COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE XXX (2010) XXX–XXX

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Strategies for health data exchange for secondary, cross-institutional clinical research

Bernice S. Elger[a], *Jimison Iavindrasana*[a], *Luigi Lo Iacono*[b,*], *Henning Müller*[a,c], *Nicolas Roduit*[a], *Paul Summers*[d], *Jessica Wright*[e]

[a] *University of Geneva, Switzerland*
[b] *NEC Laboratories Europe, Germany*
[c] *University of Applied Sciences Western Switzerland, Sierre, Switzerland*
[d] *University of Oxford, England, United Kingdom*
[e] *University of Sheffield, England, United Kingdom*

ARTICLE INFO

Article history:
Received 19 August 2008
Received in revised form
12 November 2009
Accepted 2 December 2009

Keywords:
Health data
Secondary use
Clinical research
Consent
Pseudonymisation
Reasonable anonymisation

ABSTRACT

Secondary use of health data has a vital role in improving and advancing medical knowledge. While digital health records offer scope for facilitating the flow of data to secondary uses, it remains essential that steps are taken to respect wishes of the patient regarding secondary usage, and to ensure the privacy of the patient during secondary use scenarios. Consent, together with depersonalisation and its related concepts of anonymisation, pseudonymisation, and data minimisation are key methods used to provide this protection. This paper gives an overview of technical, practical, legal, and ethical aspects of secondary data use and discusses their implementation in the multi-institutional @neurIST research project.

1. Introduction

The on-going relationship between a patient and their medical practitioners gives rise to a health record containing detailed information of medical relevance that (a) may be distributed across clinical centres, (b) involves a range of specialities and informatics systems within individual centres, (c) expands over time, and (d) is of a personal nature to the patient. Personal health data contains in the first instance information relating to the current and historical health, medical conditions, and medical tests of its subject. Secondary uses for such information arise when, for example, the records are summarized across patients in a department or institution for performance audits, or on a larger scale for epidemiological study of referrals for a given condition [1]. The low prevalence of most health conditions calls for the integration of records from multiple clinical centres in order to ensure adequate numbers of patient records are available to allow statistically relevant results to be obtained. Making more extensive use of clinical records opens the possibility to study patterns of symptoms, treatment, and outcomes, but often requires the

* *Corresponding author at*: NEC Laboratories Europe, IT Research Division, NEC Europe Ltd., Rathaussallee 10, D-53757 Sankt Augustin, Germany.
E-mail address: lo_iacono@it.neclab.eu (L. Lo Iacono).
0169-2607/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2009.12.001

data be accessible as individual rather than aggregate values. Our aim is to provide a cohesive overview of the issues encountered in creating a system that facilitates secondary use of data originating from multiple clinical centres by a consortium of research users to serve as a point of reference for other workers in the field.

Traditionally, clinical information resides within the originating structure, with extracts transmitted as necessary between clinicians in the course of patient care. In general however, unified centralized repositories of, or even means to access the various health record components within large health networks do not exist. This situation is evolving as patient information is increasingly stored and communicated as computerized records referred to as an Electronic Health Records (EHRs). Within large-scale health networks, in particular national health systems, clinical need has motivated increasing integration of health information systems and interconnection between branches of the health services. Despite this progress in the clinical domain, data entry for research generally remains parallel to that of the clinical record, requiring labour-intensive transcription of clinical notes prior to centralized data entry and compilation into a monolithic database. As evidenced in epidemiological studies based on the PHARMO database, the same infrastructure for use of EHRs could facilitate the availability of large quantities of clinical data for biomedical research where the obstacles to such use need to be overcome.

In the primary use of an EHR for the treatment of a single patient, the patient's identity is necessary and protected by medical secrecy. While the patient's identity is not normally relevant for secondary uses, the data remains personal due to its specificity to the individual, and therefore the patients' wishes regarding the use of the data must be respected. An appropriate procedure of informed consent is therefore a key requirement in the current research environment. The passage of data out of its medical setting for use in the collaborative efforts of multiple institutions, both medical and non-medical in turn raises an issue of privacy protection. The response to this issue necessitates a data protection strategy that can be clearly stated, and put into practice. Moreover, not only the handling of the medical data, but also the many activities that may form a research study must be performed in a manner consistent with the policies of privacy and data protection. Unfortunately, considerable mixing of terminology and concepts has taken place in the many documents intended to guide policy in this area. This has created a confusing environment for the researcher interested in establishing or participating in multi-centre research studies making use of primary healthcare data.

In the course of the @neurIST project (described below), a novel approach has been advanced for the federation of distributed medical records that is able to follow their evolution over time as a platform for secondary research. The project also involves an active research program that makes use of the data federation system as well as practical activities such as blood sample collection and analysis, image processing and database research. The process of implementing the systems necessary for this project and for supporting the secondary research activities has required us to face the above issues of consent and data privacy, as well as number of subsidiary and ancillary issues relating to the secondary use of clinical data in the research domain. We describe the context of ethical and legal requirements applicable to secondary research using clinical data, some of the strategies adopted for ensuring these requirements are met, and the specific solutions adopted in the course of the @neurIST project.

1.1. *The @neurIST project*

The @neurIST project is an EU-funded project that has proposed a strategy of federating data sources in clinical institutions for use in research and in advancing clinical practice [2]. The project involves seven clinical centres in five countries (England, Hungary, Spain, Switzerland, and The Netherlands), along with a further 25 institutions contributing to the technical development of the infrastructure for data federation, software systems for using the data, or making use of the data and samples collected (full list of partners available from http://www.aneurist.org/).

The specific clinical domain in which the @neurIST data strategy is being developed and demonstrated is that of intracranial aneurysms. This condition, in which segments of one or more arteries of the brain become abnormally dilated affects between 1% and 6% of the European population, and in a given year, about one percent of the affected patients can expect to experience a rupture of this altered vessel wall, leading to bleeding into the brain [3–6]. Although treatment and care are typically concentrated in larger hospitals, no one hospital can capture sufficient numbers of cases to carry out studies having statistical significance. Further, steady progress in treatment options makes it important and yet more difficult to quickly identify differences in treatment outcome while minimising the costs and overheads of the investigations. Lastly, diagnosis, treatment and post-treatment care are often managed by clinicians in different institutions, each holding a part of the patient's clinical record that may or may not be transmitted in full to the others. This necessitates repeated data entry not only for research, but also routine clinical practice. Thus the benefits to this clinical community of cooperation at national and trans-national scales that can link across data sources is strong, and can be translated to many other diseases.

Accompanying the development effort, and acting as a test-bed to demonstrate the data federation system is the @neurIST study, centring on a genetics study involving over 800 patients and 400 controls, along with investigations in transcriptomics, computational fluid dynamics, and data-mining. For the @neurIST study, data is maintained in distinct repositories, be they the original clinical data stores or mirrors holding extracts of the data in a DMZ (De-Militarized Zone), each held by the originating institution. The data are accessible through an innovative IT infrastructure [7] based on Grid and SOA (Service-Oriented Architectures) technologies developed for this purpose. Inter-relating computational tools make use of this federated data for clinical decision support, and to support the @neurIST research activities centred on cerebral aneurysms [2,8].

One of the clinical centres has fully integrated digital records covering diagnostic services, clinical history and imag-

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

ing, as well as treatment records. To the pre-existing data items captured by this system have been added fields for management of participants within the study (e.g. to record the consent, and whether the required blood sample has been taken). As the clinical data of interest to the study is in large part that routinely captured in the course of clinical care, the effort required for data collection in this setting is kept to minimum. An important further variation from most prior efforts is that here, the institution's information systems themselves interact with queries from users of the @neurIST system that arrive via a connector program in the hospital's Internet *De-Militarized Zone*. This approach is the most complete manifestation of the direct federation of data envisaged by the @neurIST architecture.

The other clinical centres, all have Picture Archiving and Communication Systems (PACS), a booking system and lab test systems as separate data domains, together with handwritten notes covering the patient's condition and care. In these centres, a traditional approach is taken in which a research nurse enters the required data to a dedicated interface from the case notes or in parallel to the routine clinical care. The interface allows other records, specifically Digital Imaging and Communications in Medicine (DICOM) images [9] to be linked in, and can then prepare the data for transfer to a local database in the hospital's DMZ where the connector program can execute the queries it receives. This architecture reflects the very restrictive attitude of most clinical centres to external connectivity and data security.

The development of the @neurIST system and its use by an active research study created a need for coordination of the development efforts with the real-world needs and constraints for healthcare research, which we wish to reflect in this document. To provide a suitable background, we start by establishing the terms on which we will rely in the course of this document.

1.2. *Personal, de-identified, pseudonymised, and reasonable anonymity — terminology*

It is important to have a common understanding of how to prepare data for use in medical research — which makes it surprising that a wide vocabulary of often overlapping terms and contradictory guidelines has developed in this area, even within Europe. This leads to ambiguous interpretations and makes it difficult to have a common understanding, causing problems from legal, ethical, and technical standpoints. This section aims to summarize some of these laws and guidelines, and the terms they use, before clarifying the terminology used in this paper.

The EC Data Protection Directive 95/46/EC [10] and the associated guidance of the Article 29 Working Party[1] [11] addresses the concepts of personal, identified and identifiable data. According to the Directive,

> 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly (Article 2a [10]).

The Article 29 Working Party [11] states that for data to be personal it must both "relate to" a natural person, who must be "identified or identifiable". Further, "particular pieces of information ... which hold a particularly privileged and close relationship with the particular individual" are termed as "identifiers" [11]. Identifiers can be separated into those likely to lead to identification of an individual, and those which act more as indirect clues to the identity of an individual. All of these concepts are explored further in Section 3.1.

The EC Data Protection Directive does not apply to data when the individual is not identified or identifiable. The authors of the Directive, foreseeing that if no possible means for de-identification exists the law could be unworkable, placed limits on the identification process in Recital 26, which states that "to determine whether a person is identifiable [directly or indirectly], account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person" [10]. The Article 29 Working Party opinion gives "anonymous data" an approachable definition: "[a]ny information relating to a natural person where the person cannot be identified, whether by the data controller or by any other person, *taking account of all the means likely reasonably to be used either by the controller or by any other person* to identify that individual [11]".[2] The terms anonymous and anonymised data are often used in relevant guidance synonymously with a spectrum of concepts like: de-identified, non-identifiable, irretrievably unlinked, irreversibly de-identified, unlinked-anonymised, or irreversibly anonymised (see Ref. [12]). Emerging standardisation activities in the privacy enhancing technologies field are attempting to clarify the underlying concepts and associated terminologies [13,14]. The unified definitions of the terms anonymisation (and pseudonymisation) produced by these efforts tend however, to be very formal and exact, resulting in complex and hard to understand constructions and wordings and so do not lend themselves to use in communicating with patients.

The EC Data Protection Directive is often interpreted as stating that data is personal if anyone, anywhere, can directly or indirectly identify the individual — ignoring the principle of reasonable means. In this interpretation, the difficulty in establishing that data has been completely de-identified causes the resulting data to be seen as personal. The authors of a 2003 World Health Organisation (WHO) document introduce the concept of "*proportional* or *reasonable anonymity*" [15] as being useful in the context of genetic databases. According to these WHO guidelines, "*proportional* or *reasonable anonymity* exists when no reasonable means of identification of specific individuals is available". Many ethical and legal documents require some level of de-identification when using data for the purposes of research, and in cases where the research subject must be re-contacted, or it is useful administratively, "cod-

[1] The Article 29 Working Group was established to – amongst other things – promote uniform application of the general principles of the Data Protection Directive.

[2] This is less specific than the 1997 Council of Europe Recommendation on the Protection of Medical Data [21], where data "shall not be regarded as 'identifiable' if identification requires an unreasonable amount of time and manpower (S.1)".

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

ing" of information is advised, rather than anonymisation. The idea is to remove identifiers and place a code or pseudonym on data passed to researchers. Coded data is also referred to in relevant documents as pseudonymised, linked, linked-anonymised or reversibly anonymised (see Ref. [12]). Some guidelines [16] use the term "coded", "reversibly anonymised", or "de-identified" to indicate that direct identifiers such as name, birth date and address are reversibly detached and can be reattached through a code or pseudonym. Other guidelines use the same terms in a stricter way [17]: coded or de-identified in these guidelines mean also that demographic data have been detached to the point that it is not possible to identify a person "easily" without knowing the code. The definition of proportional or reasonable anonymity allows that the use of linked or linkable coded information could be seen as a means to achieve anonymity, when access to the link is restricted appropriately.

