **data linkers** are introduced, allowing linkage of data bases within or between zones/subzones. Linking may be performed by **one-way coding** or by **two-way coding**. Linked databases are allocated to a zone or a subzone.

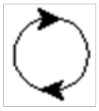
Data bases may be combined in various ways making use of privacy filters (PET) and data linkers (coding procedures). Linked databases may belong to the same zone or subzone or may have been transferred to another zone.

3.4 Actors and roles

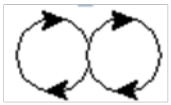
In order to allow a full representation of the framework, the different actors involved in the framework have to be defined. Major actors are clinicians, which may have the role as treating physician and/or researcher. Other actors may be EHR, a database in the non-care zone, the patient, acting staff for the treating physician or the data controller of the non-care database and the software developers of EHR's or the processor of a database acting on behalf of the controller. Persons are represented by stickman and data sources by the data symbol already introduced. The notation for data transfer, functions (privacy filter, data linker) and actors is presented in Figure 3b:

Treating physician if in care zone ,
researcher if in research zone

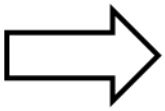
Privacy filter



Data linker, one-way coding



Data linker, two-way coding



Data transfer

Research question

Figure 3b: **Notation for data transfer, functions and actors**

4. Description of the Framework

4.1 Principles

4.1.1 *General Policy*

The framework should take into account both the privacy interests of patients participating in these research projects or whose data will be used for research and the interests of those persons who will hopefully benefit from the medical research performed in TRANSFoRm. Hence, next to privacy protection, the population has also a right that medical research will be performed from which they might profit.[3] (5). The framework should not be unnecessarily biased towards either type of interests . but should try to balance between data protection and the use of data for research for the benefit of future patients

4.1.2. *Prime responsibility for the treating physician*

The treating physician has access to his patient data in the medical care context. His primary obligation is to use these data for best possible treatment of his patients, according to the present standards for good medical care. This framework starts from the presumption that the treating physician has an obligation to raise these care standards as well. Hence, he should try to use patient data acquired in the medical care context for medical research, or at least allow such use, according to the applicable national regulatory conditions. In most systems these conditions allow for such research *by* the treating physician.

However, if the data would be transferred to a third party a different regime would apply. The treating physician is bound by the professional medical secrecy but at the same time may allow use of patient data for research if explicit consent is given, or if a research exemption (to the consent principle) would apply or if the data are rendered anonymous (whether coded or not). As data controller the treating physician retains the ultimate responsibility for allowing use of these data by third parties in whatever form.

[3] This right was stressed at the TRANSFoRm invitational conference with patient organisations on February 9 2011 in Brussels. See also Academy of Medical Sciences (UK), A new pathway for the regulation and governance of health research, London, 2011.

4.1.3 *Data chain and following data flow*

Unless the research is done by the treating physician, there is a chain of data. Data from many treating physicians are combined to acquire sufficient statistical power or data from various sources are combined like those of treating physicians with cancer registries, death registries and the like, to link original diagnosis and treatment regime with final outcomes (and to learn valuable lessons from that combination). So data are transferred from one source to another. There is a data chain or 'flow of data'. The following conditions apply to this data chain in general (next to conditions mentioned later):

- As much as possible privacy enhancing technologies (PET) (see the point 4.1.4) should be used in the chain
- The data chain should be generally transparent to the patient when consent is required
- The chain should be as short as possible for the required research purposes;
- In the chain it should always be clear who is the controller of personal data (when those are used) and who is the processor
- The chain should be embedded in Data Transfer Agreements (DTA) between the provider of the data and the recipient. Those DTA's should, however, not lead to unnecessary bureaucracy. If there is a database policy (or statutes) of the recipient and controller or holder of the database which complies to the standards mentioned in this document, it should be sufficient that the provider of the data declares that the data are transferred in accordance with that policy

4.1.4 *The use of privacy enhancing technologies*

The aim should be to use as much Privacy Enhancing Techniques (PET) as possible ensuring the optimal privacy protection for the patient. Whenever possible, anonymised data should be used. To retain their value for research, these data should in principle be coded-anonymous. The processing of such anonymous data does not fall into the scope of the Data Protection Directive. However, the decision whether they can be considered anonymous, does fall within its scope. Here we see some ambiguity between the member states. There is a limit to using PET to render data anonymous while they still can be used for research. It might happen that the aggregation level of the research data becomes that high that they cannot be used to answer the research question anymore, especially when rare diseases are involved

and/or intricate relations between environmental or work place exposure and the onset of disease. This aspect is sometimes underestimated by those who see PET leading to anonymous data as the ultimate solution for the tension between privacy and research for the public good.

But PET should not only be used to arrive at anonymous data. At the level of researchers where the research institute can be considered the controller of the personal data, researchers should only have access to those data which are needed for the research. In 'cohort studies' where people participate with answering questionnaires and might give consent to use data from health care records and other sources, a distinction is made. The administrative staff knows the names and addresses to send the questionnaires. The researchers only receive answers and other data under the unique participant number. That is using PET as well.

4.1.5 *Explicit consent*

In principle, explicit consent from the patient should be obtained whenever direct or indirect identifiable data are used by a third party. Consent seems at first hand the most adequate expression of the autonomy of the patient and hence is forwarded as basic in nearly all legal and ethical documents. However consent to treatment or participation in a trial has a different meaning than consent that someone data can be used in observational research or health research in general. Hence Directive 95/46/EC leaves room for exceptions on the consent principle for statistics and health research in the data protection regulations of the member states. As mentioned already, this possibility has been used in a very divergent way in the member states, yet most member states have some form of research exemptions in their data protection and/or health legislation.

As explicit consent can have disadvantages as well, like an undue burden for health care providers who have to ask for consent, bias in the data of those who will be included, often to the detriment of the less privileged groups. Exclusively relying on consent would jeopardise the balance mentioned at the start of these principles. Within WP3 "Legal, ethical and security framework" (WT 3.2.1) we have discussed the literature and at a session with patient organisations how patients and their organisations often want more use of data for research, also without consent and sometimes oppose the attitude of distrust (see the next principle) often seen with data protection authorities.

Using coded anonymous data is one way to solve the tension, but is not always feasible. Sometimes those data become then to crude to be used for research. A lighter form of PET should be used next to perhaps a lighter consent modality like opt-out in conjunction with research exemptions insofar as the member state allows for this exemption.

However, when explicit consent has been given, the person concerned can always withdraw his or her consent as long at the data can still be traced to this person within the chain. In such cases the data will be either destroyed or be made fully anonymous (by deleting the link between the identifiers of the person and the pseudonym).

4.1.6 *Trust (which must be earned)*

Though the emphasis is the use of PET in TRANSFoRm it should not be forgotten that neither technical solutions nor consent procedures can replace 'trust', where trust means that personal data are used wisely and diligently in the context of medical research to improve the conditions of other patients. Researchers should not use inappropriate means to retrieve the identity of patients and in Europe there are no examples of such inappropriate behaviour known to us. Researchers are dependent on that trust to receive data from patients, care givers and other sources . As said, patients have an expectancy that researchers in health research can – in general - be trusted.

Yet, in the present age more needs to be done to earn that trust. Good research governance comes into play here, amongst other thing meaning how researchers will interact with patient organisations. The other aspect is data security in the chain and in research zone especially. It is one solution that only anonymous data will reach the research zone. Many of the figures in the next chapter are based on that type of solution. We forward here that it might be a much more profitable approach to look at those software and procedural solutions that it can be reasonably ascertained that indirectly identifiable data in the research zone will be secure there *and* will not be used to identify the patient either by the researcher or anyone else. Many of the approaches to data security can be used here, but most software used at research centres for health research does not yet allow the kind of authentication, logging, etc. needed to implement those approaches.

4.1.7. *Use by third parties*

In principle third parties should only use coded- anonymous data for research. PET should lead this situation. Any exception to this principle should either be based on explicit consent of the subject or national exemptions to use data for research without consent.

4.1.8. *Pseudonymisation*

For third parties pseudonymised data can also be classified as anonymous, if adequate conditions hold (secure coding, no direct or indirect re-identification by third party possible).

4.1.9. *Data controller*

Each data controller of a database ensures compliance of the data processes with the applicable national/regional data protection legislation. In the data chain a third party receiving data should not be responsible for the regulatory compliance of the sending organization. Access to existing databases and use of such data is the responsibility of the data controller of the existing database until the third party has taken over responsibilities as data controller by forming a new database.

The same applies correspondingly if we cannot speak of a controller in the formal sense, as there are no personal data involved. In that case the responsibilities are vested in the holder of the database.

It should be mentioned that as a principle access can also be granted indirectly. The third party can ask the controller of database to perform queries on his behalf according to the policy of the database. The reply will be aggregated, anonymous data which the third party can use for its research. Those data are then transferred to the 'research zone' as defined by TRANSFoRm.

4.1.10. *Contractual agreements and database statutes*

As mentioned contractual arrangements should regulate the transfer of data in the data chain. Our aim is stipulate model Data Transfer Agreements (DTA's) for the TRANSFoRm partners. The process of making individual 'tailor made' DTA's will be greatly eased if the database to which data are transferred has a database policy or 'Statutes' stipulating the acceptable sources of the data, the PET used for collecting

those data and their 'processing' within the database, the aim of the database, the conditions under which the database can be used for research and etcetera. Once the major databases with TRANSFoRm have been distinguished, our aim is to give guidance to the drafting of the 'statutes' of these various databases insofar as this has not been done already by the holder of the database.

4.1.11. *National legislation*

Member states may introduce exceptions for reasons of substantial public interest and subject to the provision of suitable safeguards. For the use of existing databases it is necessary to examine national exemptions for scientific research. We recommend the use of national exemptions for research purposes as much as possible.

4.1.12. *Publication*

Results from analysis of data contained in databases should only be published with non-identifiable data (aggregated and fully anonymous).

4.1.13 *Data sharing*

TRANSFoRm should endorse the principles of data sharing. This means that under conditions of data protection data should be shared as widely as possible within the scientific community. This will avoid duplication of research and lead to faster availability of validated data which can be used in patient treatment. Data sharing does, however, not mean to give data away. The original researcher can require a reasonable contribution if further elaboration of the data necessary for the subsequent researcher and a fair acknowledgment, in whatever form, depending on his/her contribution also as co-author, in the publication which results from the subsequent research. By implementing adequate provenance measures traceability of data usage can be provided if requested by the data owner or controller.

4.1.14. *Genetic data*

Each person can be potentially identified by their genetic data in a data sharing and linkage context when no strict contractual arrangements are made for how data may be used, for what purpose and when the remaining or unnecessary data will be destroyed. Therefore genetic data poses a higher risk to be indirectly identifiable. Special measures to ensure privacy have to be taken (e.g. deletion of data not

necessary for research question, control of data exchange between partners, clear endpoints in the chain where the data will not processed any further) in the case of transfer and access. Processing of identifiable genetic data is prohibited if the subject has not be given explicit consent or if none of the exemptions of laws and regulations would apply.

4.2 Formal description of the framework

Using the methods and notation described in the section “methods”, different typical research scenarios are represented as a first step to illustrate the formal description of the framework (see 4.2.1). In the next step the use cases specified in WP1-WT1.1 “Development of use cases” (Version 2, 13.9.2010) are represented (see 5). Research use case 1 deals with “Type 2 Diabetes genotype phenotype study” and use case 2 with the “GORD study.

4.2.1 Representation of different research scenarios

Scenario 1: Merging data from EHRs

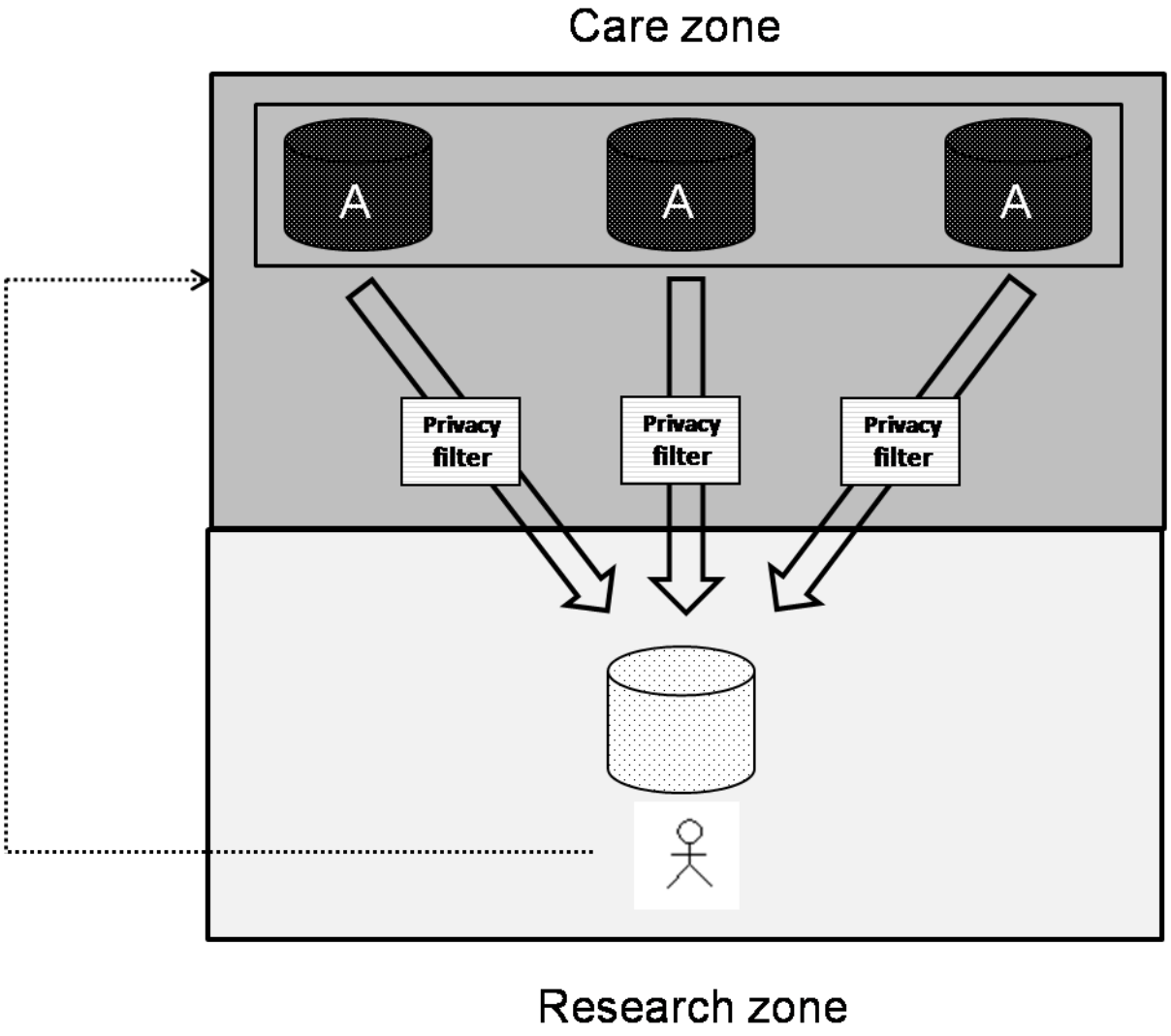


Figure 4: **Representation of scenario 1**

In this scenario data from EHRs of different GPs belonging to the same subzone A (like because the GPs are in one country) within the care zone are merged in a research database after passing a privacy filter. The data in the research zone are anonymous. No linkage on a patient-level has been performed (vertical linkage), there has been a merging of cases (horizontal linkage). As this only about anonymous data (e.g. counts), it might very well be possible to treat EHR's from various countries in a similar way. Fully anonymous data can be used for research in every country. However, some Data Protection Authorities use a higher threshold when data are not indirectly identifiable (because of the aggregation level) than

other. Next to what data are recorded in the EHR might differ from country to country. But in general it can be said that this example gives the least complications to merge data in the research zone.