Following the considerations above, the paper uses the terms personal data, de-identification, pseudonymisation, and proportional or reasonable anonymity. Personal data refers to data that is *about* an individual who can (reasonably) be identified or identifiable. De-identification is the process of removing (or modifying) identifiers from the personal data so identification is not reasonably possible. Pseudonymisation is the step where a pseudonym or code is added to this de-identified data (methods are described in Appendix A). Proportional or reasonable anonymity applies to de-identified/pseudonymised data which cannot reasonably be used to identify specific individuals.

2. Informed consent

One of the main safeguards to protect individuals while undergoing treatment or research in the medical sphere is informed consent. The principle is to provide individuals with information in a format they can understand, explaining the treatment or research and enabling them to digest it and ask questions, in order to obtain their agreement to any procedures. In practice, there are many complications including: providing adequate time for consideration in cases of individuals who arrive directly in the emergency room and may not recover; involving patients who are incapable or incompetent to give due consideration to their act of consenting; deciding who makes the initial approach to potential participants considering that often the information used to select the potential participant cannot be shared with the people outside the clinical care team. The section on consent below gives a brief overview of guidance before outlining how consent was obtained and recorded in @neurIST.

2.1. *Legal and ethical requirements*

It is a general ethical and legal principle that the informed consent of participants is required when involving them, their data or their tissues, in clinical research. International law and guidance further describes what type of information they should be provided with, how long they must be given to digest it, how many consent forms are needed, and what the specific points of consent should be. This guidance also contains provisions on how to obtain consent or authorization in relation to incompetent adults and other vulnerable groups. While these considerations are important, the relevant documents also contain exemptions to the need for informed consent for research purposes. Therefore in each case, the specific contexts in which these "research exemptions" arise will be introduced. A key conclusion is that obtaining informed consent in the context of international medical research projects is preferable to relying on a research exemption.

International guidelines specifically concerning research such as those contained in the Declaration of Helsinki and the Council for International Organizations of Medical Sciences (CIOMS) stress the importance of informed consent. The cornerstone of informed consent is adequate information. Specifically:

- Patients or their representatives (if patients are incompetent) need to be informed about risks and benefits of the study, and consent to participate therein.
- Information about the way in which data are protected and the uses to which the data is to be put needs to be made available.
- Each individual must be given as much time as needed to make a decision, including time for consultation with family members and others.
- Each site performing data and sample collection must obtain local ethical committee approval.
- The procedure for identifying individuals for recruitment needs to be well-defined.

There is an absolute requirement, both in European law and medical ethics, to obtain informed consent from individuals when it is an interventional clinical trial, or research that *involves* human subjects or interventions on human beings (this could be a physical intervention, or one which risks the psychological health of an individual) (see Refs. [18–20]).

Research solely involving personal data and/or tissues is not subject to a specific law at the European level. The Data Protection Directive 95/46/EC offers guidance in relation to personal data [10]. The Council of Europe has issued a recommendation on the protection of medical data [21] that covers the use of these data in research; as well as recommendation (2006)4 of the Committee of Ministers to member states on research on biological materials of human origin [22]. The Council for International Organizations of Medical Sciences (CIOMS) International Ethical Guidelines for Biomedical Research Involving Human Subjects [23] and the WMA Declaration of Helsinki also cover data and tissue research [19]. The documents relevant to research using data are reviewed below.

Following the Data Protection Directive, consent is one of the processing conditions for both personal and sensitive personal data (see Articles 7 and 8). Consent for personal data means any unambiguous "freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed (Article 2(h))" [10]. Where sensitive data is to be processed the consent must be "explicit (Article 8(2)a)", but this is not further

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

defined.[3] A separate processing condition for sensitive data is for the purposes of medicine, diagnosis or the provision of care or treatment (Article 8(3)).

The data protection principle that the OECD (Organisation for Economic Cooperation and Development) term "the purpose specification principle" [24] is thus relevant here: that "the purposes for which personal data are collected should be specified not later than at the time of collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible ... and as are specified on each occasion of change of purpose" (guideline 9) [10]. In the Directive elements of this are found in Article 6(1)b which requires that data must be "collected for specified, explicit and legitimate purposes". The data subject also has a right to be informed of the intended use of data whether it is collected directly from them or not — in the latter case the act of informing the data subject should be performed when the data is first disclosed or gathered (Articles 10 and 11). The information provided should include the identity of the data controller, the purposes of processing, and any other information to make the decision. Thus, as the data used for clinical research will often be of a sensitive nature, which includes details of racial or ethnic origins, religious beliefs or health (Article 8(1)) the explicit consent should be obtained where the data is collected for research directly from the data subject or where secondary use of information is foreseeable.

There are research exemptions to almost all of these requirements. Firstly, in some countries exemptions may be made where this is in the substantial public interest, for example for medical research (Article 8(4)). Second, the purpose specification principle in the Directive includes an exemption when data is further processed for scientific purposes which would not be seen as incompatible with the initial purposes (Article 6(1)b), and an additional exemption exists for scientific purposes where it would involve a disproportionate effort to provide the information (Article 11(2)).

While the results, in the form of statistical descriptions and findings may enter the public domain via publication, there is increasing interest and use for database and biobank uses of data and samples for reference purposes. As well, many research efforts form a step in the creation of products or services that will subsequently be commercialized. Where there is known potential for the passage of data from the research use into commercial application, this should be stated in the consent information provided to the data subject. Such information at the outset incorporated in the consent is protective of the interests of downstream business entities.

The Directive however, does not clearly address what happens when the purpose the personal data is used for changes; unlike the OECD purpose specification principle, which requires new information to be given in this case. Thus, following the Directive, further use of information originally collected for medical purposes would be possible for medical research without informed consent. An important argument against using this provision on an international medical research project is that both exemptions and safeguards have been interpreted and implemented differently in individual countries [25]. For example, some countries have chosen not to implement all of the available research exemptions. As a result international medical researchers cannot be sure the same research exemption applies in each country in the same way, when in fact the Directive was meant to provide harmonisation.

The Council of Europe [21] Recommendation on Medical Data states that "[m]edical data may be collected and processed ... if the data subject or his/her legal representative or an authority or any person or body provided for by law has given his/her consent for one or more purposes, and in so far as domestic law does not provide otherwise" (Section 4(3)c). The consent should be free, express and informed. The information given to the data subject includes: where the data will be collected from, if applicable; purposes of communication of the data and who would receive it; whether there is a possibility to refuse consent and the consequences of withdrawal; and the conditions under which the rights of access and rectification can be exercised (Section 5). Prior to "a genetic analysis, the data subject should be informed about the objectives of the analysis and the possibility of unexpected findings" (Section 5(4)). The information should be given at the time of collection, unless the consent was not collected from the data subject when the notification should occur as soon as possible (Section 5(2)).

The Council of Europe Recommendation also contains exemptions which may apply to research. The first is that medical data may be collected and processed without consent if permitted by law for public health reasons or an important public interest. If the data was not collected directly from the individual, and it is unreasonable or impractical to provide information to an individual, it is not necessary. Section 12 on scientific research also envisages exemptions to information and consent when it is a "defined scientific research project concerning an important public interest", but only if it is impractical to contact them, amongst other things. As with the Directive, these exemptions are more applicable in the case of further use of medical data. However, a more crucial problem for international medical research is that this recommendation is not universally applied — again, making it unwise to rely on this research exemption.

Some vaguely defined waivers for consent concerning research involving medical data have recently been introduced in the revised version of the Declaration of Helsinki, to which the following Article (25) was added in 2008:

> For medical research using identifiable human material or data, physicians must normally seek consent for the collection, analysis, storage and/or reuse. There may be situations where consent would be impossible or impractical to obtain for such research or would pose a threat to the validity of the research. In such situations the research may be done only after consideration and approval of a research ethics committee [19].

According to the CIOMS guidance "[w]aiver of informed consent is to be regarded as uncommon and exceptional, and must in all cases be approved by an ethical review com-

[3] In many countries this requirement has been translated into law as requiring written consent, for example in Belgium, Greece, Bulgaria, Hungary, Latvia, Lithuania and Poland. In Germany the consent must include an express reference. See Ref. [25].

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

mittee" (guideline 4). These principles make clear that while secondary research using medical data is possible without consent this should be an exceptional position, only taken when obtaining the consent is impossible or impractical. This reaffirms the position that research involving medical data should, as far as possible, only take place with informed consent.

It is worth noting that biological samples are increasingly seen in a different light to more conventional descriptive data. This is in part because of the as yet underdeveloped, but certainly growing potential of the genetic and proteomic data contained therein to collaterally generate relevant findings about the individual outside the research focus. Equally, genetic data has a relevance to related family members that can be considered. Thus, separate consent is recommended for the storage of samples and the use and storage of genetic information in line with current recommendations (see for example the recommendations from the UK Medical Research Council regarding informed consent [26]).

2.2. *National laws*

The legal requirements in the member countries of the project vary more than the guidelines for research ethics [27]. There are a variety of styles of legislation governing medical research, these include: clinical trials or biomedical research legislation based around individual research projects involving interventions to human subjects; national genome project legislation in Estonia and Latvia; laws protecting both tissue and data in biomedical research, for example in Hungary, Norway, Portugal and Spain [28–30]; and countries with separate regimes governing biobanks (Iceland and Sweden) or human tissue (see UK [31] and Finland) as compared to data.[4] Many countries also include legislation governing the ethics committee systems in their countries which may apply to both tissues and data, for example in Sweden and Denmark. Laws have also begun appearing on genetic testing, for example in Switzerland [32] and Germany, but these do not always apply to medical research. This varying legislation, all including requirements for informed consent (and potentially research exemptions), results in a confusing situation for international medical researchers.

2.3. *Consent in practice*

For each @neurIST participant, a clinical history is to be collected. A blood sample and any existing radiological images suitable for use in haemodynamic modelling are also sought, where possible from existing clinical records [2]. In order to achieve the ends of the secondary uses envisaged in @neurIST (such as associating gene expression with phenotype and outcome) it is necessary to retain the individual (and hence personal) nature of the data. As well, a minor physical intervention is involved to collect a blood sample. In light of the open legal debate regarding data handling, biological samples and genetic data the @neurIST project chose to follow the strictest rules in order to be able to apply the same legal framework in all partner hospitals. Thus, consent of the participants is an essential factor in @neurIST. The consent procedures related to data protection are in line with European law and confidential handling of the information is necessary [10,21].

The @neurIST Study Protocol, including the information sheets and consent forms, were drafted in line with ethical recommendations that apply in Europe [10,18,22,23,33] and thus rely on classical informed consent. Participants enter the @neurIST study on giving their informed consent. When recruiting participants to the @neurIST study, two sets of information sheets and consent forms are used. The first applies to all participants and covers involvement in the study together with the use of clinical and imaging data within the project. The second relates specifically to the genetic testing taking place in @neurIST, and is only be used if the individual is willing to donate a biological sample for research use.

Each consent form contains several items that require the prospective participant's agreement in order for them to participate in the study. These are statements acknowledging that they have read the information sheet and been able to ask questions; that the data may be used in a non-identifying form for publication; that the data may be used as part of future commercial applications and they understand they will not profit from this; that they agree to take part in @neurIST and to give access to their health records; that they understand they can withdraw at any time and request destruction of the sample; that coded data may be transferred outside the European Union; that samples may be stored and used for genetic analysis for the @neurIST project, and further transferred as necessary for @neurIST.

In addition, there are four optional items representing choices for the participant. These include an agreement for further use of the data or sample for research into other cerebrovascular diseases beyond the scope of aneurysms; an agreement that they can be re-contacted for further consent or to request other information; a request that @neurIST informs their GP about participation in the project; and a choice of whether to receive via their @neurIST clinician, genetic or other study findings which may be relevant to their health.

These different options are recorded into the @neurIST study database and as a result are linked to each participant. These options will be consulted before the @neurIST project re-contacts participants or returns any results. The samples are also labelled with the information on whether the individual has agreed to further cerebrovascular research after the @neurIST project ends. Both the samples and the data therefore provide records for the researcher as to the wishes of the patient. With regards to the means by which their privacy and data are protected, patients enrolled in the @neurIST study are simply informed that their data will be protected adequately and in conformance with local legal requirements, without providing technical details of the mechanisms. A summary of the data protection mechanisms is provided to the ethical committees as part of the research protocol. These procedures are discussed below in connection with privacy protection for the participants.

The @neurIST study has been approved by the local ethics committee of each hospital in which participants are being

[4] For an overview of relevant legislation in countries, see the website of the EC-funded PRIVILEGED project, www.privilegedproject.eu (last accessed 13 October 2009).

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

enrolled and data collected. To a certain extent, the ethical requirements placed on a research project are prone to the subjective interpretations by local ethics committees of the general principles encapsulated in the guidelines to which they work. This is reflected in the different requests for amendments of the information sheets and protocols of the project which led to a revision of the study protocols and documentation. One such point was the approach to identifying controls for recruitment. To ensure the best matching for age, gender and ethnicity, the original protocol requested that recruited patients suggest and make first contact with suitably matched individuals. One committee objected to this practice on the grounds that requesting recruited patients to suggest and make first contact was asking too much of the patients, and required open advertisement instead.

A final factor is the workload implications for the institutions, and their response to the recommended practice. Some centres argued that the costs and effort required for re-contacting patients with clinically relevant research results could not be supported, as this would necessitate confirmation testing, possibly genetic counselling and logistics beyond the realistic scope of the research and project funding. These sites were allowed to abstain from the return of results, and the relevant item excluded from the consent forms for their institution.

3. De-identification

3.1. *Legal and ethical requirements*

The majority of laws or standards relating to data (or tissue) protection in clinical research take the position that a reduction in the potential to identify an individual correlates with an increasing protection to the privacy of those individuals. Thus, any discussion of privacy protection in research involving patient data, the ethical and legal consequences associated with the validity of informed consent due to the patient's ability to judge the risks, and the need for consent in the first place depends heavily on data protection terminology. The terminology also has implications for Research Ethics Committees or other Boards assessing the research proposal as they need information on how data will be processed, stored and protected in order to judge the risks versus benefits of the proposal. Internally, the terminology is critical for system developers to implement a coherent data security strategy. Following the terminology outlined in this paper, data used in clinical research fall roughly into one of three categories: personal data; de-identified data which have been coded; de-identified data which may or may not be coded, and cannot reasonably be used to re-identify a specific individual. While this seems straightforward, many questions arise, for example: What are the identifiers that need to be modified or removed? What is the difference between direct and indirect identifiers? Who may hold the key to a code? Can I ever be sure re-identification is impossible? What does "reasonable" mean in the context of potential re-identification? Can data from the electronic patient record ever be reasonably anonymous? Can genetic data? Do I need to do a risk analysis? This section offers tentative answers to these questions though outlining the ethical and legal requirements in relation to de-identification.

Disentangling the concept of personal data is vital to understanding identifiers. According to the Article 29 Working Party [11], for data to be personal it must both "relate to" a natural person, who must be "identified or identifiable". The words "relate to" create a requirement that the data must be "about" an individual, and not an object.[5] It is clear that medical information and other information used in clinical research would fall under this category. It is, of course, precisely the "information relating to a natural person" aspect of medical records that is of interest for many of the secondary uses foreseen in @neurIST and other areas of biomedical research. A second important aspect of clinical data is its providence — "where did it come from?", "what conditions was it collected under?".

Where distinct data about individuals is not relevant to a secondary use, "depersonalisation" could be achieved through aggregating or grouping the data together by using summary values such as mean and standard deviation. This happens for example when medical institutions provide summary descriptive statistics for epidemiological or audit use. This has the secondary effect of making it impossible to identify one individual out of the group despite knowing the centre at which the data was collected. A list of values, one for each subject will lend itself to more general use, but together with even basic information about the location where the data was collected, may allow some individuals to be recognized, particularly for the extreme values. This leads us to consider what the terms "identified or identifiable" in the Directive mean.

The Article 29 Working Party opinion [11] states that if enough identifiers are present in the dataset, they may alone, or in combination, lead to identification — and this will depend on the context of the particular situation. The Data Protection Directive mentions both the possibility of direct and indirect identification. A direct identification could occur when enough identifiers are present in the dataset for someone (anyone) to be able to associate that record to a specific individual. It is also conceivable that combinations of the data items with additional data sources, either intentionally or unintentionally — could achieve identification. The information sources may be incidental (for example the recognition of a known person based on facial features) or externally referenced (based for example on data reduction against publicly available sources). The potential for indirect identification should not be dismissed lightly; Sweeney [34] found that 87% of the population in the United States, could potentially be uniquely identified by their five-digit ZIP code, combined with their gender and date of birth.

Identifiers can be separated into those likely to lead directly to identification of an individual, and those which act more as indirect clues to the identity of an individual, aiding potential indirect identification. Direct identifiers include identification or unique numbers (for example national identity, social secu-

[5] The Article 29 Working Group [11] gave the example of the value of a house, which is information about an object, and therefore does not "relate to" an individual. However, this could become personal data when it becomes information about the financial assets of a person.

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

rity or hospital numbers), names, address, telephone number, profession, place of birth and dates more specific than a year (including dates of birth or death, treatment or admission dates). The US privacy rule [35] includes "ages over 89" as an important element to be removed to protect health information, presumably because the number of people who survive past this age is so low that individuals in a particular geographic area may be directly identifiable. The UK Medical Research Council [36] guidance in this area includes data such as a photograph or image of the face, a noticeable disease or disability, and ethnicity as potentially leading to an identification. Indirect identifiers include data such as: a rare disease or treatment, an unusual height or weight, family medical history, doctor responsible for care of a patient, the name of the hospital where the patient received treatment, medication use, religion, salary details or the DNA profile. Any of these identifiers could directly or indirectly, alone or in combination, potentially or actually, lead to the identification of a natural person.

In the context of clinical care and treatment, the patient's identity is protected by medical confidentiality norms. In light of the above-mentioned possibilities for identification, the principle that data is only anonymous if identification is not possible by anyone, anywhere is problematic for de-identifying medical data for secondary use in research for at least three reasons. Firstly, it is unlikely that the original clinical record will ever be deleted, and so this avenue of identification must be considered ever present. Secondly, there is ample scope for indirect identification based on combinations of values unique to individual patients in the clinical data. Thirdly, in the context of medical data, it may be that the clinician or members of the clinical team are sufficiently familiar with a given patient to recognize them from peculiarities of their medical condition. These problems are compounded by the fact that in practice, few national data protection laws have explicitly incorporated the idea of identification as being limited to reasonable means.[6] A further consideration arises when genetic data is included in the research. Recent commentators have said that "[w]henever genetic samples are involved re-identification will be possible" [37]. The solution for this group was to accept that in genomics research, privacy and confidentiality is impossible to protect completely.

Complete anonymisation of the extract of clinical records or a genetic sample that is to be used for secondary research is thus, unlikely to be feasible while the original clinical record persists. De-identification should nonetheless be as complete as possible such that the extract in isolation from the original record would be reasonably anonymised. This can be strengthened by eliminating any link between the extract data and the original record. As noted above, there are reasons that support the existence of such a link or code. The Organisation for Economic Cooperation and Development (OECD) Guidelines for Human Biobanks and Genetic Research Databases [38] state that "[u]nless strictly necessary, researchers should be provided access only to ... data or information that is coded such that the participant cannot be identified (7.D)". The WHO document interpreted the concept of "*proportional* or *reasonable anonymity*" to be in line with the Data Protection Directive [10], and considered therefore acceptable that proportional anonymity be used to secure genetic data. The use of linked or linkable coded information that has been suitably de-identified is consistent with reasonable anonymity, when access to the link is restricted appropriately. However, it depends on how this is done, who is to have access, and the uses which have been consented to.

Codes placed on the data for research use may not always be reversible (see Appendix A, "one-way single-pass"), they may be one-way and yet still allow updates to the research record. However, in clinical research it can be useful to allow re-contact of participants under certain circumstances, and so reversible systems are often preferred (see Appendix A). Following the implementation of the EC Data Protection Directive, data that is reversibly coded is viewed as "personal data" in the majority of countries, and thus falls under the data protection law in those countries. This is because it has been interpreted that data is personal if anyone, anywhere could re-identify the individual. There are exceptions to this, for example in the UK Data Protection Act 1998 [39] the only person who is important in terms of indirect identification is the data controller.[7] Thus, if the researcher is the data controller and holds coded data, but cannot re-identify the individuals, this would not fall under the Act. This means that in some countries *who holds the code* may be important in terms of the law. The questions of *who holds the code*, and indeed whether there *is* a code which enables re-identification, may blur the distinction between coded and reasonably anonymised data.[8]

The OECD guidelines [38] suggests other policy options, such as that "researchers should be required not to attempt to re-identify participants (7.D)", if various partners collect the data "each partner holding these could use their own code with none of them holding the totality of the codes (55)", or honest broker systems could be used to remove or modify identifiers. Reviews of pseudonymisation techniques mention (independent) Trusted Third Parties (TTPs) who apply a code to the data and/or hold the identifiers [40]. The European Medicines Agency (EMEA) suggest that the TTP could be "an external entity, such as governmental agency, legal counsel, or other qualified third party not involved with the research" [41]. An example of one approach is the Icelandic Health Sec-

[6] The German Federal Data Protection Act (of the 20 December 1990 (BGBl. I 1990 S. 2954), as amended) states that depersonalisation means "the modification of personal data so that the information concerning personal or material circumstance can no longer or only with a disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual (S.3(6))". Other countries to include similar provisions to German law include the Czech Republic, Poland and Slovenia. All refer to time and cost as being relevant factors. Manpower, effort and material resources are also referred to. (Information gathered as part of the PRIVIREAL project, www.privireal.org.)

[7] Similar provisions are found in Ireland and Finland.

[8] According to the 2006 Council of Europe Recommendation on Research on Biological Materials of Human Origin [22] there is a difference between linked anonymised (reversibly anonymised in the French version) and coded. In the first case the researcher using the samples or data does not have access to the code, whereas the latter (coded) means that the researcher using the material/information knows the code.

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

tor Database which put in place a "third-party encryption system" in collaboration with the Icelandic Data Protection Commission [42]. The International Subarachnoid Aneurysm Trial used a "24-h telephone randomisation service" provided by a Clinical Trials Service Unit at the University of Oxford [43].

The WHO guidelines remind us also that "mere compliance with legal standards of anonymity may not be enough, if, for example, anonymisation procedures (even if only to the stage of reasonable anonymisation) are inadequate, without review" [15]. Hence, the anonymisation process should be overseen by an independent body that would have the obligations to maintain standards and keep anonymisation processes under review.

In line with the concept of "reasonable anonymity", an alternative way to ensure de-identification according to the US federal medical privacy rule is to prove that "a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods" determines that the risk of identification is "very small" and provides documentation of "the methods and results of the analysis" [37,44,45]. This amounts to conducting a risk assessment, technology assessment, or privacy impact analysis. There is a wide literature on risk and technology assessment, and some guidelines on how to conduct a privacy impact assessment [46,47]. Simplistically, a risk assessment involves characterising the potential hazards, as well as quantifying the impact or effect these could have. Two major problems are that quantifying the impact of an event can be a highly subjective process, and in the context of data security it is unclear what "the methods" of risk assessment are. For example, should someone try to re-identify individuals and let us know how they get on? Should this be a third party auditor? Malin and Sweeney evaluate anonymity protection systems through using trail re-identification though specific algorithms [48]. Taking publicly available "anonymised" genomic data in the US, in some cases they achieved 100% re-identification of individuals — for example with single-gene disorders such as Refsum's disease. They state that "features about the data, as well as the environment in which the data are shared, [should be] ... taken into account before the data can be declared anonymous" [48]. There are also standardisation activities in the area of data security, which suggest methods for evaluating information security [49].

Potential hazards in clinical research include that the clinician, members of the clinical team, or people working in the hospital may be able re-identify individuals as the original clinical record will never be deleted, or they may be sufficiently familiar with a patient to recognize them from the peculiarities of their medical condition. While this would do little more than demonstrate identification was possible as the data from the medical record would already be in the hands of those doing the identification, it may be that this would also give them access to derived data in the research domain (genetic analyses, or in a more general context drug trial results). A specific consequence of identification of research data would be that it is not yet validated and may lead to inappropriate clinical decisions. Another hazard is that the scope for indirect identification based on combinations of values unique to individual patients is often high in clinical research, especially in epidemiological studies where large amounts of data are collected. This leads to a higher possibility that the data can be combined with other data outside the study record to re-identify individuals. Further hazards arise if the research project includes genetic data, which may give rise to higher and more reliable possibilities of re-identification of individuals or their families. Along the same lines, research projects may include images of the body or face which must be carefully de-identified as a full face picture is seen as personal data. The final hazard is that the growing federation of data access and its use by third parties outside of hospitals raises the likelihood of interception and subsequent attempts to identify the source of the data.

Characterisation of the risks to data security on a research project is useful for applications to the research ethics committees, who must balance risks and benefits, and for the information sheets given to the research participants, with the idea that they give a valid consent to take part considering all the risks. Finally, as there is a legal and ethical obligation to protect medical confidentiality whilst undertaking secondary use of health data, it is a legal and ethical responsibility to implement measures to minimize the risk of patient privacy being violated. This section has introduced some of the general considerations for measures moving towards de-identification. The following sections will focus on experience on the @neurIST project, introducing the model of the project, the identifiers removed from the clinical database, the de-identification of medical images, pseudonymisation and access restriction to certain data types.