Scenario 2: Linkage and merging of data from different subzones in the care-zone

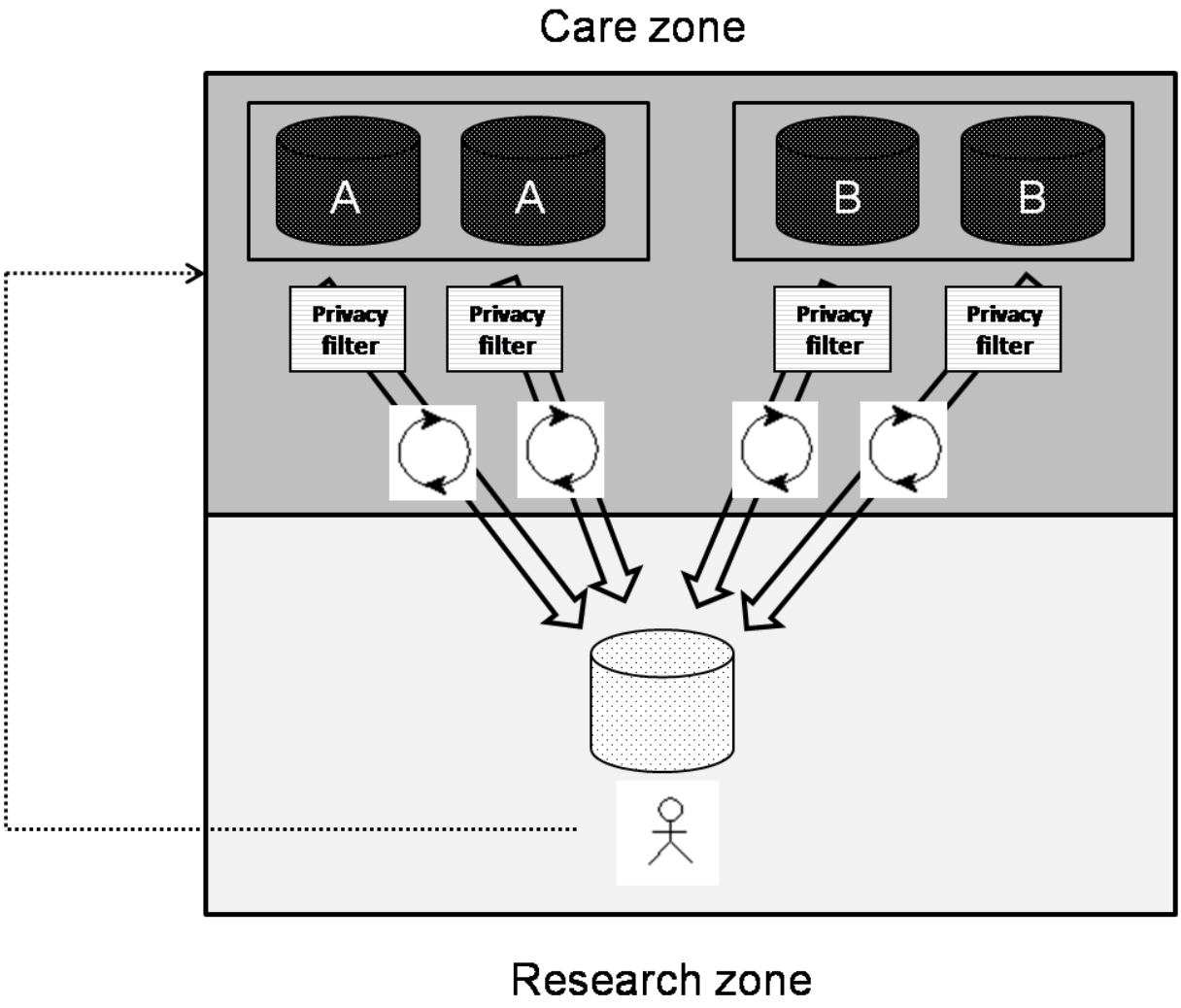


Figure 5: **Representation of scenario 2**

In scenario 2, privacy filtering and data linkage with one-way coding is applied to two EHRs in Subzone A (i.e. in one country) of the care zone. A similar procedure is applied to two databases in subzone B (.e. in another country or from different care providers within one country) of the care zone. The two linked databases are merged for the research database.

Scenario 3: Data transfer from non-care database to the research zone

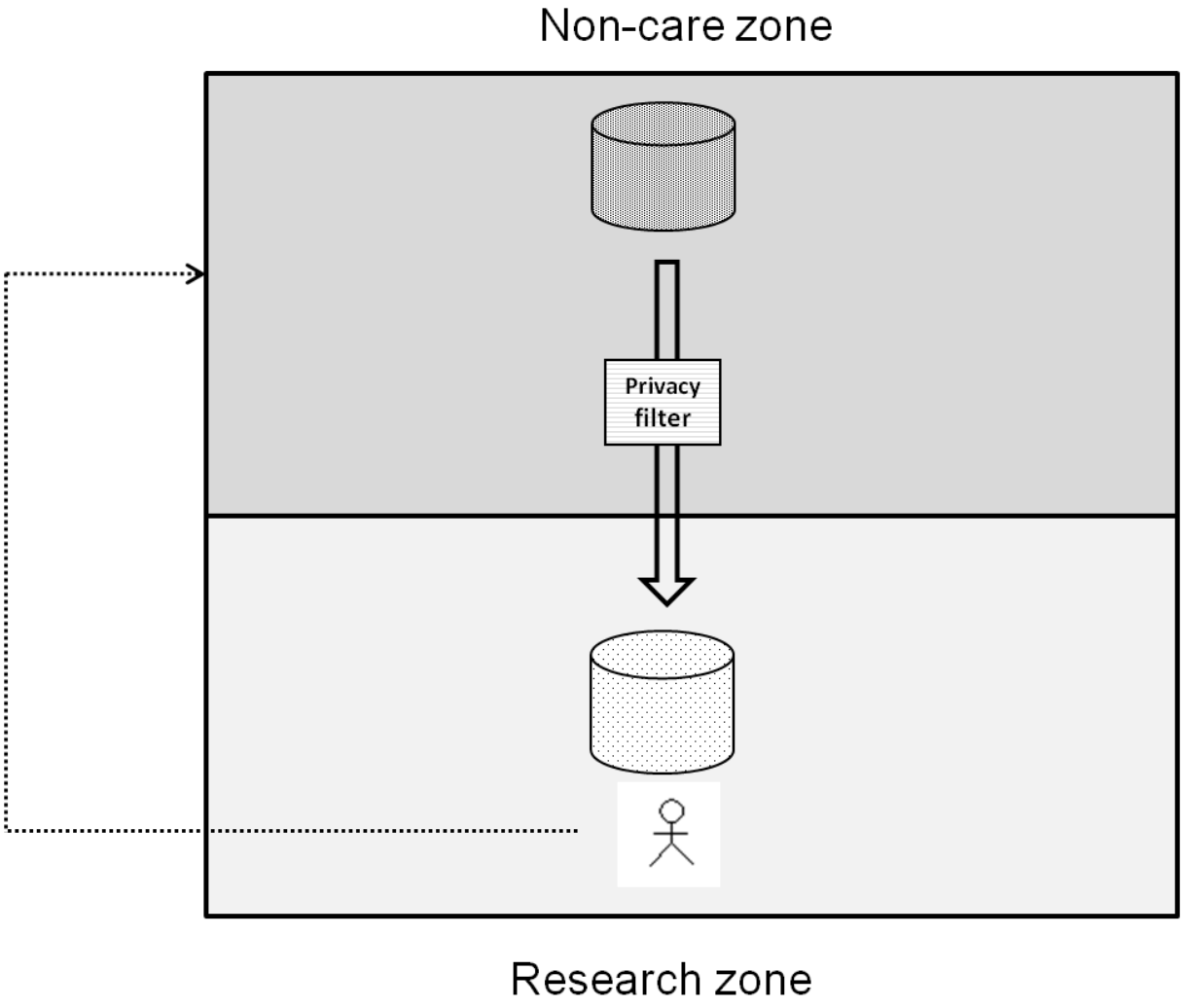


Figure 6: **Representation of scenario 3**

Scenario 3 describes a simple data transfer from a secondary database (e.g. cancer registry, research database) in the non-care zone using privacy filtering to a database in the research zone.

Scenario 4: Linkage and enrichment of data bases in the non-care zone

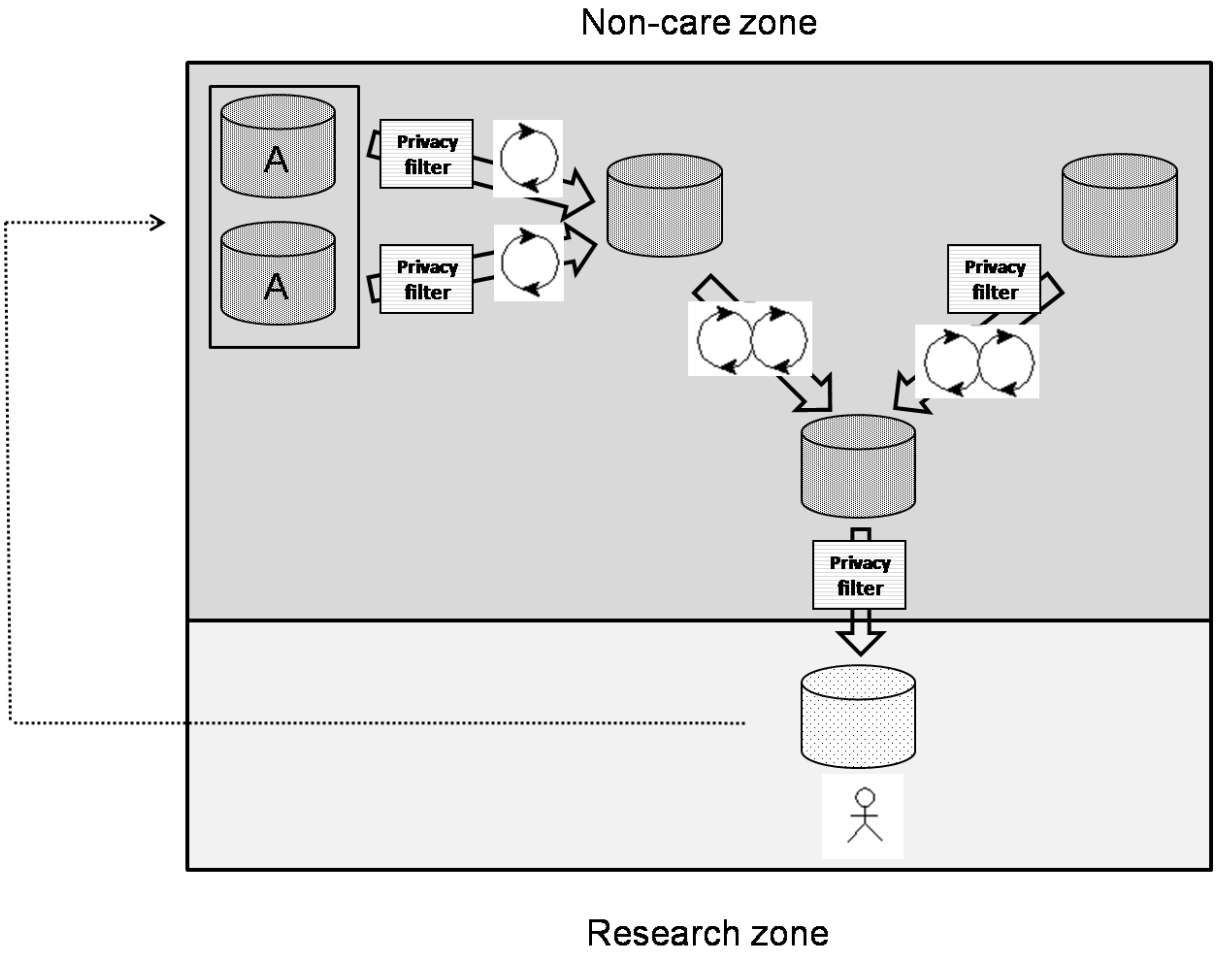


Figure 7: **Representation of scenario 4**

Scenario 4 is a bit more complicated. It represents a situation where two data bases in the same subzone of the non-care zone, such as cancer registry, 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 passed a privacy filter. The resulting database is transferred to the research zone. The aim of this scenario is the enrichment of a linked data base with additionally data from a secondary database.

Scenario 5: Linkage of a cohort study database with data from the care and non-care zone

To explain scenario 5 the type of data bases to be used later is characterized in a first step. This is illustrated in figure 8a. Here, no research zone is added. Instead the present situation in many countries in Europe is described. The scheme gives *examples* of those databases in the Netherlands. The fact that these databases exist, does not mean that they can be used for research without further conditions applied to that research.

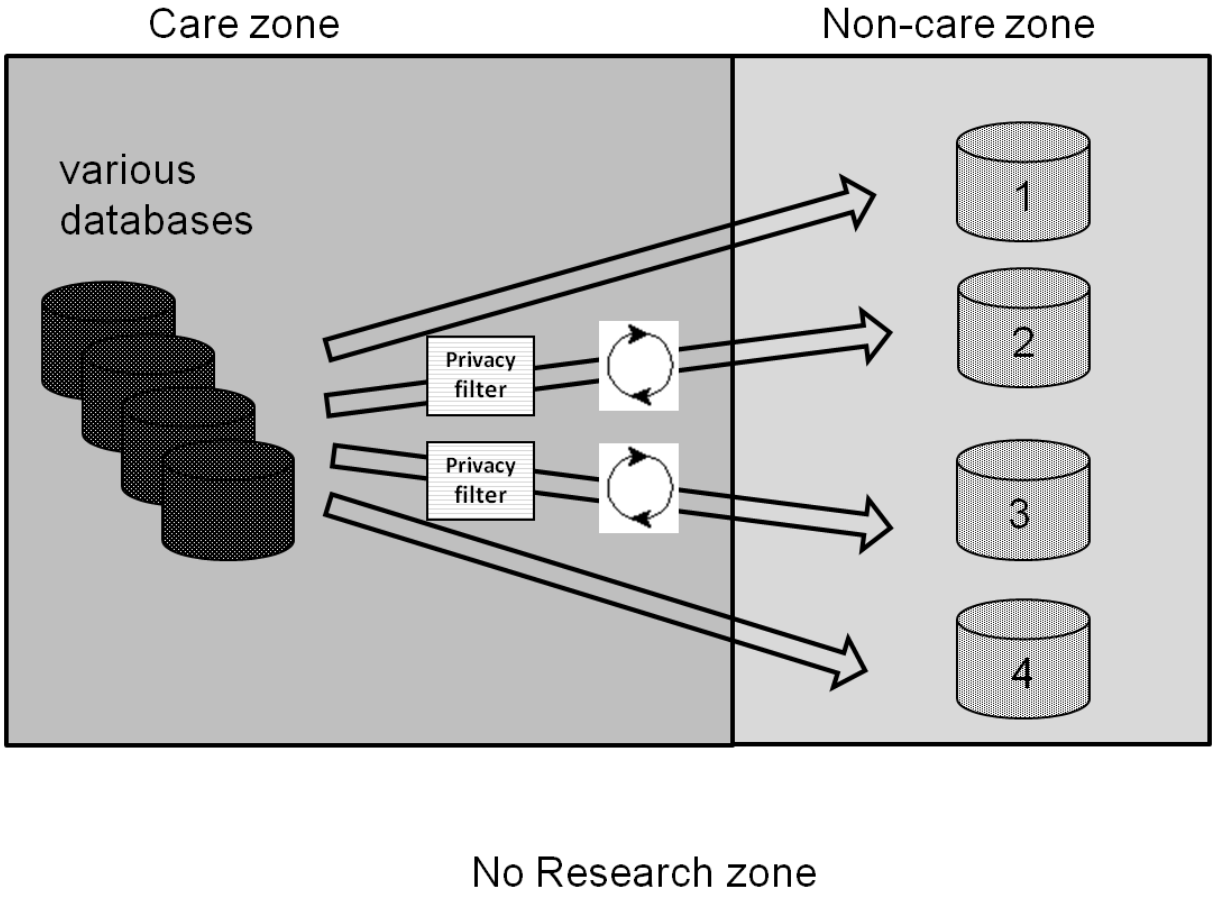


Figure 8a: **Illustration of data bases in the non-care zone**

1 Data about medical consumption at the health care insurers (specified format, rather general, but of more than 99,9 % of the Dutch population, though spread over the respective insurers). Statutory (for the health care providers) and contractual (for the insured) obligation to report, necessary for the reimbursement scheme of the health care system. Directly identifiable by the insurer.
2 LINH, detailed though only of general practitioners participating in LINH, coded-indirectly identifiable

3 National cancer registry, covering about 99% of all cancer patients in the Netherlands, indirectly identifiable.
4 Death registry at the Statistic Netherlands. Statutory obligation to report cause of death by physician declaring the death of a person. Directly identifiable by Statistic Netherlands. Can be used for research if a person has given consent when alive and in that case by using those directly identifiable data. Otherwise data must be one way coded (or fully) anonymous for research.

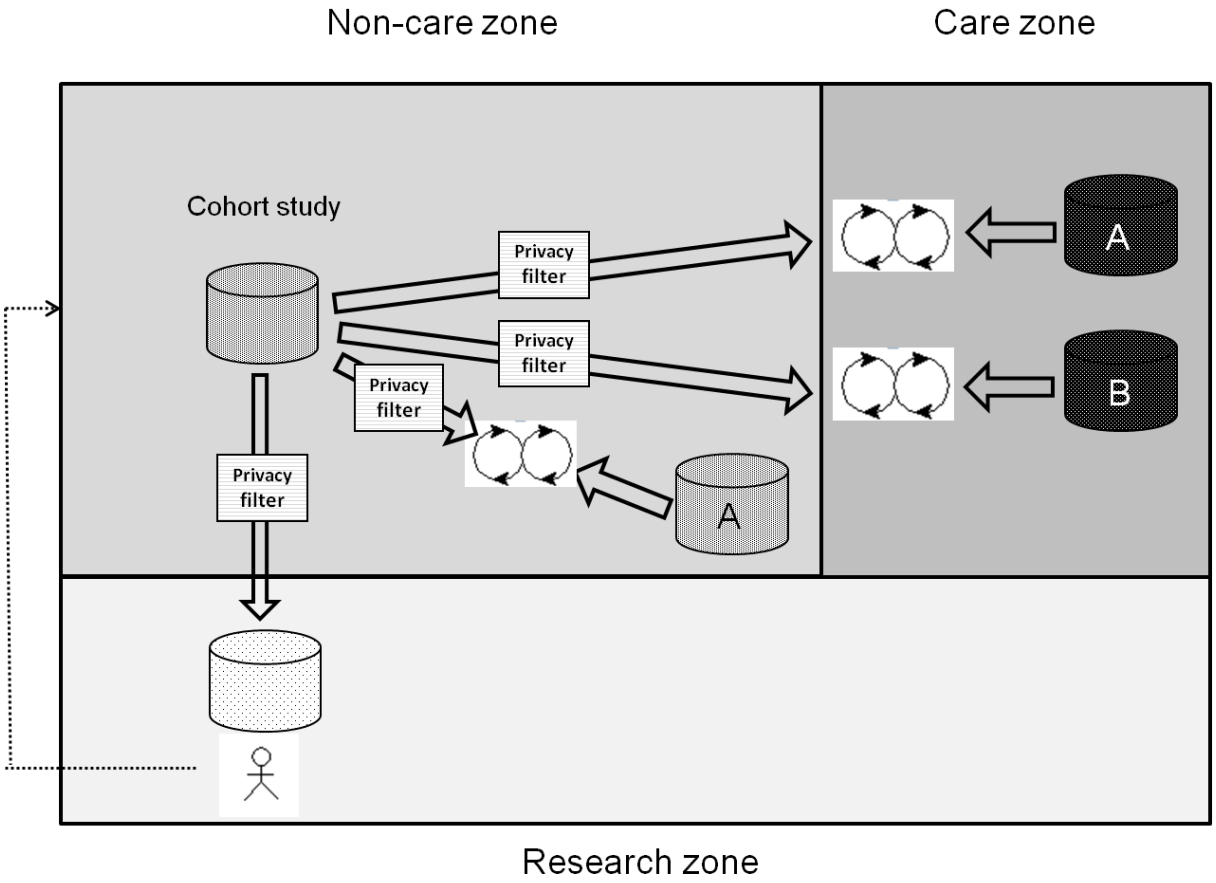


Figure 8b: **Representation of scenario 5**

Scenario 5 is even more complicated than scenario 4. This is a cohort study, where participants have given informed consent to link their data with other sources. The data are collected in the cohort study database. From this data base data are sent to the other databases. There is a privacy filter when sending out the data, as the cohort database only gives those data which are necessary for the linking. There is no privacy filter back, it will get all the needed data (based on the informed consent by the participants) from the care zone or from a database in the non-care zone. The enriched database will be used to fill the end research database using privacy filters, but no linking is necessary as that has been done already in the cohort database. In this scenario data sources from the care zone and from the non-care zone are linked and transferred to the research zone after privacy filtering.

5. Framework applied to TRANSFoRm use cases

The TRANSFoRm project aims to radically advance the understanding of ICT system interoperability relevant to healthcare and clinical research and to develop an EU-wide system capable of integrating Primary Care electronic health records systems and research systems. To develop and validate the software and services two representative research use cases (a genomic-phenotype study and an RCT) are being developed. The use cases have been described in WP1-WT1.1 “Development of use cases” (Version 2, 13.9.2010). Research use case 1 deals with diabetes and use case 2 with the gastrooesophageal reflux disease:

Use Case 1:

TRANSFoRm Genotype-Phenotype association study in type 2 diabetic complications using eHRs and other health related databases

Use Case 2:

Quality of care and evidence for treatment of gastrooesophageal reflux disease

For the diabetes study the following use cases have been defined (WP1-WT1.1 “Development of use cases” (Version 2, 13.9.2010):

1. present system options
2. authorize data selection, extraction and linkage
3. select patients
4. extract information
5. link and reintegrate data
6. present data

For the GORD study the following use cases have been defined (WP1-WT1.1 “Development of use cases” (Version 2, 13.9.2010):

1. find patients for randomized clinical trials (RCT)
2. eCRF
3. measure QoL
4. randomize
5. matching
6. extract information
7. linkage (general), including anonymisation storage (general)
8. storage

5.1 Classification of use cases for this document

Use case 1 (present symptom options) from the diabetes study is dedicated to meta information about databases, EHR, etc.. This information is not relevant for the confidentiality and data privacy network and will therefore not be considered in this document. Use case 2 (authorize data selection, extraction and linkage) from the diabetes study is a prerequisite for every use case. Therefore, the authorisation procedure will be part of all use cases discussed in this document. In addition, the following use cases have been defined but will not be discussed in this document: randomize trial patient (use case 4, GORD), data collection by GP (use case 2, GORD) and data collection by patient (use case 3, GORD). These use cases are related to clinical research therefore fall under the rules and regulations of clinical trials (e.g. EU Directive 2001/20/EC).

The following use cases will be considered in this document:

1. **counts of patients with a defined pattern**
   This is not explicitly contained in the use cases.

2. **find patients for clinical research**

3. **select patient for research question**

4. **extract information of selected patients.**

5. **linkage of data**

The relationship between the original use cases (diabetes, GORD) and the use cases discussed in this framework is presented in table 2.

Use cases		**Use cases decribed in this framework**
Diabetes study	**GORD Study**	
1. Present smptom options	-	Not relevant for this framework
-	-	**1.Counts of patients with a defined pattern**
2. Authorization data selection, extraction and linkage	-	Relevant for every use case

3. Select patients	1.Find patients for randomized controlled trials (RCT) 5.Matching (partly)	**2.Find patients for clinical research**
3.Select patients		**3.Select patients for research question**
4. Extract information	6.Extract information	**4.Extract information of selected patients**
5. Link and reintegrate data	7.Linkage (general) including anonymisation storage (general)	**5.Linkage of data**
-	2.eCRF	Not tackled in this framework (related to clinical trials
-	3.Measure QoL	Not tackled in this framework (related to clinical trials
-	4.Randomize	Not tackled in this framework (related to clinical trials

Table 2: **Relation between original use cases (diabetes, GORD) and the use cases considered in this framework**

5.2 Use case: counts of patients with a defined pattern

The aim of this use case is to identify the number of suitable patients in a database, fulfilling a specified pattern of inclusion- and exclusion criteria. The identification of these patients may be done in the research context by the researcher searching data bases in the non-care zone (e.g. research database, register) or by treating physician searching his EHR database in the care zone. Therefore the use case is divided into two sub cases.

Possible actors:	researcher treating physician (e.g. GP) database in the non-care zone EHR database in the care zone
Preconditions:	list of criteria (archetype) defining search pattern (inclusion- and exclusion criteria) data available in EHR or in selected databases in the non-care zone
Trigger:	search initiated by researcher
Post-condition:	number of cases fulfilling the search pattern
Open issues:	quality of data, suitability of archetype

5.2.1 Counts of patients in EHR

Search for eligibility in EHR database is triggered by the researcher and conducted by the treating physician.

Context:	research and medical care
Zone:	care zone
Actors:	researcher treating physician EHR
Process:	search is triggered by the researcher and conducted by the treating physician search within EHR displays number of cases fulfilling search pattern number of cases is transferred to researcher
Data privacy:	**assured**, treating physician is allowed to assess EHR database and there is normally no privacy issue with transferring counts

Potential problems: EHR data not suitable, not complete

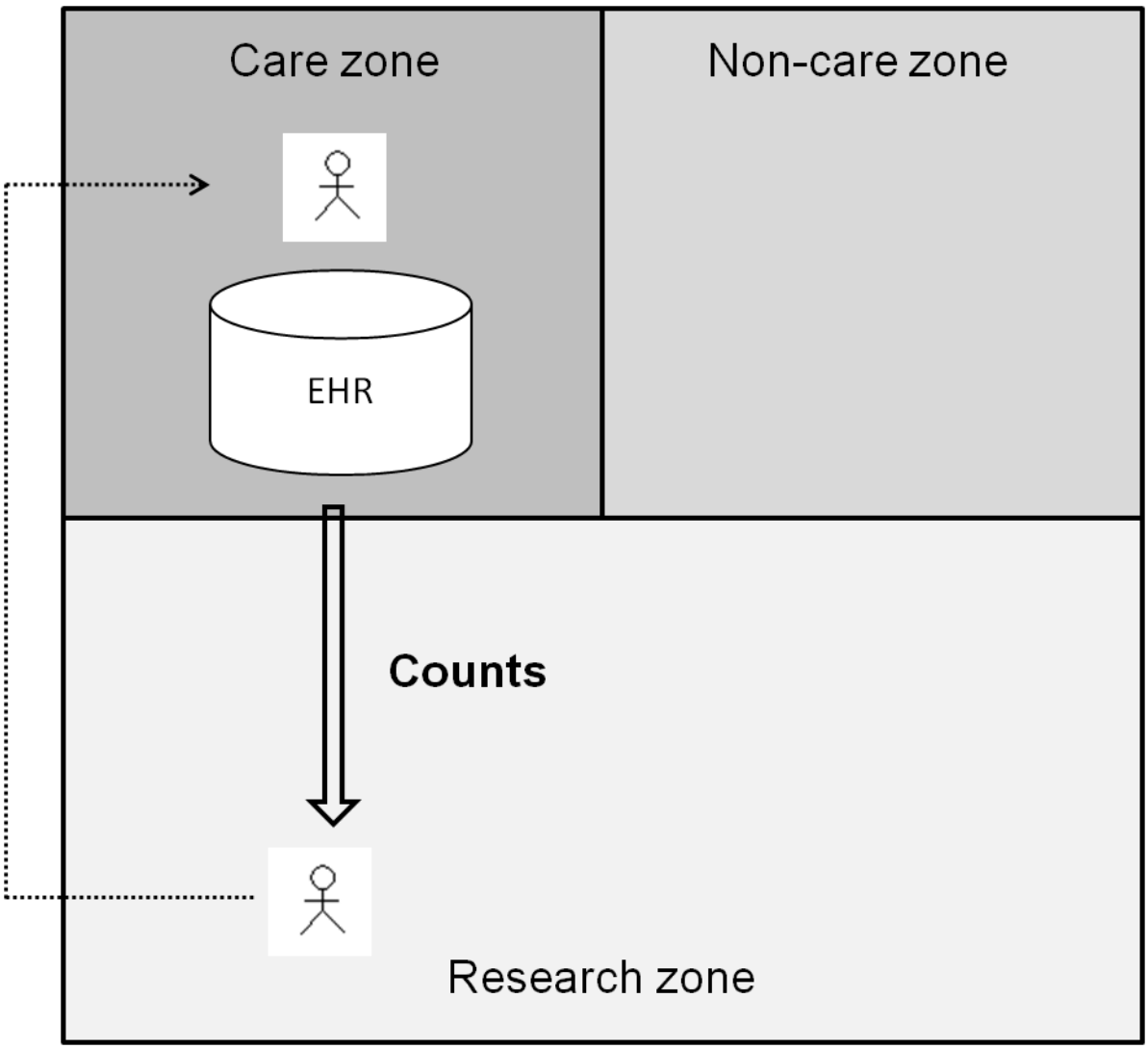


Figure 9: **Representation of use case "Counts of patients with defined criteria in EHR"**

5.2.2 Counts of patients in non-care database

Search for eligibility is triggered by the researcher and performed in the target database in the non-care zone.