3.2. *De-identification in practice*

The following three points factored in the development of the @neurIST study privacy protection policy. First is the size of the study database. The clinical reference information model for aneurysm patients contains over 500 unique data fields likely to emerge in the course of diagnosis, recruitment, treatment and follow-up of a patient. At least 100 items (for example aneurysm location, size, patient age and gender) are considered essential to all the downstream users within @neurIST and 25 have established relevance for rupture risk reported in a Cochrane review [72]. Many of the remaining data items are of interest for database research via data-mining to identify the next generation of likely risk factors for investigation. Consolidating these records into categories rather than the patient specific values would have reduced the possibility for re-identification, but relevant categories were not known *a priori*. Simply due to the number of different variables contained in the records of distinct individuals it is likely that re-identification of many patients would be possible if compared against full or partial extracts of the original medical record, which still exists.

The second major consideration is that @neurIST eliminates the centralized, monolithic database typical of traditional research data collections. Clinical centres may choose to allow the local interface "connector" program that communicates with external @neurIST programs to directly forward requests to the EHR servers within their institution (the idealized vision of the project) or to servers placed in a "de-militarized" zone of their network that contain extracts specific to the patients who have consented (the pragmatic

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

requirement of many clinical centres). Both approaches have been adopted within the project, and so the system must be able to link distributed, multi-format, and multi-scale data from active clinical domain sources (lab reports, images, clinical history). Thus, the @neurIST project must include real-time implementation of the privacy protection measures, often termed "on-the-fly" [50].

A third important point is that follow-up data must be incorporated into the system, and a mechanism for potentially returning results relevant for the health of participants is required. The @neurIST data structure envisages the arrival of follow-up data at intervals ranging from months to years after recruitment, and incorporation of this data must be performed in such a way as to maintain the association with the original dataset obtained for the same patient. At the same time, some research ethics committees required that the ability to return relevant results to individual patients be maintained. Thus, the mechanism for releasing data for secondary use must remain consistent for all subsequent data from a given participant, and a means must exist for identification of individual patients under specific conditions at a later date.

In consideration of these points, complete anonymisation is not a realistic procedure for the @neurIST project, and we have therefore sought to achieve reasonable anonymisation in the sense of the WHO guidelines. De-identification, data minimisation, aggregation, and pseudonymisation feature in the strategy for privacy protection in @neurIST (illustrated in Fig. 1), which would result in data that do not contain any direct identifiers, and a reduced set of possible indirect identifiers (generically referred to as personal identifying information (PII)) but can thereafter be associated or linked to a certain individual. The overall definition of these processes was developed in accordance with the above-mentioned standards and initiatives. The application of these procedures to the major data entities in @neurIST, structured clinical data and images, are considered in the following sections.

3.2.1. *Clinical reference information model and crop-list*

The current therapy of cerebral aneurysms is based on very limited information such as the size and localisation of the aneurysm. To ensure the widest possible applicability of the available data, we took the approach of developing a Clinical Reference Information Model (CRIM); that is, a data model containing all the clinical information regarded necessary for clinicians and researchers involved in the project. Set up as a relational database for storage purposes, the CRIM currently (October 2009) contains 90 tables and 1272 attributes, falling into three categories: administrative items, clinical findings, and treatments. Administrative items include records such as payment of travel expenses, whether the GP has been informed of the patient's participation in the project. The clinical findings range from personal history, signs and symptoms, vital signs, laboratory examinations, to image-associated findings acquired at the patient recruitment and at each follow-up. All CRIM items are formally defined using an ontology [51] and available for consultation and download.[9]

[9] http://www.imbi.uni-freiburg.de/aneurist/ontology/ (last accessed 24th September 2009).

Whereas the HIPAA rules set a guide for de-identification [35], it must be recognized that these apply to directly identifying data. An important concept for minimising the risk of indirect identification is the principle of data minimisation — the limitation of data made available to that required for the task at hand (also included in the European Data Protection Directive 95/46/EC [10]). The less data exposed, the lower the risk of unauthorized and unwanted (indirect) identification. @neurIST combines these considerations through the creation of a crop-list specifying the items of the CRIM to be removed or modified before disclosure of information. The guiding principle for this list was that the default mode for all data fields should be not to allow disclosure via the data infrastructure. Those items considered directly identifying or collected solely for clinical or administrative use (for either the clinical care team or the database) are locked in this state so that no directly PII or administrative data is communicated to researchers. Of the remaining data items, those required for research use were then identified in discussion with the data users and removed from the crop-list. In the cases of PII being required for research purposes, transformations were introduced to reduce the potential identifiability while still allowing its release. This overall approach ensures that the data disclosed to researchers is "adequate, relevant and not excessive in relation to the purposes ..." (Data Protection Directive Article 6.1.c [10]).

The transformation rules applied to data during de-identification fall into four categories: conversion (transformation of the data value into another one), truncation (only a part of the data is transmitted), access restriction (only authorized users can query the information) and removal (the information is removed from the result). As specific examples, the hospital number is converted into a pseudonym and serves as a unique identifier of the participant for the secondary use of the data from across the federated database (details of this conversion are considered in Section 3.2.4). Other project-specific ID numbers, are either removed or access-restricted. These include local study identifiers to count each patient, sample ID numbers and barcodes, links to internal participants, clinical site IDs, and image IDs. In place of the date of birth, the participant's age at recruitment is expressed in complete years since birth and all other dates are expressed in terms of days from this event. This permits coding the age of the patient with respect to other events. For postal (ZIP) code data, only the two first digits are kept to provide regional information. The truncated zip code, country and ethnicity values are only accessible for authorized users such as the epidemiologists.

The crop-list in @neurIST is relatively strict. However, it has the flexibility to adapt based on justifications made to an oversight committee within @neurIST over the course of the project. For those data fields where the balance between disclosure for research and removal or modification as they are potentially identifying factors has been struck in favour of disclosure for research: there may be a warning attached and it may be preferable to restrict access to these fields.

3.2.2. *Medical images — DICOM headers*

Outside the data model, structured data can also be found as part of what is mainly unstructured data. An important

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

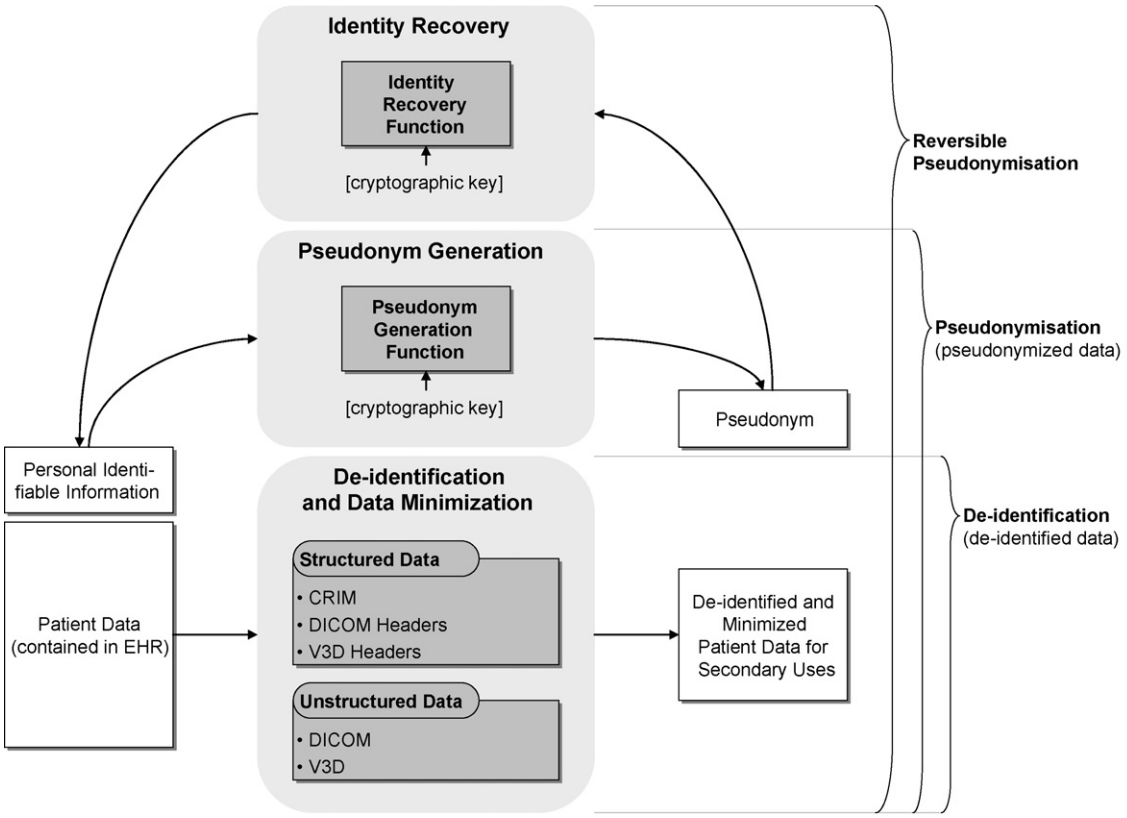

Fig. 1 – Functional view and classification of the terms De-identification, Pseudonymisation, and Pseudonym Reversal.

example of this in @neurIST is the header in medical image files. The DICOM data format is an open and a well documented standard widely used to exchange medical image data. A DICOM structure is organized in Data Elements. Each element is defined by a Name, a Tag and a Value Representation (VR) that defines the type and the length of a DICOM data element value. Even the image data itself is a particular DICOM Element named "Pixel Data". It is common practice to refer to the data that is not the image as the header though by convention DICOM stores both the header and image data within a single file.

The DICOM standard has provided a document entitled "Supplement 55: Attribute Level Confidentiality (including De-identification)" [52] that recommends a list of data elements to be protected in order to provide a minimal level of confidentiality for identification. De-identification of DICOM data elements for research purposes is also discussed by Noumeir et al. [53]. The list proposed by the DICOM standard, along with all private data items in image headers was adopted as the initial crop-list for treatment of imaging data in @neurIST.

The information contained in most of the crop-listed DICOM elements (see Table 1) is not relevant for research purposes. Thus, the default action for these attributes is to clear them (set to a null value) if the element is required by DICOM, or remove them (where the value cannot be null according to the DICOM standard).

Several DICOM data elements however, have a role in the secondary use of the data that warrants their release in a manner that respects the requirements of preventing direct and minimising their utility for indirect identification:

- The Patient ID is replaced by the reversible pseudonym (see Section 3.2.4) to maintain the linkage with other data items in the research domain.
- The values of Study Description, Series Description and Protocol Name are free-text entries that often contain metadata regarding the scan method that are not readily available from other DICOM fields and are relied upon by researchers for making decisions about image processing. It is also possible that these hold identifying information like names or dates. For this reason, a text filtering algorithm is applied [54] that removes any text strings from this field that were removed in the process of depersonalising the CRIM fields, as well as any numeric strings conforming to common date formats.
- Unique Identifiers (UIDs) are used in DICOM to unambiguously relate objects such that a study, series, image, or result to a specific scanner, patient and time of creation. Some of these UIDs (e.g. SOP Instance UID, Study Instance UID, Series Instance UID and Frame of Reference UID) are mandatory fields in DICOM, and must be present in order to conform to the DICOM standard [9] and allow the use of many software tools available for handling DICOM data, or for their storage

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

Table 1 – List of DICOM data elements to be removed/modified for pseudonymisation.

Data Element Name	Tag	Action
Patient ID	(0010,0020)	Replaced by a pseudonym
Protocol name	(0018,1030)	Unchanged
SOP Instance UID	(0008,0018)	Replaced by new UID
Study Instance UID	(0020,000D)	Replaced by new UID
Series Instance UID	(0020,000E)	Replaced by new UID
Frame of Reference UID	(0020,0052)	Replaced by new UID
Instance Creator UID	(0008,0014)	Cleared
Accession number	(0008,0050)	Cleared
Institution name	(0008,0080)	Cleared
Institution address	(0008,0081)	Cleared
Referring physician's name	(0008,0090)	Cleared
Referring physician's address	(0008,0092)	Cleared
Referring physician's telephone numbers	(0008,0094)	Cleared
Station name	(0008,1010)	Cleared
Study description	(0008,1030)	PII is removed from the text
Series description	(0008,103E)	PII is removed from the text
Institutional department name	(0008,1040)	Cleared
Physician(s) of record	(0008,1048)	Cleared
Performing physicians' name	(0008,1050)	Cleared
Name of physician(s) reading study	(0008,1060)	Cleared
Operators' name	(0008,1070)	Cleared
Admitting diagnoses description	(0008,1080)	Cleared
Referenced SOP Instance UID	(0008,1155)	Removed
Derivation description	(0008,2111)	Cleared
Patient's name	(0010,0010)	Cleared
Patient's birth date	(0010,0030)	Cleared
Patient's birth time	(0010,0032)	Cleared
Patient's sex	(0010,0040)	Cleared
Other patient Ids	(0010,1000)	Cleared
Other patient names	(0010,1001)	Cleared
Patient's age	(0010,1010)	Cleared
Patient's size	(0010,1020)	Cleared
Patient's weight	(0010,1030)	Cleared
Medical record locator	(0010,1090)	Cleared
Ethnic group	(0010,2160)	Cleared
Occupation	(0010,2180)	Cleared
Additional patient's history	(0010,21B0)	Cleared
Patient comments	(0010,4000)	Cleared
Device serial number	(0018,1000)	Cleared
Study ID	(0020,0010)	Cleared
Synchronization Frame of Reference UID	(0020,0200)	Removed
Image comments	(0020,4000)	Cleared
Request attributes sequence	(0040,0275)	Cleared
UID	(0040,A124)	Removed
Content sequence	(0040,A730)	Cleared
Storage media file-set UID	(0088,0140)	Removed
Referenced Frame of Reference UID	(3006,0024)	Removed
Related Frame of Reference UID	(3006,00C2)	Removed

in a PACS. Moreover, these UIDs must be internally consistent, meaning that while it is possible to generate UIDs that do not relate to a specific real-world scanner, the new UIDs also have to preserve the object hierarchy and their respective relationships. For instance, each series (scan) has to have a single Series Instance UID, while multiple Series within a study may share the same Study Instance UID. In @neurIST, we generate new UIDs based on the method described by Kamauu et al. [55]. A file is temporarily kept on the server performing the depersonalisation at the hospital containing the mapping between the original Study ID and a part of the new UID (uniqueness stamp). This mapping allows building the requisite new UIDs, while preserving the consistency between DICOM files. Other UIDs are removed and when they are in a sequence attribute, all the sequence is removed.

A further consideration is that the DICOM standard reserves several available fields for use according to the discretion of the manufacturers (private fields). There is no reason for identifiable information to be placed in such fields but the possibility for information been echoed in these fields exists, and so they are removed.

3.2.3. *Medical images — facial images*

Identifying information in unstructured data is much more difficult to accurately distinguish than in structured data. As a consequence, de-identification or depersonalisation of

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

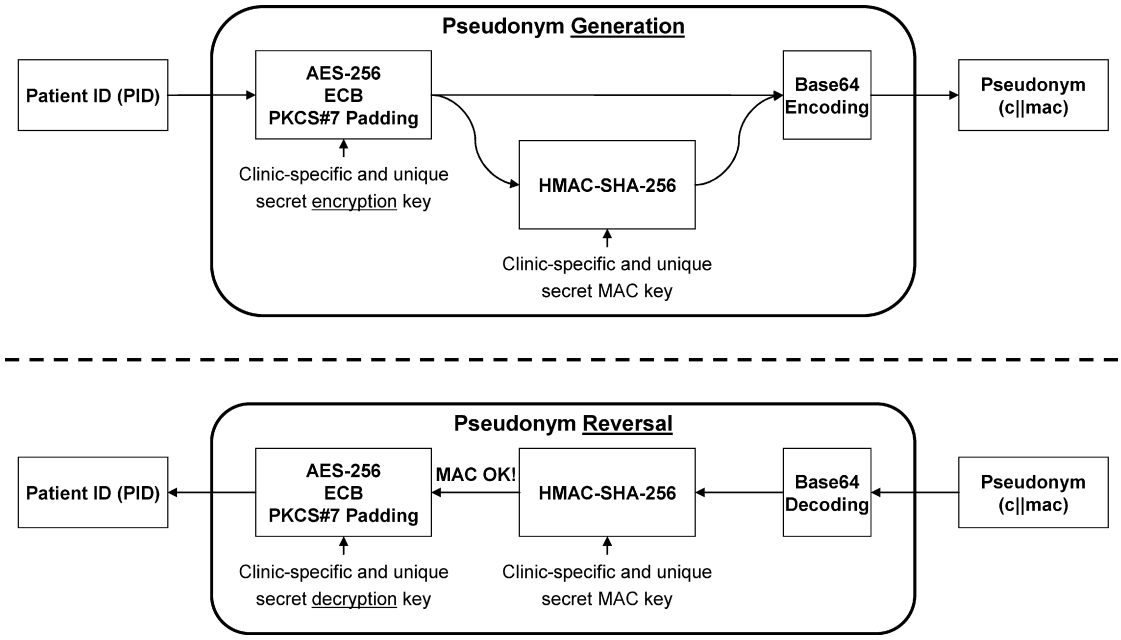

Fig. 2 – Pseudonymisation (upper) and reversal (lower) scheme implemented for @neurIST test-bed.

unstructured data is a complex task, which can lead to two undesirable consequences: firstly, it may destroy the utility of data for secondary use by researchers; secondly, some elements of identifying information might not be removed and allow for indirect identification. Thus, each data type requires a particular method. In the context of @neurIST, the unstructured data requiring transfer for secondary use are images in DICOM format. The two most important types of identifying information contained in the pixel data of DICOM images are:

- DICOM images particularly those generated by secondary capture (screenshots) or modalities such as ultrasound commonly include patient and institution information in the pixel data by replacing, or storing "inscribed" values (comparable to writing on a photograph).
- Images produced by CT, MRI, and in some cases by XA can often be reconstructed to view the patient face in 3D.

Inscribed data is particularly difficult to deal with as it is machine and vendor dependent and requires over-writing or cropping regions of the image. In @neurIST, image types containing pixel data information are not transmitted to researchers unless a template has been applied to over-write specific regions of the image. The size and location of the region to be replaced when applying the template is defined specifically for the modality and manufacturer of the device used to capture the image.

Despite the fact that hair, eyebrows, and skin pigmentation are missing or may be poorly depicted, it is possible to identify a person from rendered views of the tomographic image data to be used in @neurIST, particularly when this person is familiar to the viewer [56]. Bischoff-Grethe et al. [57] proposed an automated "defacing" algorithm that removes only identifiable facial features while preserving brain tissue. A variation of this general strategy of removing the frontal face features was developed in the project, and is applied to images at the time of de-identification.

3.2.4. *Pseudonymisation*

Completing the process of data release for secondary use in @neurIST is the association of a pseudonym to the de-identified data. The actual degree of de-identification achieved within @neurIST having been described in the previous sections, it remains to consider the pseudonymisation step itself.

The motivations for adopting pseudonyms include adhering to ethical and legal guidance on return of clinically relevant results, the needs of developing a data quality system where resources for complete "at source" testing are not available and the need to integrate future clinical records into the same patient record in the secondary domain also considering inter-clinic patients.

Available pseudonym generation systems can roughly be classified depending whether they are reversible or not and the number of steps taken to generate the pseudonym (see Appendix A). One-way pseudonyms are suited for the provision of follow-up data at a later point in time, the withdrawal of samples or data upon specific request by a patient, and quality control of the data such as the checking for double entries. In essence, all requests or data relevant to a particular person that are initiated at the source of the clinical data can undergo the same pseudonymisation process to receive the same pseudonym as any previous or subsequent data for that individual.

Reversible pseudonyms allow a flow of data in a converse sense. The classical examples here are the contacting of those individuals for whom the research activities identify a clinically relevant result, and the correction of a data inconsistency that requires action by the source centre to modify or correct the records of a specific individual.

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

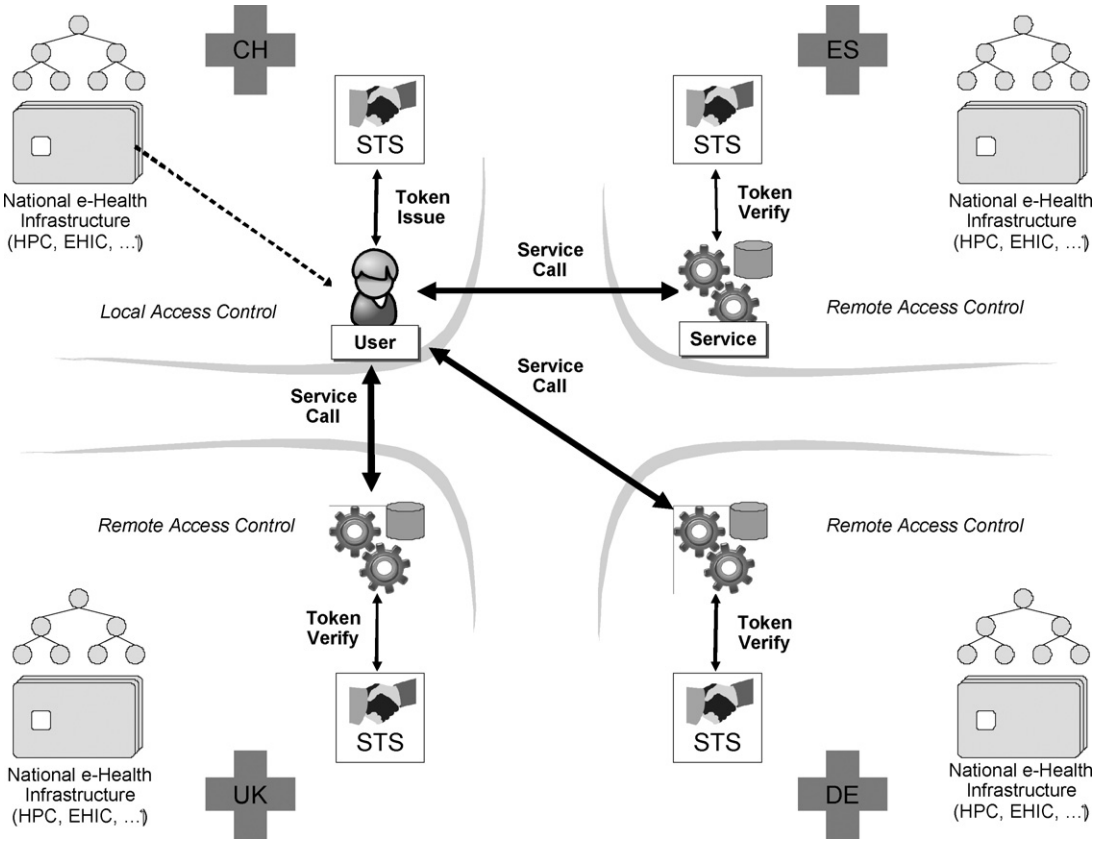

Fig. 3 – Hybrid Access Control Model.

A number of strategies exist for the generation and reversal of pseudonyms, a brief overview of which is provided as an Appendix A (a more rigorous review can be found in Ref. [42]). Pseudonyms may be produced in one step (or single-pass) by either the clinical centre or through a central Trusted Third Party (TTP) service. In either case, sufficient unique PII about the patient is required from which to generate the pseudonym. For multi-institutional studies this step has to be performed by a TTP in order to obtain pseudonyms, which are universal to the study and uniquely link all information on one patient. In order for a TTP to generate such pseudonyms however, the requisite PII must pass from the clinical centre to the TTP.

Within the @neurIST test-bed, patients are treated in distinct medical institutions and the case of inter-clinic patients does not arise. The approach to pseudonym generation implemented in the test-bed reflects this by relying on local creation of reversible pseudonyms (Fig. 2). The standard symmetric encryption algorithm AES [61] is used to generate the pseudonym. Each participating clinical centre has a unique secret key for the @neurIST study which is only known to the particular centre. The pseudonyms are then generated by encrypting the Patient ID (PID) used by the clinic centre for their patients with their unique and secret @neurIST study key. In addition to the encryption of the PID, integrity protection is incorporated into the pseudonym in order to have a proof that the pseudonym is unaltered before reverting it. The reversal of the pseudonym back to the PID can only be performed by the clinical centre holding the proper decryption key.

To support future inter-clinic patients in the @neurIST reference architecture a dual-pass pseudonymisation scheme was developed, which avoids the need to have PII sent to a central TTP pseudonymisation service in order to obtain pseudonyms within a particular study [58]. Instead, the first pseudonymisation step is performed locally and the resulting pseudonym, which does not contain any PII, passed to a TTP that performs a second pseudonymisation step to produce the unique, inter-clinic pseudonym. The proposed scheme enables a multi-centric universal pseudonymisation, meaning that a patient's identity will result in the same pseudonym, regardless of which participating study centre collects the patient data. The developed scheme is based on the discrete logarithm problem (DLP), a number theoretic problem commonly used to construct asymmetric cryptosystems [59]. A more in-depth introduction to the underlying cryptography as well as the developed approach can be found in [60].

3.2.5. *Access restriction*

From a conceptual viewpoint, the access control system for multi-institutional research should follow the common patterns and principles for distributed cross-domain or cross-border information systems, in which heterogeneous environments of multiple distinct institutions need to be interconnected. Generally, clinical centres have their own security system, access rights management and privacy protection pol-

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

icy according to the role of the user [60]. Additionally, they have the know-how concerning data access and communication using standards such as HL7 (Health Level 7), DICOM, IHE (Integrating the Healthcare Enterprise), or business components such as Web Services [61].