Context: research

Zone: non-care zone

Actors:	researcher database in the non-care zone
Process:	search is triggered by the researcher and conducted in the target database with permission of the data controller search within the database displays number of cases fulfilling search pattern number of cases is transferred to researcher
Data privacy:	**assured**, only numbers are displayed to researcher
Potential problems:	research database not suitable, not complete

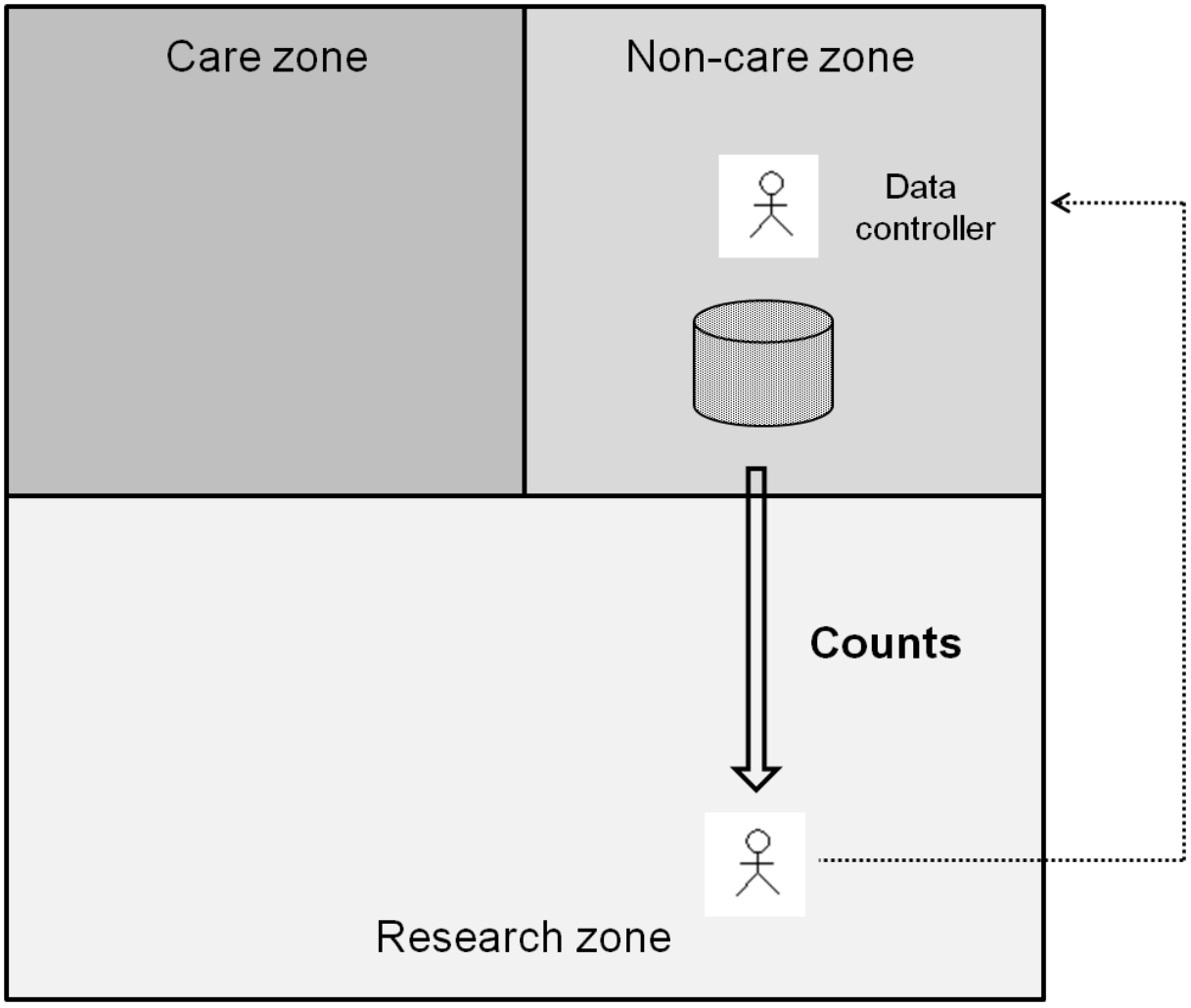


Figure 10: **Representation of use case "Counts of patients with defined criteria in non-care database"**

5.3 Use case: Find patients for clinical research

The target of this use case is to identify potential participants for clinical trials. The identification may be done purely in the care context by the treating physician (e.g. at the patient visit) or by searching data bases in the non-care zone (e.g. research database, register). The use case is divided into four subcases.

Possible actors:	researcher treating physician (e.g. GP) staff acting for the treating physician database in the non-care zone EHR patient
Preconditions:	list of criteria (inclusion-, exclusion criteria) for clinical trial available data availability known for EHR or selected databases in the non-care zone
Trigger:	search initiated by treating physician or trigger in EHR search initiated by researcher
Post-condition:	potential trial patient selected for evaluation of trial participation
Open issues:	linkage of non-care databases necessary to evaluate criteria

5.3.1 Identification of trial patients by treating physician within treatment context

Search for eligibility is triggered within the treatment context by the treating physician (e.g. during visit at the GP) within EHR (data triggered).

Context:	research and medical care
Zone:	care zone
Actors:	treating physician patient
Process:	eligibility search is triggered by the treating physician or triggered within EHR (e.g. data trigger) search within EHR identifies potential trial patient identified patient data transferred to treating physician treating physician invites patient for trial participation patient gives informed consent (or not)
Data privacy:	**assured**, treating physician is allowed to assess data of his patients and to invite his patients for trial participation
Potential problems:	EHR data not suitable biased patient selection for trials (selection bias)

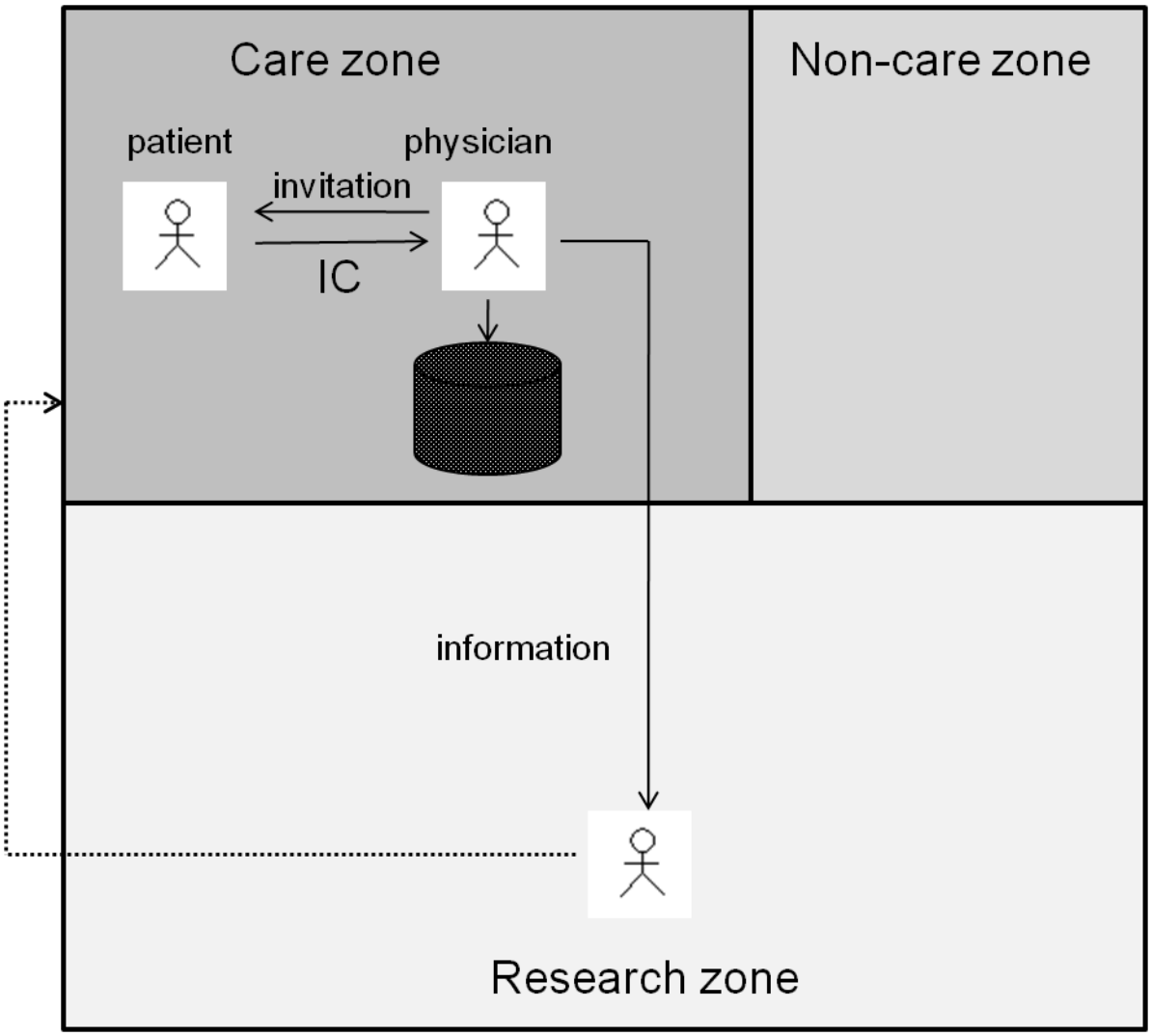


Figure 11: **Representation of use case "Find patients for clinical research by treating physician"**
(IC = informed Consent)

5.3.2 Identification of trial patients by extra staff acting for the treating physician

Search for eligibility is triggered by extra staff working for the treating physician. Extra staff may be, for example, a research nurse working for the treating physician.

Context:	research and medical care
Zone:	care zone
Actors:	treating physician patient extra staff acting for the treating physician
Process:	eligibility search is triggered by extra staff acting for the treating physician search within EHR identifies potential trial patient identified patient data transferred to treating physician treating physician invites patient for trial participation patient gives informed consent (or not)
Data privacy:	**assured**, if extra staff is acting for the treating physician, which will then function as a data processor. The treating physician will be the data controller. Treating physician is allowed to assess data of his patients and to invite his patients for trial participation
Potential problems:	EHR data not suitable biased patient selection for trials (selection bias)

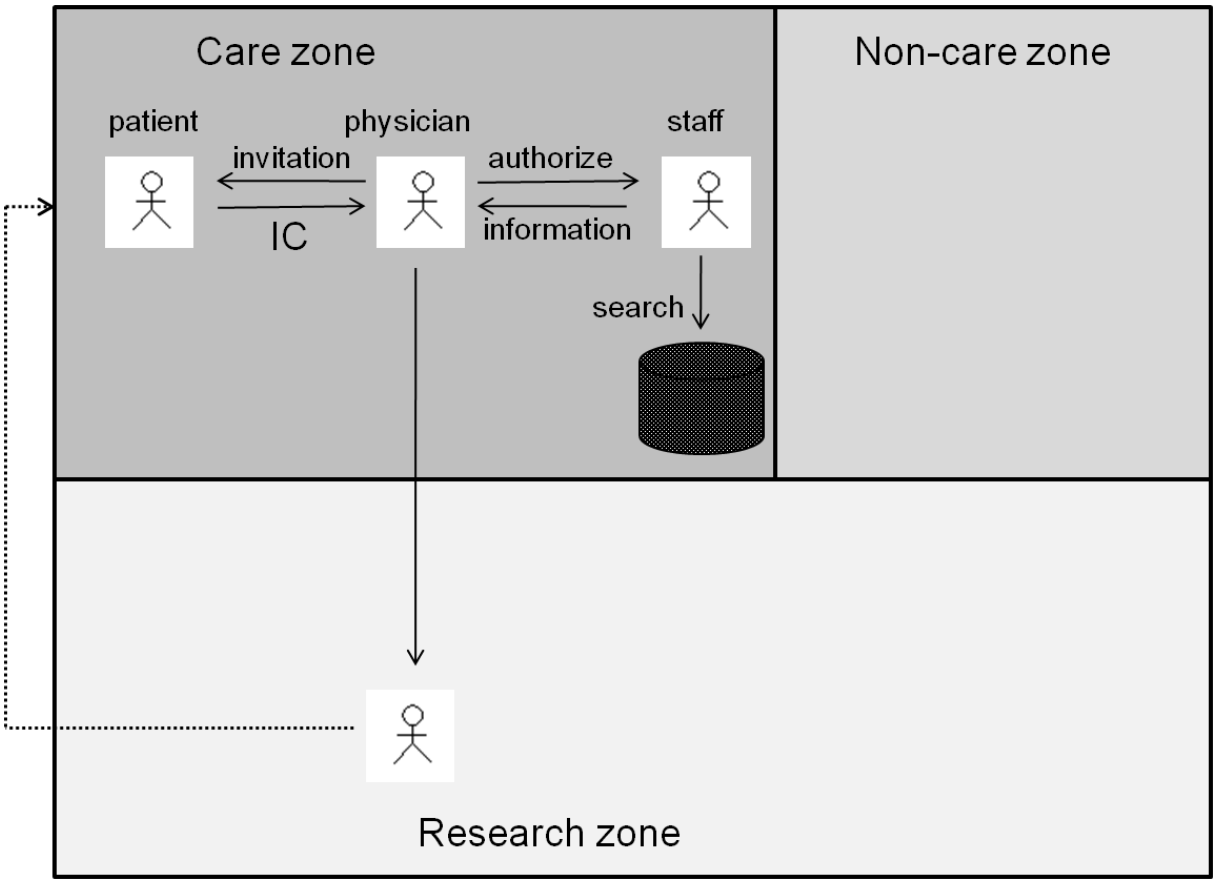


Figure 12: **Representation of use case "Find patients for clinical research by extra staff acting on behalf of the physician"**
(IC = informed Consent)

5.3.3 Identification of trial patients in non-care databases (e.g. research database, register)

Search for eligibility is triggered by researcher. Search for potential trial patients is performed in a non-care database with pseudonymous data. Pseudonym of identified patient is transferred to the treating physician. The treating physician is able to identify the patient and invites the patient for trial participation.