Hybrid security models underlie the combination of local and distributed models across various security domains and their different (and often heterogeneous) security policy implementations. In other words, within a security domain all the security is concentrated and placed under the responsibility of this domain whereas between different security domains local credentials are mapped to inter-domain credentials that can be exchanged. Fig. 3 shows such a distributed architecture consistent with the health application domain. Here, the local security model relies on national e-health infrastructures such as Health Professional Cards (HPC) and local authorization systems to obtain a security claim from the local STS (Security Token Service). This claim can be used to access the distributed services residing in a distinct security domain or even in a distinct country.

The approach chosen for the @neurIST project consists of creating a designated security entity in each domain that will be in charge of issuing and verifying short-term security tokens with the entities of the other security domains. This entity is known as Security Token Service.

The attributes contained in security claims have a system-wide scope, since the claims are accepted as a valid measure for user authentication and the attributes are recognized throughout the system and are used in the attribute based access control to resources. The end-users' attributes currently consist of their role inside @neurIST (e.g. clinician, researcher, nurse, etc.).

In detail, the authentication and authorization process includes the following steps:

1. Steps performed at the client-side (locally)
   (a) request an @neurIST token from the local STS,
   (b) incorporate the @neurIST token into the service request,
   (c) call the service.
2. Steps performed at the service-side (remotely)
   (a) call the remote STS for @neurIST token verification,
   (b) verify the service request,
   (c) check the authorizations (role based access control, RBAC),
   (d) invoke the service.

The client requests a security claim from the local STS using mechanisms defined in the WS-Trust [62] standard specification. This request contains the originator's username and the desired relationship for the claim. The STS retrieves which attributes the user was assigned to (for the given relationship) and issues an SAML [63] Attribute Assertion containing these attributes and the user's certificate. The certificate is the one used by the originator inside the respective institution for local access control. This may be for example the certificate for the key stored on the HPC. Finally, the whole token is signed by the STS.

The SAML token uses the "holder-of-key" confirmation method. This means, that the token's assertion is valid (within the token lifetime of 10 min) for anyone proving possession of the private key corresponding to the contained public key certificate. This has the advantage that the STS does not necessarily need to perform access control for token issuing. Anyone can request a token, but only the holder of the private key can make use of the issued token for authentication and authorization, since only he is in possession of the private key, which is required to pair the security token to the service request. The following XML fragment shows an example of such a SOAP message request containing WS-Security [64] conforming security elements.

The client creates such a request for an @neurIST service (e.g. starting an aneurysm simulation) and embeds the SAML token inside the SOAP message header. The SOAP message body is signed by the user's private key (by making use of the HPC). This pairs or binds the security token and the SOAP message together, since only if the signature of the SOAP body can be verified with the certificate contained in the security token, the message will be considered as authentic.

The signed SOAP message containing the relationship token is then used to call the service. The STS on the service-side first checks whether the SAML token was issued by one of the trusted STS belonging to the addressed relationship, i.e. the SAML token is signed by one of the STS certificates exchanged at contract signing (see Fig. 3). If the token could be verified, the contained attributes and especially the contained user certificate can be considered as authentic. The @neurIST service then verifies the SOAP message signature with the obtained user certificate and thus checks, if the "holder-of-key" confirmation is fulfilled. Additionally, the verification checks include the timestamps and other context information such as the validity of the relationship and so forth. Finally, a policy enforcement point checks if the attributes present in the SAML token entitle the user to access the requested resource. @neurIST uses XACML [65] for its policy framework.

It is important to note that although the credentials are issued to users and administered by the local entity to which they belong, the access to data and compute services is controlled based on policies set by the entities that own them. Therefore, even if a local entity issues a credential to its users that enable them to use an @neurIST service, the authorization decision to allow such a use is finally controlled by the policies set by the entity owning the service.

3.2.6. *Security evaluation*

A final step in the process of rendering data reasonably anonymous is to review and monitor each link in the chain of steps for weakness, and taking appropriate risk reduction steps. This is part of the WHO guidelines [15], and should be performed independently of the development of the process.

Within the @neurIST project, the vulnerabilities are in:

- dependence on trust and contractual obligations of those with access to the source data,
- possibility of cracking the pseudonym generation mechanism,
- possibility of establishing an indirect identification; in small communities, any rare characteristic or rare combinations might permit one to identify a person, necessitating a more aggressive approach to stripping of data items than in a

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

larger community; hospitals equipped to undertake the specialized treatment of aneurysms generally have rather large catchment populations so that, in general, fewer characteristics will need to be excluded from the records; however, it may depend on the frequency of a certain characteristic in the community; in any case, it must be remembered that for a database that prospectively follows the course of a patient's clinical history, data entering the record at a later point could contribute to identification by creating a rare combination of data values, misuse of data access — password/id sharing, etc.

4. Specific challenges encountered in @neurIST

This section details particular challenges that came up during the development of the @neurIST project, as well as the preparation and early data collection for the @neurIST study. Not all of these are universal in nature as the @neurIST study, in addition to being a research study in its own right, is a test-bed for the systems and tools developed contemporaneously in the @neurIST project. As such, some issues encountered in the course of the study may reflect the constraints of decisions made in achieving the aims of the wider project.

4.1. *Handling of sensitive data*

Not considered in the process of crop-list generation for the CRIM is that some data items of clinical or research interest may be of a sensitive nature in a given societal setting, including income, religion, heritage (nationality, and ethnic background), details of contraceptive or non-prescription drug use, and information on some co-morbidities such as HIV/AIDS. This point was not commented on by any of the ethical committees who approved the study. However, the research personnel undertaking CRIM-based interviews for data collection in clinical centres where data collection was performed in parallel to the clinical workflow identified the awkwardness for the participant and themselves that these questions could raise.

In the interest of better understanding these issues, the researchers making secondary use of the CRIM data were asked to provide justifications for the collection of these data. For contraceptive use, drug use, and co-morbidities, the justifications were accepted by the clinical research staff in full. In respect to income, the relevance of this item as a factor in healthcare access (epidemiology) was accepted, and together with the use of fairly large categories, the question was rendered sufficiently nonintrusive to be similarly accepted. With respect to ethnicity, while the cases made for use in genetics were generally accepted, one site has maintained a policy of not asking this question, in part because there remains debate in the genetics community over how to define and uniformly capture this item [66]. This debate is all the more relevant in light of the viewpoint in some countries (e.g. France) where the need to collect information on ethnicity is questioned.

4.2. *Free-text data*

Free-text data are common in medical records, and many such documents still kept in hand-written form and not readily useful in the database structure of @neurIST. As clinical information and reporting systems become more common, if not universal, the potential exists to consider these as a further part of the clinical record for digital inclusion in secondary use. As with other unstructured data, there is the potential that free-text fields may contain personal information.

Due to the widespread use of free-text in clinical practices, there is nonetheless interest, particularly from a data-mining perspective, in having access to clinical reports for further analysis. Different approaches are proposed in the literature for automatically removing identifying information in free text. Uzuner et al. [73] have evaluated several methods and have shown that some of them are close to removing all the private health information. The position within @neurIST is that free-text fields are not to be released into the database without justification. This has meant configuring the data collection software to avoid the use of free-text.

4.3. *Architecture-related considerations*

As described in the introduction, a central design concept underlying the information technology development in the @neurIST project is its reliance on a distributed architecture for storage of clinical data, including medical images, i.e. each participating clinical centre is managing the clinical data of patients recruited locally. Ideally, the storage would be in the original clinical information systems used during the care of the patients (as in one clinical centre taking part in the study), with derived data (e.g. the results of a research driven genetic or proteomic investigation, or of a computational fluid dynamics simulation) stored in separate data repositories at the centre where the data is generated.

We have found, however, that most clinical centres are resistant to such an implementation. In response, two models of clinical data storage are supported in the @neurIST project: (1) an anonymised model where the data are depersonalised and stored in a repository accessible for researchers (DB model), (2) an on-the-fly (OTF) model where the live Clinical Information System (CIS) is chosen to store the entire @neurIST data [67]. In the first model, the clinical data are stored in a relational database based on the data model described above. In the second model the data are structured and stored based on the CIS storage system and the data are depersonalised on the fly prior to leaving the hospital intranet [7,69]. The data held with these two models are accessible using a single query with a uniform query interface (using OGSA/DAI — Open Grid Services Architecture/Data Access Interface) and have the same depersonalisation level.

For images in the OTF model, the computational demand of churning large numbers of large data sets has led to the adoption of the pseudonymised images being stored in the DMZ after a first query ready for any future query.

The architectural choice for database federation rather than consolidating from an array of sources to a single monolithic database gives considerable control to the sites of origin of the data. It also limits the abuse potential by making only

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

part of the data available to a potential intruder at any time. This strategy comes with several costs. First, it places greater technical demands on the clinical centres for implementation and maintenance of the system. Second, there is little scope for redundancy of data, such that if a site is off-line at the time a request is made, the response from that site is not accounted for, or filled in with data from a prior time point. This places the end-user at risk of appearing to receive inconsistent responses to the same query even if the data in the servers has not changed. Still, such off-times can be expected to become rarer in the future with many services relying on online available services.

Using the OTF model has the additional advantage that always the latest data from the patient record is used thus limiting potential data inconsistencies when data of a patient are corrected.

4.4. *Data quality*

As with any research study, a persistent concern is data quality. The federated data source design conceived in @neurIST requires close liaison with the clinical centres hosting the data sources, as they must effect any revisions and corrections to the data if these are to be persistent (i.e. not over-written with erroneous data in a later export, or query).

A significant consideration is that while the clinical centres receive financial support from the project for the required server system and manpower to carry out data collection these funds are limited and focussed on recruitment and collection of an adequate number of blood samples with accurate and adequate clinical history being subsidiary in order to carry out the planned genetics study. The funds are explicitly not intended to, nor are they sufficient to carry out additional testing (e.g. MRI screening of controls, or tests for genetic disorders or co-morbidities) by the clinics. By the same measure, @neurIST has no means to provide financial incentives for good quality data collection by the clinical centres involved, nor conversely to discourage the collection of poor quality data. Thus, we are reliant on the clinical information routinely available in the course of the patient's clinical care.

That said the data obtained follows each sites consideration of best practices in routine clinical record-keeping. Moreover, there is some scope for fostering and elevating the data quality achieved. Firstly, when the data source is in fact the hospital information system (as is the case in one of the partner sites) and not merely an ancillary research-only dataset, errors in the source data are errors in the clinical record of the patient. The potential for these errors to cause confusion or to lead to a medical mishap provide a strong motivation for the host institution and the members of the clinical team to be watchful for and take steps to correct errors. The on-going use of the clinical records over the course of the patient's care also mean that many "sets of eyes" pass over the data in connection with the patient, raising the possibility of at least significant errors or gaps in the record being spotted. The formulation of the CRIM, together with the identification of risk factors within the CRIM from literature and work-practice reviews provide a further means of fostering good data collection practices. In the review of the crop-list for release of clinical data, the end-users of secondary data in @neurIST were unanimous in requesting the risk factors and phenotype information from the CRIM be made available for analysis. It is precisely this same information, which is most useful in clinical decision-making. Thus, by encouraging universal collection of those items recognized as affecting clinical decision-making, both clinical practice and secondary research use benefit.

The moratorium on supporting additional scanning or testing protocols from the project funds also meant that specifications for scanning or screening could not be demanded of the clinical centres. This is contrary to the needs of most secondary users of clinical data where minimum requirements (resolution in images, traits covered for phenotype characterisation, etc.) exist for successful testing or accurate results to be obtained. This disparity is a weakness of any system seeking to use routine clinical information to support secondary use in research, and must not be seen as specific to the @neurIST system or strategy. It should be emphasized that while the @neurIST system could perhaps be used to support controlled clinical trials (e.g. by pre-populating fields based on clinical records prior to the research nurse performing a structured interview, and in depersonalising data to be made available to the trial controller), that is not its primary objective. Our express purpose is in making routine clinical information available for secondary use. In this section we describe some of the procedural approaches to maximizing data quality achieved.

4.4.1. *Education*

In the course of initiating the study, the staff members from each centre responsible for recruitment, data collection, and entry were brought together for two training sessions. The first covered the recruitment process, informed consent, and blood sample collection. A preliminary form of the questionnaire for gathering CRIM data items to a paper record was necessary at this point The second provided an opportunity for feedback on experience to date, and introduced the digital data collection tool used for collection at most sites, and the extensions to the routine clinical information system at the site where this is used directly. This session also provided an opportunity to review and provide explicit clarification of points of confusion regarding terminology and practice.

A concern for the fully embedded clinical collection strategy is its reliance upon a network of busy personnel entering data in the clinic. This can only be addressed by appropriate training and awareness being developed within the clinical centres.

4.4.2. *Typographical/transcription errors, invalid entries and misinterpreted fields*

In the course of the study, most of the centres will be working from existing paper clinical notes or printed questionnaires completed at the bedside and then transcribing them to the @neurIST data entry software. Each of these steps can be expected to introduce typographic errors. One site however will not follow this route, and in fact the data collected comes directly from the electronic health record limiting transcription errors.

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

The data collection and preparation tools used by the clinical centres do not check for the validity of data beyond the data type which is specified for an entity in the CRIM. This raises the possibility for instance that although a treatment date is entered, it is not realistic (e.g. treatment prior to birth date). Similarly, there may be differences in interpretation between users or even centres regarding the meanings of some questions. The data collection interface includes links to an ontology and balloon help for each item to aid reducing this inconsistency in recording data. While it would be preferable to catch these types of errors at data entry, this is not possible within the present study. Thus we are reliant on post-hoc cleansing to catch such errors.

4.4.3. *Linkage errors*

In the passage of the data from the clinical domain via the DMZ to the researcher in response to database queries, we are reliant on the depersonalisation and transfer processes to maintain the continuity of an individual participant record. The study participant UID, determined based on a hospital code is the basis for this link, and for all further links of derived data. If the hospital code is changed (e.g. due to a correction), a participant whose data was formerly associated with one UID will afterwards be associated with a new UID. In the OTF approach, future queries will only receive results based on the latter UID. Analyses and derived data associated with the former UID however will be orphaned from any subsequent updates of data. This creates a risk of re-representing the same individual. Much the same happens more immediately in the DB model, where updates of the DMZ records will in fact leave the record originally exported under the first UID and an updated replicate of that data under the second UID. Because images under the OTF model are exported once and stored to avoid load in responding to later queries, the same applies to these. Removing these double entries may be possible with a periodic clearing of the DMZ servers and full export. However, the effort required for doing this may be prohibitive unless all the images are kept within the hospital, or are being retrieved automatically from the PACS system. A more efficient and elegant approach making use of the XML record for each individual, and when a change of hospital code is identified, updating the DMZ records can be envisaged. This would involve removing the original clinical data in the DMZ and its images re-linked to the new UID. A further mechanism is needed for propagating this change in UID to derived data.

A linkage step also exists in the connection of the analysis results derived from physical samples (in the present study: blood samples) to the phenotype, clinical history and haemodynamic analyses. To avoid the risk of transcription errors here, a barcode is fixed to each sample and its code recorded in the original data file. This code is exported in encrypted form along with the participant's other data. Only the genetics and transcriptomics testing groups have the key for decrypting this field.

The creation of the CRIM marked a new process for cerebral aneurysms by integrating all information necessary to gather on a patient to fully describe their health journey. The CRIM involves a large number of data items, making it awkward and time consuming to bring together manually as it is being performed at most of the clinical centres (estimates from two centres put the recruitment and collection time at 2–3 h per participant). Where the CRIM is populated directly from the clinical information systems this time is distributed over the course of the patient's care, and it does not represent a duplication of effort. This is particularly attractive for researchers as it shows the potential of the system to contribute to reducing data collection costs.

A critical consideration however, is that many of the items may not be available within the EHR, and depending on local practices, it may or may not be routine to collect and record any particular item. To some extent, we anticipate this being a target for progressive improvement by encouraging clinical best practices. Within the @neurIST study, the completeness of records being submitted is to be monitored to identify the common practices and variations in practice relative to the CRIM. In order to support the planned secondary use research, the sites will receive feedback on incomplete data collection specifically for the priority data items of risk factors identified in a Cochrane review, and on primary phenotype traits such as rupture status, aneurysm size, location, shape, age, gender and ethnicity. Where gaps are identified, the pseudonym and missing priority item(s) will be reported to the source clinical centre in a request for completion. As noted however, there is no obligation on the source centre to carry out this process.

5. Conclusion

This article presents the complexity of following legal regulations in multi-centre medical research projects involving secondary use of medical data with, as a concrete example, the strategy taken in the @neurIST research project that includes seven clinical centres in five European countries, as well as a further 25 non-clinical partners. Secondary use of medical data is currently a topic of attention as the data could yield considerable information that can be used to improve effectiveness and efficiency of health processes. Such multi-centre studies allow, for example, obtaining a statistically meaningful number of cases quicker than studies in a single centre, so the applied research results can be used quicker as well and particularly for rare diseases a critical mass of cases can be obtained in sufficient quality that no single institution could obtain. On the other hand it is important to protect the private data of the patient, and that across the sites, the data conform to a uniform set of standards. Legal regulations vary in European countries and thus studies over several institutions have to take into account a large number of legislations and different situations. It becomes clear that a global security strategy for the project had to be taken into account.

In @neurIST, and likely many other studies making secondary use of clinical data, the possibilities for anonymisation in the strict sense of the Data Protection Directive are limited by the persistence of the source medical record and the importance to the research use of the individual personal data. A more achievable goal therefore is that of reasonable anonymity as defined by the WHO [15]. The criteria of reasonable anonymity allow for useful mechanisms such as patient re-identification for the purpose of relaying relevant research findings, but do so with the premium of demanding careful review of the data protection process. Greater clarity and con-

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

sistency across countries and contexts in defining the legal framework for secondary data use would serve the global community. Highlighting this point is one of the goals of this article.

De-identification, data minimisation, aggregation and pseudonymisation feature in the @neurIST privacy protection strategy for reaching reasonable anonymity. Each stand of any such strategy must be both workable within the context of the study, and provide a high level of protection against direct and indirect identification if the target of reasonable anonymity is to be achieved.

In keeping with good practice in the consent process, it is important to better inform the patients who participate in a study about the security process, so they can have a choice regarding the use of their data. Explaining them the risks and potential benefits of research studies is thus an important part of a project's security strategy. Unfortunately, the diversity and subtlety of terms currently in use for privacy protection do not foster the ease of communication required for many patients. A limited vocabulary, used consistently would facilitate uniformity across studies and in the understanding of risks by individuals who, at best, are likely to have a rudimentary grasp of the topics. Careful use of terms in other contexts, including ethical, legal and procedural documents and discussion is also to be recommended.

Large multi-centre research studies will surely become more common in a Europe that wants to be a leader in clinical research. Having clear ideas on the legal constraints to be taken and the technologies to use to satisfy the requirements of law can help other projects that will start in similar situations.

Conflicts of interest

There are no conflicts of interest.

Acknowledgements

This work was generated in the framework of the @neurIST Integrated Project, which is co-financed by the European Commission through the contract no. IST-027703. The authors would like to thank Ivan Periz and Professor Deryck Beyleveld for many fruitful discussions.

Appendix A. Pseudonymisation

A.1. *One-way single-pass*

One-way pseudonymisation systems generate pseudonyms, which are not reversible per-se. (Pseudo-)Random number generators [61] as well as keyed cryptographic one-way or hash functions [68] (also referred to as one-way encryption in this context) are common technologies to implement this property.

The main advantage of such pseudonym generation functions is the high level of privacy protection they provide. The natural disadvantage is that it does not allow re-contacting the patients in case relevant findings are made during the course of a study (to enable the reverse mapping with such schemes, additional mechanisms such as a mapping database are required, effectively making them reversible pseudonym generation schemes, which are discussed below). Some technical realizations, such as the usage of a random number generator to generate the pseudonym preclude the addition of follow-up data during the course of a study. Specifically, while the entire updated record, including the past content and follow-up could be sent anew, there would not be a way to selectively remove any previously released records for that individual, or to link previously derived secondary usage results to the newly pseudonymised clinical record as they would have a different pseudonym.

To avoid this shortcoming, keyed hash functions may be used instead. An essential prerequisite here is some form of PII, which identifies the patients unambiguously and is used as input to the pseudonym generation function. New information on the same patient would produce the same pseudonym and therefore enable updates to the data.

A.2. *One-way dual-pass*

In order to prevent the provision of PII of patients to external parties even though they are trusted for a certain purpose such as the creation of pseudonyms, the pseudonym generation step can be performed twice: once at the clinical centre and once at the TTP. Thus, the TTP only receives a pseudonym and uses this in turn as input to an additional pseudonym generation step resulting in the universal pseudonym being generated. Rather than return the universal pseudonym, the TTP can attach it to the de-identified data to avoid the situation of members of the clinical centre having direct access to both the clinical and research domain identifiers. Equally, having a single TTP service for all the clinical centres simplifies the task of ensuring unique links through all patient data in a multi-institutional study. Such schemes are also termed dual-pass or dual-coding schemes.

A.3. *Reversible single-pass*

Reversible pseudonymisation systems allow the patient to be re-identified from its pseudonym to enable re-contacting a patient in case clinically relevant findings arise through the course of the study. Conceptually, the simplest way to enable re-identification is to make use of a one-way scheme supplemented by some form of mapping database which stores the association between PII and pseudonyms. Such a mapping database is, however, an attractive target for attackers and violates the medical secrecy [52]. Thus, additional safeguards need to be in place in order to protect the sensitive mapping database. Encryption of the mapping database with a proper key management system as well as careful configuration of the underlying system on which the mapping database is used (e.g. to prevent clear-text portions of the mapping data being swapped to temporary storage) are commonly deployed for this task.

Alternative approaches rely on cryptographic mechanisms only and do not require a mapping mechanism. In the case of symmetric cryptosystems – which use one and the same key for encryption and decryption – the entity holding the key

can generate pseudonyms by encrypting some PII and retrieve this PII again back from the pseudonym by decryption. This creates a level of security very much like the patient/clinician confidentiality relationship. If the clinical centre holds such a key, it is able to re-identify participating patients.

To ensure that re-identification of individuals is strictly controlled, asymmetric cryptosystems may be used instead of symmetric ones, where the public encryption key is used to generate the pseudonyms, but the private decryption key is, e.g. only known to and in possession of the study's medical and ethical advisory board [69]. Thus, the usage of asymmetric encryption schemes allows for a separation of duties.

A.4. *Reversible dual-pass*

As for the one-way pseudonym generation, a second pseudonymisation step can be introduced in such a way that the two pseudonymisation procedures are independent from and do not know anything about each other. This "additional privacy safeguard" allows for much stricter control over linking research information back to a patient. The situations warranting working back through the two pseudonyms can be regulated by strict operating procedures. In single centre studies the first pseudonymisation step can be performed within the study centre [70]. For multi-centric studies, however, both have to be generated by distinct TTPs in order to obtain the correct linkage between data from different sources.

LIPA (Linkable Identity Privacy Algorithm) is an example of such architecture [71]. It relies on asymmetric cryptographic techniques and two distinct TTPs in order to calculate a universal pseudonym. However, the first TTP receives the patient's National Health Service (NHS) number which reveals PII to the TTP since the NHS number is the common and unique identifier for patients in England and Wales.

An improvement to LIPA has been developed within @neurIST, which allows computing a universal pseudonym without the need to reveal PII to a TTP outside the participating clinical centre [60]. The approach relies on asymmetric cryptographic techniques and one TTP to calculate this study-global pseudonym. A clinical centre first computes a local pseudonym using a study- and clinic-specific private key. Only the TTP is afterwards in possession of the required additional keys to compute the universal pseudonym. Note that the same patient in distinct clinics will result in distinct local pseudonyms but in one unique universal pseudonym. In-depth details on the underlying number theoretic and cryptographic foundations of the developed scheme are available in [60].

REFERENCES

[1] American Medical Informatics Association (AMIA), A Taxonomy of Secondary Uses and Re-Uses of Healthcare Data, 2007 Invitational Conference on Secondary Use of Health Data, September 2007. Available online at: http://www.amia.org/inside/initiatives/healthdata/2007/index.asp (last accessed 21 October 2009).

[2] H. Rajasekaran, P. Hasselmeyer, L. Lo Iacono, J. Fingberg, P. Summers, S. Benkner, G. Engelbrecht, T. Arbona, A. Chiarini, C.M. Friedrich, M. Hofmann-Apitius, B. Moore, P. Bijlenga, J. Iavindrasana, H. Müller, R.D. Hose, R. Dunlop, A. Frangi, K. Kumpf, @neurIST—towards a system architecture for advanced disease management through integration of heterogeneous data, computing, and complex processing services, in: Proceedings of 21 IEEE International Symposium on Computer-Based Medical Systems 2008, June 2008.

[3] V.L. Feigin, C.M. Lawes, D.A. Bennett, C.S. Anderson, Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century, Lancet Neurol. 2 (1) (2003) 43–53.

[4] V.L. Feigin, G.J. Rinkel, C.M. Lawes, A. Algra, D.A. Bennett, J. van Gijn, et al., Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies, Stroke 36 (12) (2005) 2773–2780.