Context:	research and medical care
Zone:	non-care zone and care zone
Actors:	treating physician data controller of non-care database patient
Process:	eligibility search is triggered by 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)
Data privacy:	access to research database depends on **country regulations** (e.g. research exemption, opt-out; pseudonymous data treated as anonymous) treating physician is allowed to assess data of his patients and to invite his patients for trial participation
Potential problems:	data from research database not suitable biased patient selection for trials (selection bias)

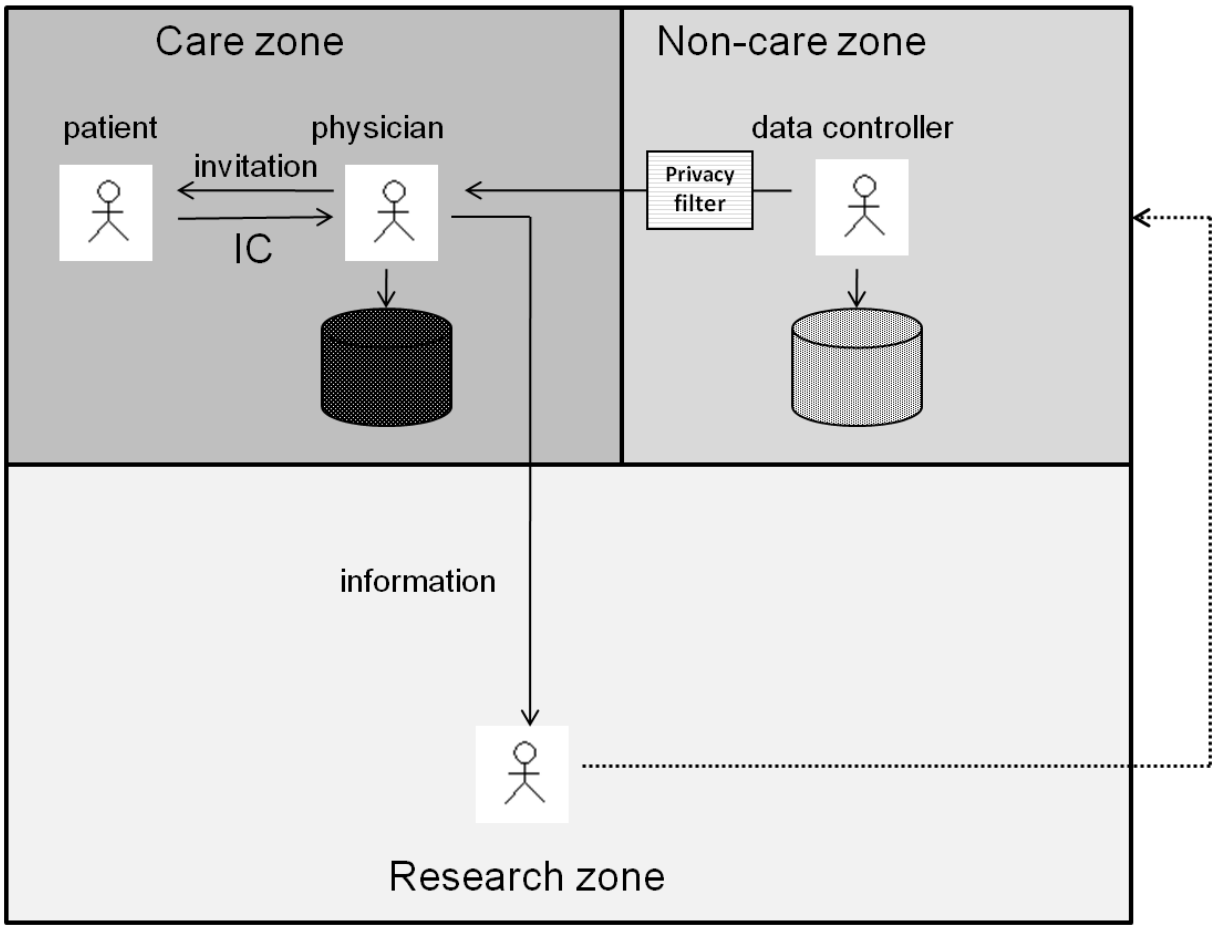


Figure 13: **Representation of use case "Find patients for clinical research by in non-care database"**
(IC = informed Consent)

5.3.4 Identification of trial patients in care or non-care database with the possibility to identify patients by the researcher

Search for eligibility is triggered by researcher. Search for potential trial patients is performed in a) care database (EHR) or b) non-care database (e.g. research database) with pseudonymous data. Pseudonym of identified patient is transferred to the researcher, who is able to identify the patient. The researcher invites the patient for trial participation.

Context:	research and medical care
Zone:	**researcher operates with personal identifiable data of the patient** (identification of potential trial patient in the care zone or via TTP in the non-care zone)
Actors:	researcher data controller of non-care database patient
Process:	eligibility search is triggered by researcher two scenarios: a) care database (EHR) allows search for researcher. Search within care database (EHR) identifies potential patient. Researcher is able to identify potential trial participant b) data controller of non-care database allows search (consistent with policy to use non-care database). Search within non-care database (e.g. research database) with pseudonymous data identifies potential patient researcher is able to identify potential trial participant researcher invites patient for trial participation patient gives informed consent (or not)
Data privacy:	**explicit consent** from patient necessary for researcher to search patients and to invite patients for trial participation
Potential problems:	explicit consent not available data from EHR/research database not suitable biased patient selection for trials (selection bias)

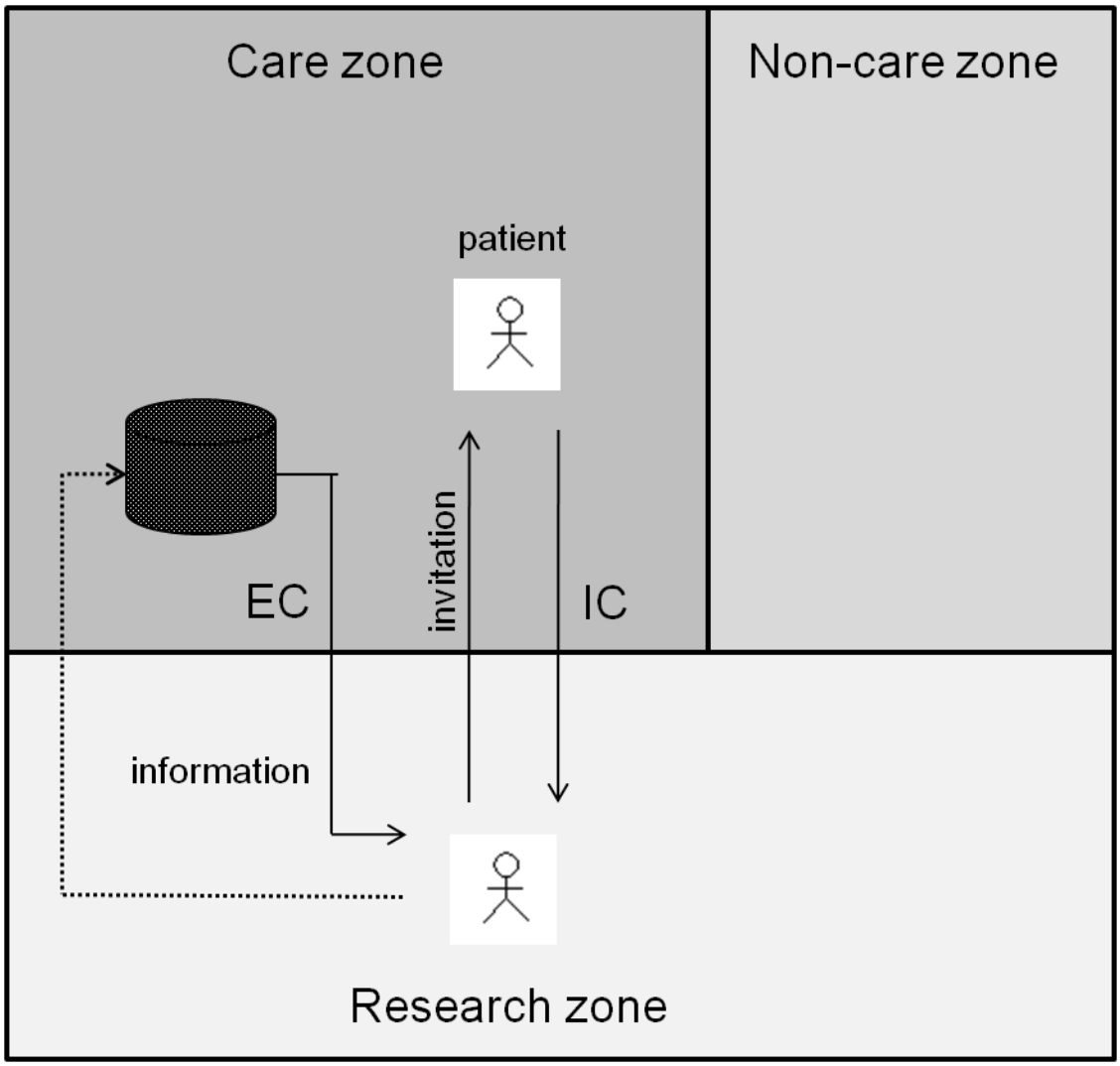


Figure 14a: **Representation of use case "Find patients for clinical research by researcher in care database"**
(IC = informed consent, EC = explicit consent, invitation = invitation for informed consent)

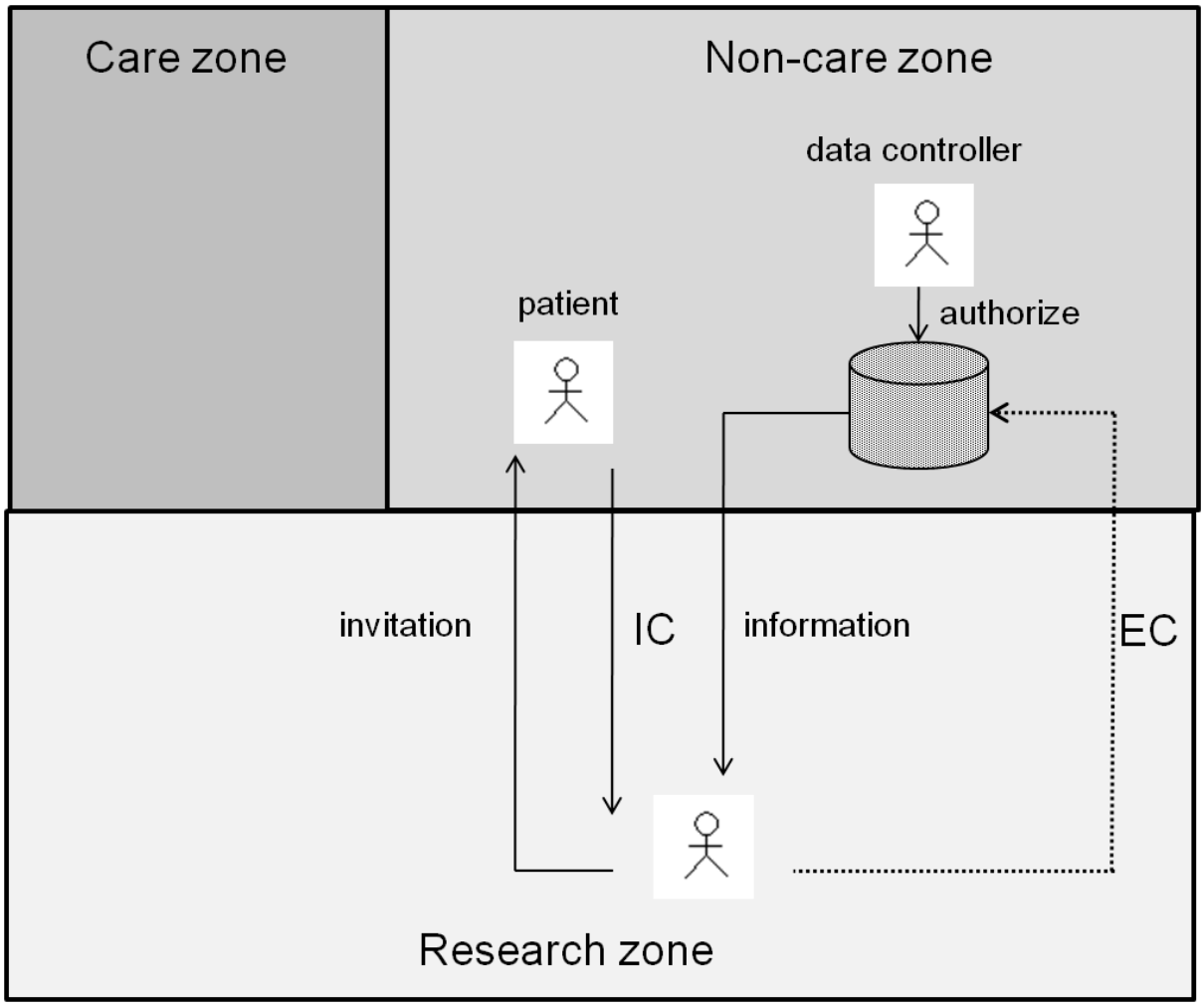


Figure 14b: **Representation of use case "Find patients for clinical research by researcher in non-care database"**
(IC = implicit consent, EC = explicit consent, invitation = invitation for informed consent)

5.4 Use case: Select patient for research question

The target of this use case is to identify and select potential cases for clinical research. This use case has overlap with use case 5.3 ("find patients for clinical research"). The target of the use case 5.3 is to identify potential participants for clinical trials. In this use case patients are also selected according to inclusion- and exclusion criteria but the identification is not related to invitation for participation in clinical trials but to provide data for research projects in the next step. The selection may be done in the care zone by the treating physician or by searching data bases in the non-care zone (e.g. research database, register). The use case is divided into two subcases.

Possible actors:	researcher treating physician (e.g. GP) database in the non-care zone EHR database
Preconditions:	list of criteria (archetype) for identifying patients data availability known for EHR or selected databases in the non-care zone
Trigger:	search initiated by treating physician or trigger in EHR search initiated by researcher in none-care database
Post-condition:	selected cases for clinical research
Open issues:	quality of data, suitability of archetype

5.4.1 Selection of research cases by treating physician in the care zone

Search for suitable cases is triggered within the treatment context by the treating physician (e.g. during visit at the GP).

Context:	research and medical care
Zone:	care zone
Actors:	treating physician researcher EHR database
Process:	search for suitable cases is triggered by the treating physician or triggered within EHR (e.g. data trigger) in response to request by researcher

search within EHR identifies potential cases for research
potential case privacy filtered
pseudonym is sent to researcher

Data privacy: **assured**, treating physician is allowed to assess data in EHR, identified cases are pseudonymised

Potential problems: EHR data may not be suitable or may be incomplete

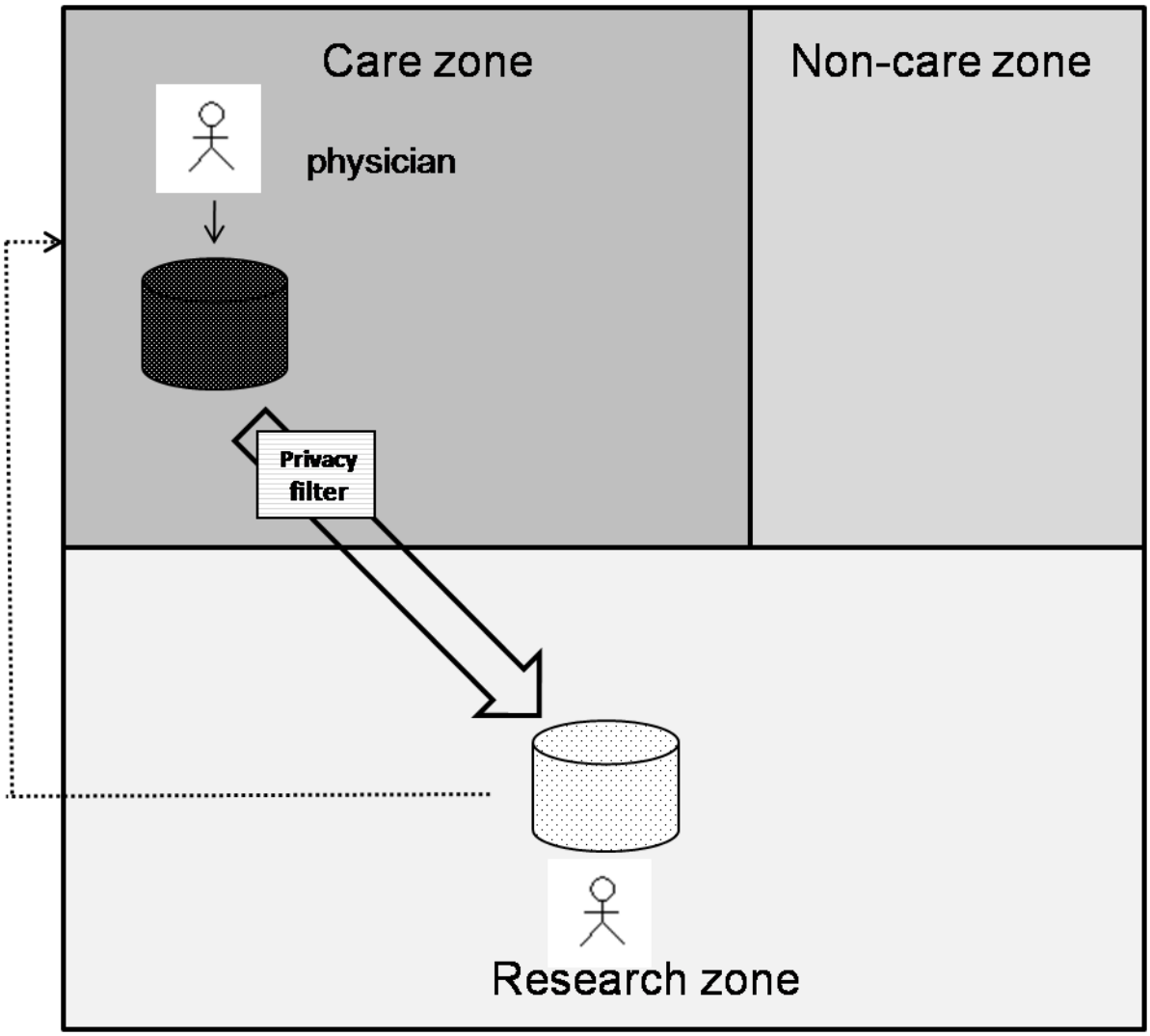


Figure 15: **Representation of use case "Select patients for research question in the care zone"**

5.4.2 Selection of research cases in none-care zone

Search for suitable cases is triggered by the researcher. Research is performed in a database in the non-care zone.

Context:	research and non-care zone
Zone:	none-care zone
Actors:	researcher non-care database
Process:	search for suitable cases is triggered by the researcher search within the non-care database (e.g. research database) identifies potential cases for research potential case privacy filtered pseudonym is sent to researcher
Data privacy:	**assured**, researcher receives filtered data, identified cases are pseudonymised
Potential problems:	non-care database not suitable for identifying patients for the research question at hand

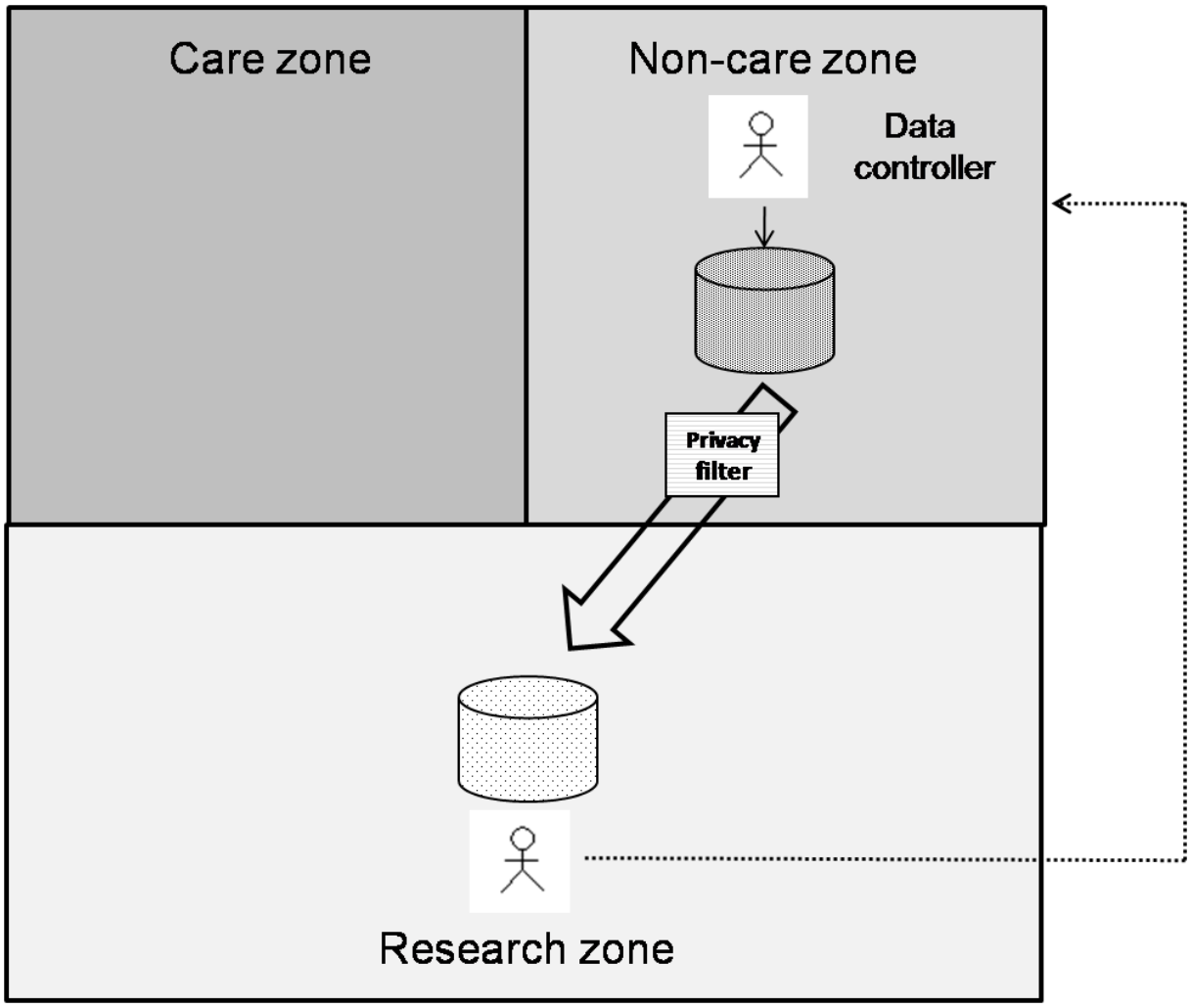


Figure 16: **Representation of use case "Select patients for research question in the non-care zone"**

5.5 Use case: Extract information of selected patients

The aim is to extract medical information from selected cases (see use case 5.4). The extraction may be done from an EHR database by the treating physician or by transfer from databases in the non-care zone (e.g. research database, register). The use case is divided into two subcases.

Possible actors:	researcher treating physician (e.g. GP) database in the non-care zone EHR database
Preconditions:	patient has been selected according to use case 5.4 data availability in EHR or selected databases in the non-care zone
Trigger:	extraction by treating physician or trigger in EHR in response to request by researcher extraction initiated in non-care database (e.g. research database) in response to request by researcher
Post-condition:	extracted data sets for research in database located at research zone
Open issues:	quality of data

5.5.1 Extraction of information of selected patients from care zone

Extraction of data sets of selected patients is triggered within the treatment context by the treating physician (e.g. during visit at the GP).

Context:	research and medical care
Zone:	care zone
Actors:	treating physician researcher EHR database
Process:	extraction of data sets of suitable patients is triggered by the treating physician or triggered within EHR (e.g. data trigger) data sets are pseudonymised (privacy filter) data sets are transferred into research zone data sets reside in research database with researcher
Data privacy:	**assured**, only treating physician is allowed to assess data in EHR, transferred data sets are fully anonymised or at least pseudonymised (or coded) anonymous
Potential problems:	EHR data may not be suitable or may be incomplete

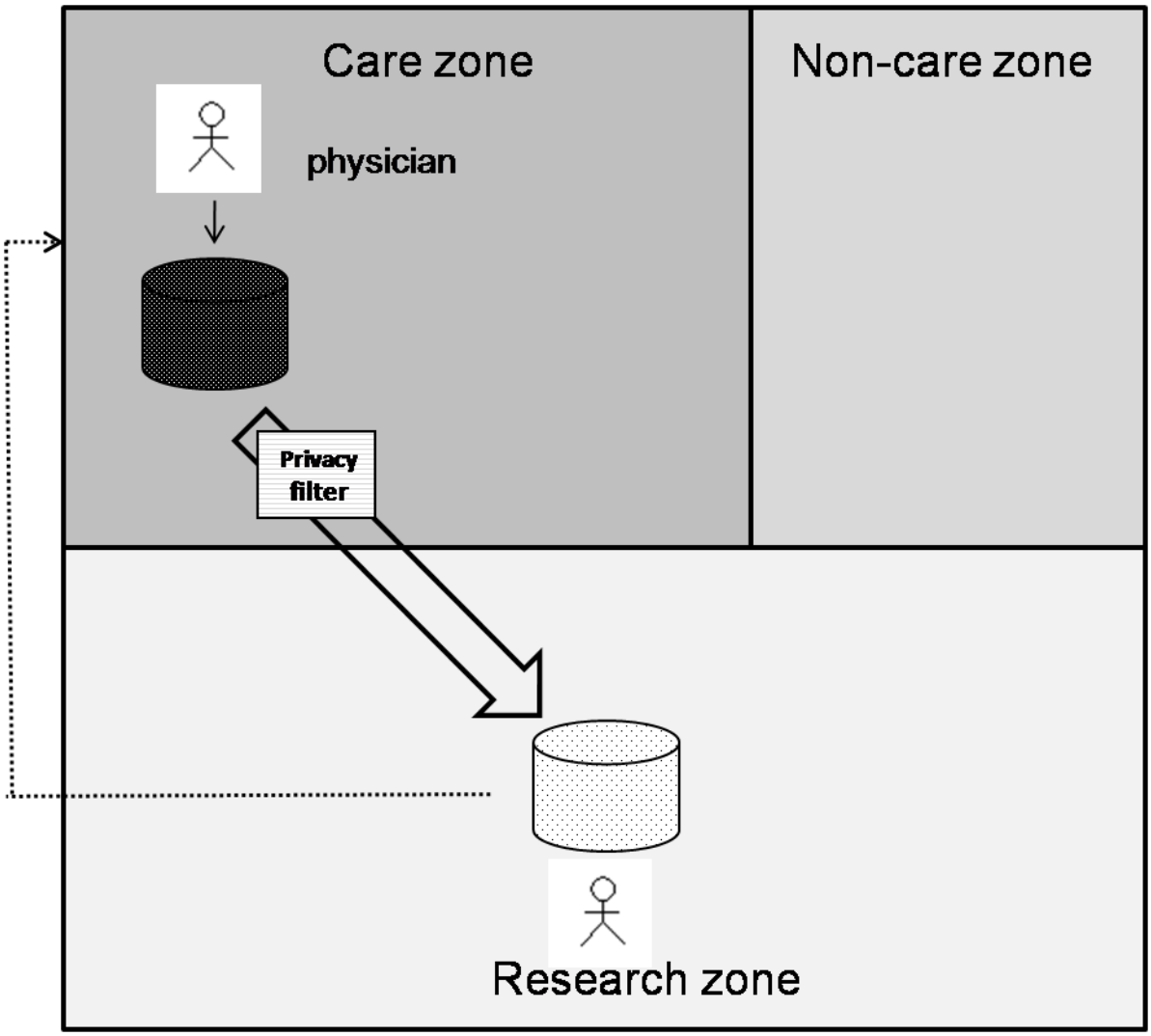


Figure 17: **Representation of use case "Extract information from selected patients in the care zone"**

5.5.2 Extraction of information of selected patients from none-care zone

Extraction of data sets of selected cases is triggered by the research database in request to the researcher

Context: research and non-care zone

Zone: none-care zone

Actors: researcher
database in non-care zone

Process: extraction of data sets of suitable cases is triggered in the non-care database
data sets are pseudonymised (privacy filter)
data sets are transferred into research zone
data sets reside in research database with researcher

Data privacy: **assured**, researcher receives transferred data sets that are fully anonymised or at least pseudonymised (or coded) anonymous

Potential problems: research database may not suitable

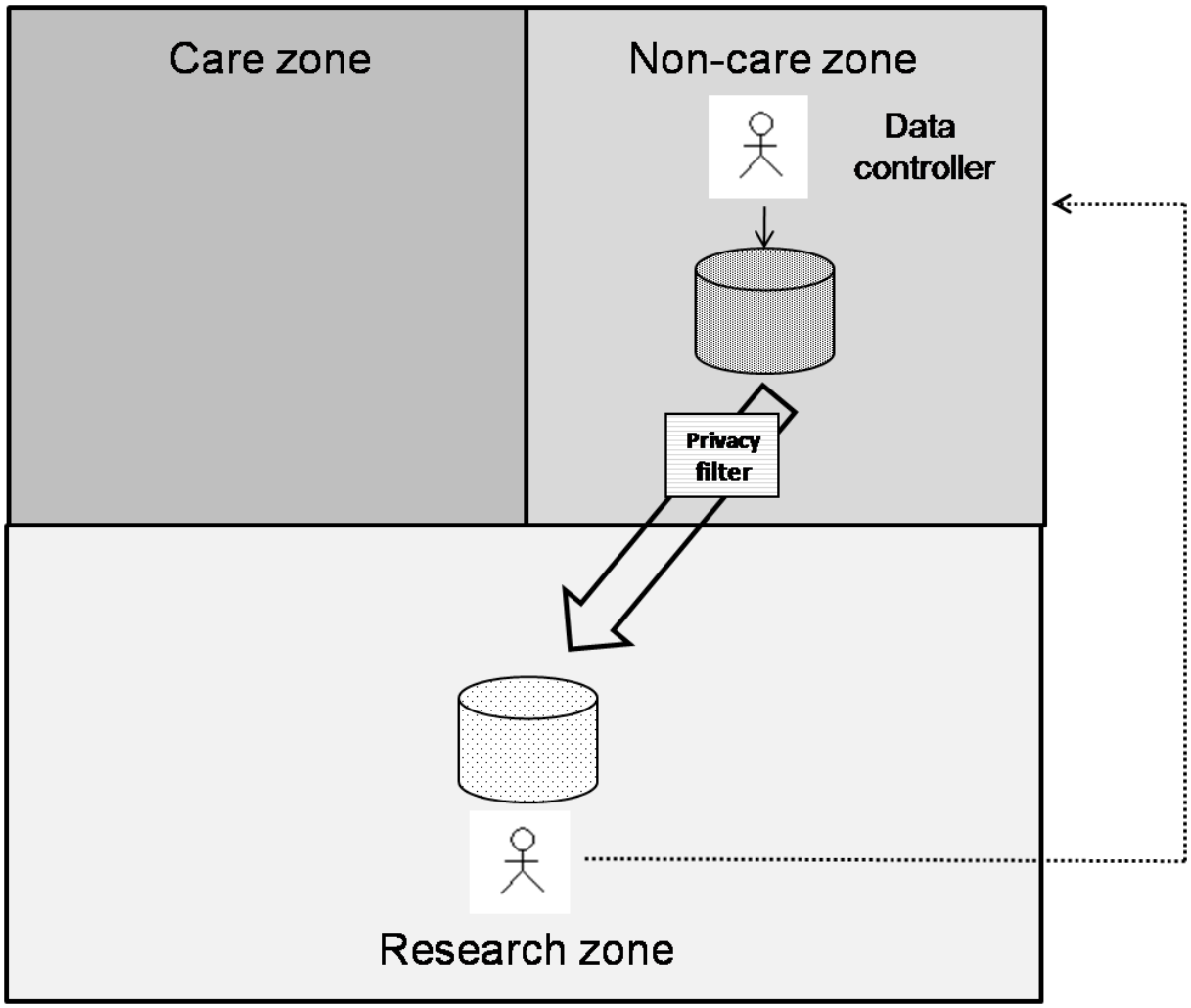


Figure 18: **Representation of use case "Extract data from selected patients in the non-care zone"**

5.6 Use case: linkage of data bases

The aim is to extract information about selected cases from linked databases. The linking may be done within a zone or subzone or between zones or subzones. The use case is divided into three subcases, according to involvement of databases in the care zone (e.g. EHR) and databases in the non-care zone (e.g. research database, register)..

Possible actors:	researcher treating physician (e.g. GP) database in the non-care zone database in the care zone
Preconditions:	data potentially linkable (on individual-level) linking possible according to rules and regulations and authorised
Trigger:	linking initiated by researcher
Post-condition:	linked database transferred to research zone with pseudonymised (or coded) anonymous data (if re-identification is necessary) or fully anonymised data
Open issues:	risk of re-identification in linked database different from risk of re-identification in the individual databases potential identifiable data in linked database requiring explicit consent

5.6.1 Linking between two databases in the none-care zone

Linkage is performed between databases in the non-care zone (e.g. research database, register).

Context:	research
Zone:	none-care zone
Actors:	researcher none-care databases
Process:	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 committee, data protection committee) preparation of linkage procedure in both none-care databases performance of linkage (e.g. use new pseudonym) linked database coded a second time (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
Data privacy:	**assured**, if data linkage is allowed according to rules and regulations and authorised. Linking is done with pseudonyms; transferred data sets are pseudonymised (or coded) anonymous data (if re-identification is necessary) or fully anonymised data.
Potential problems:	linking is not possible or incomplete

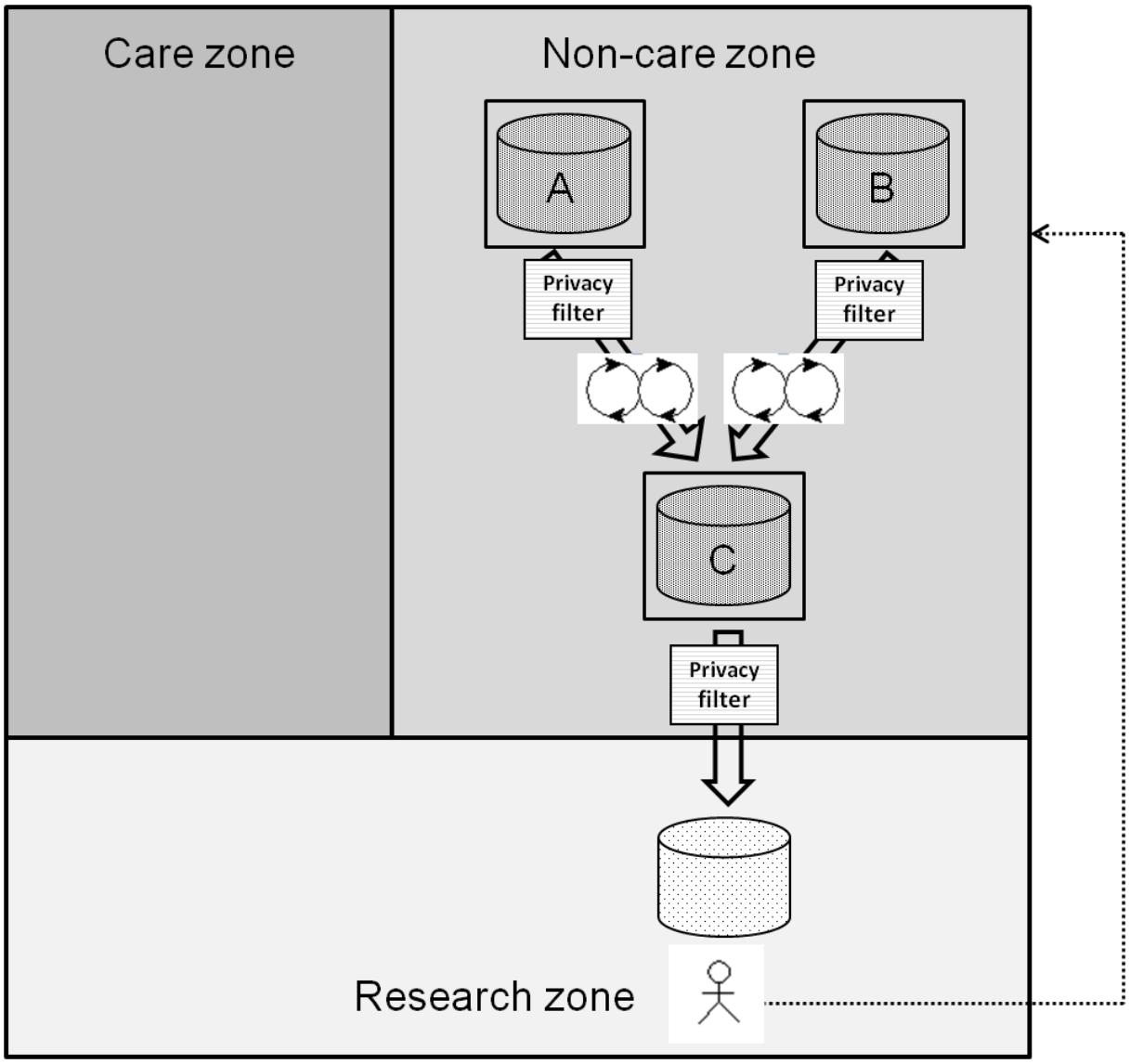


Figure 19: **Representation of use case "Linkage of databases in the non-care zone"** (researcher is provided with pseudonymised (or coded) anonymous data (if re-identification is necessary) or fully anonymised data according to the research question and authorisation)

5.6.2 Linking between two databases in the care zone

Linkage is performed between two databases in two different subzones of the care zone (e.g. EHR database, HIS database, hospital data warehouse).

Context:	medical care
Zone:	care zone and non-care zone
Actors:	treating physician researcher databases in the care zone
Process:	trigger of linkage by researcher checking the permissibility of linkage of the care databases by the data controllers of these databases authorisation of data linkage (e.g. etics committee, data protection committee) preparation of linkage procedure in both care databases performance of linkage in non-care zone (e.g. using new pseudonym) linked database coded a second time (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
Data privacy:	**assured**, if data linkage is allowed according to rules and regulations and authorised. Linking is done with pseudonyms in the non-care zone; transferred data sets are pseudonymised (or coded) anonymous data (if re-identification is necessary) or fully anonymised data.
Potential problems:	linking is not possible or incomplete

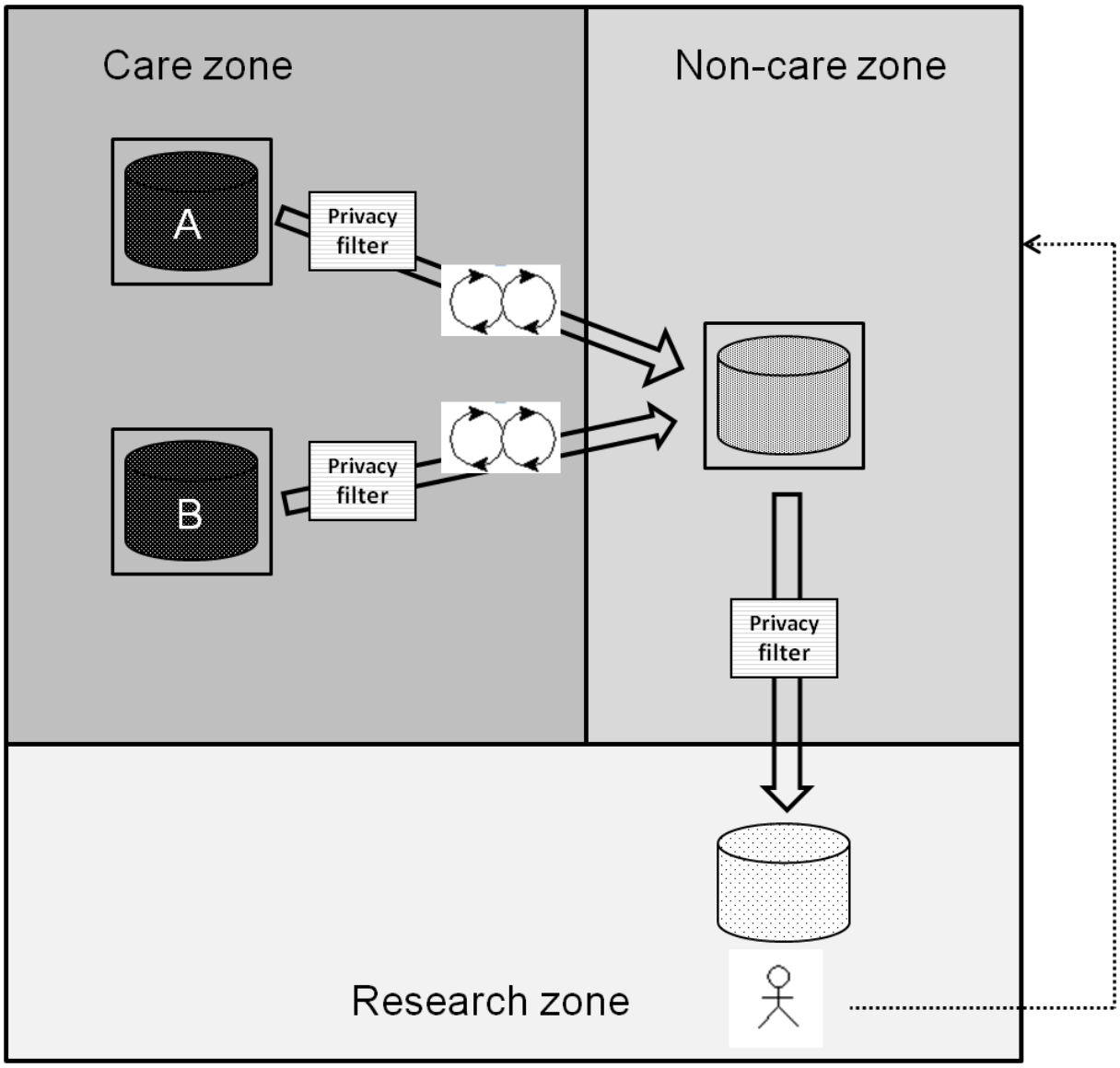


Figure 20: **Representation of use case "Linkage of databases in the care zone"** (researcher is provided with pseudonymised (or coded) anonymous data (if re-identification is necessary) or fully anonymised data according to the research question and authorisation)

5.6.3 Linking between database in care zone with database in none-care zone

A link is generated between the database in the care zone (e.g. EHR database, HIS database, hospital data warehouse) with a database in the non-care zone (e.g. research database, register).

Context: research and medical care

Zone: care zone and none-care zone

Actors:	treating physician researcher care database none-care database
Process:	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 a second time (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
Data privacy:	**assured**, if data linkage is allowed according to rules and regulations and authorised. Linking is done with pseudonyms in the non-care zone; transferred data sets are pseudonymised (or coded) anonymous data (if re-identification is necessary) or fully anonymised data.
Potential problems:	linking is not possible or incomplete

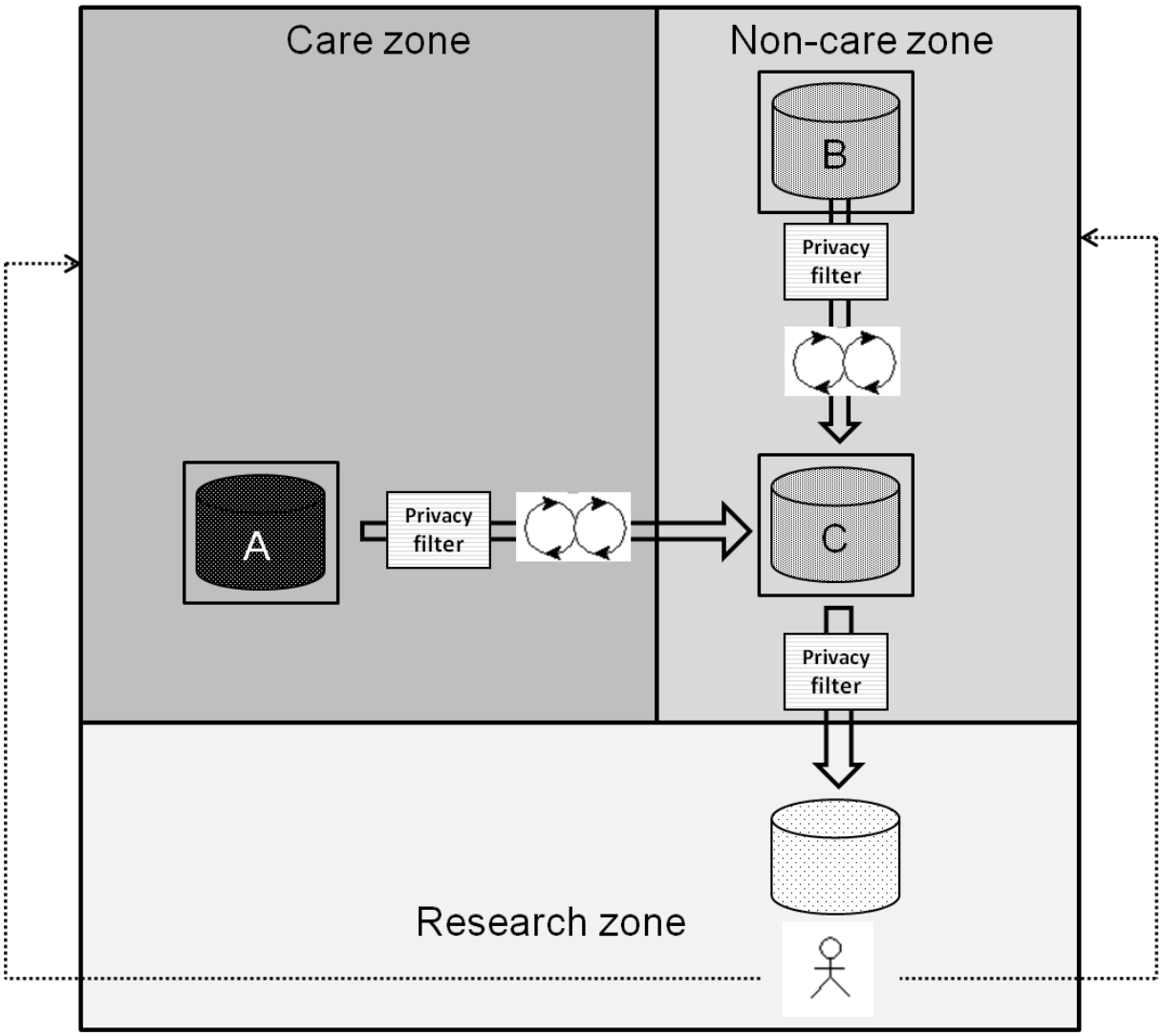


Figure 21: **Representation of use case "Linkage of database in the care zone with database in the non-care zone"**
(researcher is provided with pseudonymised (or coded) anonymous data (if re-identification is necessary) or fully anonymised data according to the research question and authorisation)

6. Concluding remarks

6.1 Discussion

Existing confidentiality and data privacy frameworks proved to be largely limited by scope, target and application areas. Often they are dedicated to specific countries (only US), specific diseases (e.g. cancer), specific situations (e.g. only research context with informed consent by the patient) or to specific data sources (e.g. only use of secondary data). In some countries, like the US and Germany, specific solutions have been developed, however these are not extendable to all countries (Safran, 2007; Office of the National Coordinator for Health Information Technology, 2008). In addition, the situation for TRANSFoRm is special. The framework should be applicable across medical care (EHR) and clinical research, should take into consideration different data sources (primary and secondary data) from different fields (phenotype, genotype data), should deal with different diseases (diabetes, GORD, cancer) and should be applicable across countries and regions in Europe.

Being aware of the divergent implementations and applications of laws and regulations for data protection and confidentiality in Europe, the primary objective within this work package of TRANSFoRm was to analyse and structure the given heterogeneity to come to an approach working out similarities between different scenarios (e.g. between countries, data sources, context), and thus providing practical help in a highly complex environment. The idea behind was to build upon existing means of confidentiality and data privacy measures (e.g. PET, TTP, data transfer agreements) and to incorporate it into the model. In order to be of practical help, the framework should be not too simple (because then it would be of no benefit) but also not too complex in order to allow at least some relevant structuring. The primary motivation for the zone model was to structure according to categories of sensitivity. Similar to the caBIG-approach three categories of sensitivity were defined: directly identifiable data, pseudonymised data and anonymous data (caBIG, 2011). These types of data can be largely (but not exactly) allocated to different zones, the medical care zone, the non-care zone (research database, registers, studies) and the research zone. Data bases from the care zone mean always directly identifiable data, data bases from the non-care zone usually indirectly identifiable data (whether pseudonymised or not) and data from the research zone anonymous data (whether pseudonymised or not).

Starting with zones on this abstract level allows a first but crude structuring with practical consequences. In the care zone the care provider is responsible for the data collected from the patient and is the ultimate data controller. For a specific research project, the researcher in the research zone needs data for actual research on a patient level without a need to identify patients. For TRANSFoRM it was felt that the main zones are insufficient to be of practical value. Therefore the concept of subzones was introduced, which contain data that are comparable, can be used for the same or a similar purpose and with similar applicable rules and regulations for their use. With this concept a second layer for structuring according to similarities was introduced with the possibility to bring together databases from different countries/regions, different sources and application areas,

provided similar rules and regulations have to be applied. This concept was defined on a rather general level and only a few examples were given for illustration. Within TRANSFoRm it has to be sorted out by practical examples to what extent this concept can be applied to data sources used in the project. If scenarios with similar conditions could be treated in a similar way, significant improvement would have been achieved.

6.2 The next steps

The last remark of the previous section brings us to the follow-up of this Framework. In a way this approach of this framework is not unique. In the literature, these kind of frameworks are usually represented by some kind of data flow model, describing actors and roles, institutions involved, data sources used and actions/functions performed (e.g. Forgo (ACGT) 2007; Kaira, 2005; ISO, 2008). The Unified Modeling Language is a standardized general-purpose modelling language in field of object-oriented software engineering. UML includes a set of graphic notation techniques to create visual models of object-oriented software-intensive systems. UML provides many tools (e.g. use case diagrams, activity diagrams) that can be used for modelling of a framework, for example used for data flow between entities in the workflow model in ISO/TS 25237. We did not use UML for modeling but instead a simple data flow model derived from structured analysis (Yourdon, 1989). In this approach the system is viewed from the perspective of data flowing through it and this flow is graphically represented. The data transfer between zones/subzones is described by flow diagrams. To make the flow possible, specific functions/processes are applied to the data. These functions are called privacy filters (PET), which allow transformation from personal identifiable data via pseudonymous data to anonymous data and thus may move data from one zone to another. Whether the formal description of the framework has to be amended or expanded has to be discussed when first experiences with concrete applications are available.

The first application of the framework also will also be a test of the principles. The introductory chapter described how privacy versus health research can be considered a highly contested area. Will the balance we have set in our principles hold? To what extent will it still be necessary to use national exemptions to the consent principle within the zone model and with the aid of TRANSFoRm software? In this respect we conclude with forwarding a new idea for the software development, already briefly mentioned in the text. That is not to only concentrate on how data will arrive in the research zone coded anonymously, but also how software (by logging, audit trails of data etc.) can guarantee that researchers can substantiate their claim that they do not re-identify a patient by indirectly identifiable data and which still gives sufficient flexibility for researchers to allow their complex statistical analyses of those data .

In addition to the more general principles, specific aspects have to be evaluated and worked out (Arning, 2009). This includes terminology and semantics, such as the interpretation of anonymous data. Though the Working Party 2007/4 (Working Party 2007) document already gives much guidance, there will always remain a ‘grey area’ between anonymous and indirectly identifiable data Contrary to the HIPAA legislation in the US, which works with fixed datasets to determine the level

of anonymity and possible re-identification, Directive 95/46/EC uses a general term as described earlier, leading to a more flexible but also more ambiguous and context driven approach. Technical specifications, such as the ISO Technical Specification on Pseudonymisation (ISO, 2008), may support implementation of a privacy framework for a concrete project. Guidance derived from research projects may be used to support data linkage of heterogeneous data sets, preserving anonymity by preventing re-identification (Lyons, 2009). New methodological approaches may suggest possibilities to perform a pooled analysis of individual-level data without sharing the data (Wolfson, 2010). Proposals for evolutionary transition from traditional EHRs to a tagged data element model by means of a universal exchange language may support data sharing in heterogeneous environments in the future (Executive Office of the President, 2010). These and other specific aspects need to be considered for implementation of a confidentiality and privacy framework in a concrete application.

7. References

Academy of Medical Sciences (AMS): A new pathway for the regulation and governance of health research, January 2011-03-02. http://www.acmedsci.ac.uk/download.php?file=/images/project/129468115924.pdf. Last visited: 31.3.2011

Arning M., N. Forgó and T. Krügel: Data protection in grid-based multicentric clinical trials: killjoy or confidence-building measure? Phil. Trans. R. Soc. A 13 2009; 367: 2729-2739

Article 29 Data Protection Working Party: Opinion 4/2007 on the concept of personal data. 01248/07/EN, WP136, adopted on 20th June http://ec.europa.eu/justice_home/fsj/privacy/index_en.htm. Last visited: 31 March 2011

Beyleveld D., and D. Townend: When is personal data rendered anonymous? Interpreting recital 26 of Directive 95/46/EC. Medical law international, 2004; 6: 73-86.

Beyleveld D. et al (ed): Research Ethics Committees, Data Protection and Medical Research in European Countries. Aldershot/Burlington, Ashgate 2005

Beyleveld D.et al (ed): Implementation of the Data Protection Directive in Relation to Medical Research in Europe. Aldershot/Burlington, Ashgate 2004

cancer Biomedical Informatics Grid (caBIG): Data Sharing and Security Framework. https://cabig-kc.nci.nih.gov/DSIC/KC/index.php/Data_Sharing_and_Security_Framework. Last visited: 31.3.2011

Council for International Organizations of Medical Sciences (CIOMS), International Ethical Guidelines for Epidemiological Studies, Geneva: CIOMS 2009

Declaration of Helsinki (World Medical Association) Ethical Principles for Medical Research Involving Human Subjects. 59th WMA General Assembly, Seoul, October 2008

EU Directive 2001/20/EC of the European Parliament and the council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use. Official Journal of the European Communities, No. L121/34-121/43

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. Official Journal of the European Communities, 1999; No. L281/31-281/39

Executive office of the President, President's Council of Advisors on Science and Technology: Report of the president realizing the full potential of health information technology to improve healthcare for Americans: the path forward, December 2010. http://www.whitehouse.gov/administration/eop/ostp/pcast. Last visited: 31.3.2011

Follesdal F., R.A. Wessel, J. Wouters (ed): Multilevel regulation and the EU: the interplay between global, European and national normative processes, Leiden/Boston: Martinus Nijhoff Publishers 2008

Forgó N. (Ed.): The ACGT ethical and legal Requirements. ACGT Deliverable 10.2., 13.03.2007. http://eu-acgt.org/uploads/media/ACGT_D10.2_IRI_Final_01.pdf. Last visited: 31.3.2011

General Practice Research Database (GPRD). http://www.gprd.com. Last visited: 31.3.2011

Kaira D., R. Gertz, P. Singleton, H.M. Inskip. BMJ 2006; 33, 196-198

Kalra D., 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 2005; 44:193-197

Islam S., H. Mouratidis, S. Wagner: Towards a framework to elicite and manage security and privacy requirements from laws and regulations. Lecture notes in Computer Science, 2010, Volume 6182/2010, 255-261

ISO Technical Specification 25237: Health informatics – pseudonymisation, 2008 (E)

Lyons R.A., K.H. Jones, G. John, C.J. Brooks, J.P. Verplancke, D.V. Ford, G. Brown, K. Leake: The SAIL databank: linking multiple health and social care datasets. BMC Med Inform Decis Mak. 2009;9:3

Netherlands Information Network of General Practice (LINH). http://www.nivel.nl/oc2/page.asp?PageID=8599&path=/Startpunt/NIVEL%20international/R

esearch/Data%20bases%20and%20information%20systems/National%20Information%20Network%20of%20GPs%20(LINH). Last visited: 31.3.2011

Netherlands Institute for Health Services Research (NIVEL). http://www.nivel.nl. Last visited: 31.3.2011

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 http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__privacy___security_framework/1173. Last visited: 31.3.2011

Organisation for Economic Co-operation and Development (OECD): Guidelines for human biobanks and genetic research databases 2009, http://www.oecd.org/document/12/0,3746,en_2649_34537_40302092_1_1_1_1,00.html. Last visited 31.3.2011

Safran C., 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. 2007; 14: 1-9

Unidied Modelling Language (UML). http://www.uml.org. Last visited: 31.3.2011

Universal Declaration on Bioethics and Human Rights adopted by UNESCO on 19th October 2005

van Veen E.B.: Obstacles to European research projects with data and tissue: solutions and further challenges. Eur J Cancer. 2008 ; 44: 1438-50

van Veen E.B., P.H. Riegman, W.N. Dinjens, K.H.Lam, M.H. Oomen, A. Spatz, R. Mager, C. Ratcliffe, K. Knox, D. Kerr, B. van Damme, M. van de Vijver, H. van Boven, M.M. Morente, S. Alonso, D. Kerjaschki, J. Pammer, J.A. Lopez-Guerrero, A. Llombart Bosch, A. Carbone, A. Gloghini, I. Teodorovic, M. Isabelle, A. Passioukov, S. Lejeune, P. Therasse, J.W. Oosterhuis: TuBaFrost 3: Regulatory and ethical issues on the exchange of residual tissue for research across Europe. Eur J Cancer. 2006; 42: 2914-23

Verschuuren M., G. Badeyan, J. Carnicero, M. Gissler, R.P. Asciak, L. Sakkeus, M. Sternbeck, W. Devillé, For the 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. European Journal of Public Health 2008; 18: 550-551

World Health Organization (WHO): Human gentic databases: towards a global ethical framework. http://www.who.int/ethics/topics/hgdb/en/print.html. Last visited 31.3.2011

Wolfson M., S.E. Wallace, N. Masca, G. Rowe, N.A. Sheehan, V. Ferretti, P. LaFlamme, M.D. Tobin, J. MacLeod, J. Little, I. Fortier, B.M. Knoppers, P.R. Burton: DataSHIELD: resolving a conflict in contemporary bioscience – performing a pooled analysis of individual-level data without sharing the data. International Journal Epidemiology 2010; 39: 1372-1382

Yourdon E.: Modern Structured Analysis, Yourdon Press Computing Series, 1989

8. Abbreviations

CRF	Case Report Form
DB	Data Base
CLEF	Clinical E-Science Framework
DTA	Data Transfer Agreement
EC	Explicit Consent
eCRF	electronic Case Report Form
EHCR	Electronic Health Care Record
EHR	Electronic Health Record
GORD	GastroOesophageal Reflux Disease
GP	General Practitioner
GPRD	General Practitioner Research Database
HIPAA	Health Insurance Portability and Accountability Act
HIS	Hospital Inormation System
IC	Informed Consent
ISO	International Standards Organisation
LINH	Netherlands Information Network of General Practice
MCR	Medical Research Council
NIVEL	Netherlands Institute for Health Services Research
PET	Privacy Enhancing Technique
QoL	Quality of Life
RCT	Randomised Clinical Trial
TRANSFoRm	Translational Research and patient Safety in EuRope
TTP	Trusted Third Party
WP	Working Package
WT	Working Task