[5] W.I. Schievink, Intracranial aneurysms, N. Engl. J. Med. 336 (1) (1997) 28–40.

[6] J.I. Suarez, R.W. Tarr, W.R. Selman, Aneurysmal subarachnoid hemorrhage, N. Engl. J. Med. 354 (4) (2006) 387–396.

[7] J. Iavindrasana, A. Depeursinge, P. Ruch, S. Spahni, A. Geissbuhler, H. Müller, Design of a decentralized reusable research database architecture to support data acquisition in large research projects, Stud. Health Technol. Inform. 129 (2007) 325–329.

[8] R. Dunlop, A. Arbona, H. Rajasekaran, L. Lo Iacono, J. Fingberg, P. Summers, S. Benkner, G. Engelbrecht, A. Chiarini, C.M. Friedrich, B. Moore, P. Bijlenga, J. Iavindrasana, R.D. Hose, A.F. Frangi, @neurIST—chronic disease management through integration of heterogeneous data and computer-interpretable guideline services, in: Proceedings of HealthGrid, June 2008.

[9] The Association of Electrical and Medical Imaging Equipment Manufacturers (NEMA), DICOM Part 3: Information Object Definitions. PS 3.3-2008. Available at: ftp://medical.nema.org/medical/dicom/2008/08_03pu.pdf (last accessed 21 October 2009).

[10] European Parliament and Council, Directive 95/46/EC of the European Parliament and of the Council of October 24, 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, Off. J. Eur. Commun. No. L. 281/31 (1995).

[11] Article 29 Working Party, Opinion No. 4/2007 on the concept of personal data, WP 136 (2007).

[12] B. Knoppers, M. Saginur, Correspondence: the Babel of genetic data terminology, Nat. Biotechnol. 23 (8) (2005) 925–927.

[13] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), Technical Standard on Health Informatics: Pseudonymisation Practices for the Protection of Personal Health Information and Health Related Services, ISO/TS 25237 (2008).

[14] A. Pfitzmann, M. Hansen, Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management—A Consolidated Proposal for Terminology, February 2008. Available online at: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml (last accessed 21 October 2009).

[15] World Health Organization (European Partnership on Patients' Rights and Citizens' Empowerment), Genetic Databases, Assessing and the Impact on Human Rights and Patient Rights, World Health Organization, Geneva, 2003.

[16] Swiss Academy of Medical Sciences, Biobanks: obtainment, preservation and utilisation of human biological material—medical-ethical guidelines and recommendations, Journal of International Biotechnology Law 4 (5) (2007) 210–214.

[17] Human Genetics Commission, Whose hands on your genes? A discussion document on the storage protection and use of personal genetic information, Human Genetics

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

Commission, London, 2000. Available online at: http://www.hgc.gov.uk/UploadDocs/DocPub/Document/business_consultations 2maintext.pdf (last accessed 21 October 2009).
[18] European Parliament and the Council, Directive 2001/20/EC relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use, Off. J. Eur. Commun. L121/34 (2001).
[19] World Medical Association (WMA), Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects (WMA General Assembly, Seoul, October 2008). Available online at: http://www.wma.net/en/30publications/10policies/b3/index.html (last accessed 3 August 2009).
[20] Council of Europe, Additional Protocol to the Convention on Human Rights and Biomedicine, Concerning Biomedical Research, ETS 195 (2005).
[21] Council of Europe Committee of Ministers, Recommendation No. R (97) 5 of the Committee of Ministers to Member States on the Protection of Medical Data (1997).
[22] Council of Europe Committee of Ministers, Recommendation Rec(2006)4 of the Committee of Ministers to member states on research on biological materials of human origin. Available online at: https://wcd.coe.int/ViewDoc.jsp?id=977859 (last accessed 21 October 2009).
[23] Council for International Organizations of Medical Sciences (CIOMS), International Ethical Guidelines for Biomedical Research Involving Human Subjects, CIOMS, Geneva, 2002. Available online at: http://www.cioms.ch/frame_guidelines_nov_2002.htm (last accessed 21 October 2009).
[24] OECD, Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, OECD, Paris, 1980.
[25] D. Beyleveld, D. Townend, S. Rouillé-Mirza, J. Wright (Eds.), See The Data Protection Directive and Medical Research Across Europe, Ashgate Publishing Ltd., Aldershot, 2005.
[26] Medical Research Council (MRC) (UK), Human Tissue and Biological Samples for use in Research—Operational and Ethical Guidelines, MRC, London, 2001.
[27] E. Zika, T. Schulte In den Baumen, J. Kaye, A. Brand, D. Ibarreta, Sample, data use and protection in biobanking in Europe: legal issues, Pharmacogenomics 9 (6) (2008) 773–781.
[28] A.C. Da Rocha, J.A. Seoane, Alternative consent models for biobanks: the new Spanish law on biomedical research, Bioethics 22 (8) (2008) 440–447.
[29] H. Nys, G. Fobelets, The regulation of biobanks in Spain, Law Hum. Genome Rev. 29 (2008) 169–188.
[30] J.A. Seoane, A.C. Da Rocha, Consentimiento, biobancos y ley de investigacion biomedica (Consent, biobanks and biomedical research law), Law Hum. Genome Rev. 29 (2008) 131–148.
[31] B.J. Clark, Tissue banking in a regulated environment—does this help the patient? Part 1. Legislation, regulation and ethics in the UK, Pathobiology 74 (4) (2007) 218–222.
[32] Swiss Federal Law on Genetic Analysis of Human Beings (Loi fédérale sur l'analyse génétique humaine) (8.10.2004). Available online at: http://www.admin.ch/ch/f/rs/810_12/index.html (last accessed 21 October 2009).
[33] Council of Europe, Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine (Convention on Human Rights and Biomedicine) Oviedo, ETS 164 (1997).
[34] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, Carnegie Mellon University Laboratory for International Data Privacy, Pittsburgh, 2000. Available online at: http://privacy.cs.cmu.edu/dataprivacy/papers/LIDAP-WP4abstract.html (last accessed 21 October 2009).
[35] National Institutes for Health (US), De-identifying Protected Health Information Under the Privacy Rule, 2007. Available online at: http://privacyruleandresearch.nih.gov/pr_08.asp#8a (last accessed 21 October 2009).
[36] Medical Research Council (MRC) (UK), Personal Information in Medical Research, MRC, London, 2003.
[37] J.E. Lunshof, R. Chadwick, D.B. Vorhaus, G.M. Church, From genetic privacy to open consent, Nat. Rev. Genet. 9 (2008) 406–411.
[38] OECD Guidelines for Human Biobanks and Genetic Research Databases, OECD, Paris, 2009.
[39] UK Data Protection Act, Chapter 29, Office of Public Sector Information, London, 1998.
[40] K. Pommerening, M. Reng, Secondary use of the electronic health record via pseudonymisation, in: L. Bos, S. Laxminarayan, A. Marsh (Eds.), Medical Care Compunetics, vol. 1, IOS Press, Amsterdam, 2004, pp. 441–446.
[41] European Agency for the Evaluation of Medical Products (EMEA) Committee for Proprietary Medicinal Products, Position Paper on Terminology in Pharmacogenomics (EMEA, London, 2002) 4.
[42] J.R. Gulcher, K. Kristjánsson, H. Gudbjartsson, K. Stefánsson, Protection of privacy by third-party encryption in genetic research in Iceland, Eur. J. Hum. Genet. 8 (10) (2000) 739–742.
[43] ISAT Trial Collaborative Group, International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: a randomised trial, Lancet 360 (2002) 1267–1274.
[44] J. Kulynych, D. Korn, The effect of the new federal medical-privacy rule on research, N. Engl. J. Med. 346 (3) (2002), pp. 201-4, 202.
[45] G.J. Annas, Medical privacy and medical research—judging the new federal regulations, N. Engl. J. Med. 346 (3) (2002), pp. 216-20, 217.
[46] A. Stirling, et al., On Science and Precaution in the Management of Technological Risk, The EC Joint Research Centre (JRC), Institute for Prospective Technological Studies, Seville, 1999.
[47] Information Commissioner (UK), Privacy Impact Assessment (PIA) Handbook (version 2), Information Commissioner's Office, Wilmslow, 2009.
[48] B. Malin, L. Sweeney, How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems, J. Biomed. Inform. 37 (2004) 179–192.
[49] Common Criteria, The Common Criteria for Information Technology Security Evaluation v3.1, July 2009. Available online at: http://www.commoncriteriaportal.org (last accessed 21 October 2009).
[50] P. Hunter, N. Smith, J. Fernandez, M. Tawhai, Integration from proteins to organs: the IUPS Physiome Project, Mech. Ageing Dev. 126 (1) (2005) 187–192.
[51] S. Hanser, M. Boeker, K. Kumpf, P. Bijlenga, S. Schulz, Design of an ontology on cerebral aneurysms: representing the conceptual space of the @neurIST project, in: Proceedings of MedInfo, 2007.
[52] NEMA, DICOM Supplement 55: Attribute Level Confidentiality (including De-identification), February 2000. Available online at: http://medical.nema.org/Dicom/supps/sup55_03.pdf (last accessed 21 October 2009).
[53] R. Noumeir, A. Lemay, J.M. Lina, Pseudonymisation of radiology data for research purposes, J. Digit. Imaging 20 (2006) 284–295.
[54] P. Ruch, R.H. Baud, A.M. Rassinoux, P. Bouillon, G. Robert, Medical document anonymization with a semantic lexicon, in: Proceedings of AMIA Symp., 2000, pp. 729–733.
[55] A.W. Kamauu, S.L. Duvall, D.E. Avrin, Using Java to generate globally unique identifiers for DICOM objects, J. Digit. Imaging (2007).

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001

[56] A.M. Burton, S. Wilson, M. Cowan, V. Bruce, Face recognition in poor-quality video: evidence from security surveillance, Psychol. Sci. 10 (1999) 243–248.
[57] A. Bischoff-Grethe, I.B. Ozyurt, E. Busa, B.T. Quinn, C. Fennema-Notestine, C.P. Clark, S. Morris, M.W. Bondi, T.L. Jernigan, A.M. Dale, G.G. Brown, B. Fischl, A technique for the deidentification of structural brain MR images, Hum. Brain Mapp. 28 (9) (2007) 892–903.
[58] L. Lo Iacono, Multi-centric Universal Pseudonymisation for Secondary Use of the EHR, in HealthGrid, Geneva, Switzerland, 2007.
[59] A.J. Menezes, P.C. van Oorschot, S.A. Vanstone, Handbook of Applied Cryptography, CRC Press, Florida, 1996.
[60] C. Lovis, S. Spahni, N. Cassoni-Schoellhammer, A. Geissbuhler, Comprehensive management of the access to a component-based healthcare information system, Stud. Health Technol. Inform. 124 (2006) 251–256.
[61] A. Geissbuhler, C. Lovis, A. Lamb, S. Spahni, Experience with an XML/http-based federative approach to develop a hospital-wide clinical information system, Medinfo 10 (2001) 735–739.
[62] A. Nadalin, M. Goodner, M. Gudgin, A. Barbir, H. Granqvist (Eds.), WS-Trust 1.3 OASIS Standard, March 2007. Available online at: http://docs.oasis-open.org/ws-sx/ws-trust/200512 (last accessed 21 October 2009).
[63] S. Cantor, J. Kemp, R. Philpott, E. Maler (Eds.), Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) Version 2.0 OASIS Standard, March 2005. Available online at: http://saml.xml.org/saml-specifications (last accessed 21 October 2009).
[64] A. Nadalin, C. Kaler, R. Monzillo, P. Hallam-Baker, Web Services Security: SOAP Message Security 1.1 OASIS Standard Specification, February 2006.
[65] T. Moses (Ed.), Extensible Access Control Markup Language (XACML) Version 2.0 OASIS Standard, February 2005. Available online at: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml (last accessed 21 October 2009).
[66] T.R. Rebbeck, P. Sankar, Ethnicity, ancestry, and race in molecular epidemiologic research, Cancer Epidemiol. Biomarkers Prev. 14 (2005) 2467–2471.
[67] J Iavindrasana, L. Lo Iacono, H. Müller, I. Periz, P. Summers, J. Wright, Access to clinical information systems for research in the life sciences: security and privacy considerations, in: Proceedings of HealthGrid, Chicago, 2008.
[68] S. Bakhtiari, R. Safavi-Naini, J. Pieprzyk, Keyed hash functions, in: E. Dawson and J. Golic (Eds.), Cryptography: Policy and Algorithms, Lecture Notes in Computer Science, vol. 1029, 1996, pp. 201–214.
[69] K. Pommerening, M. Miller, I. Schmidtmann, J. Michaelis, Pseudonyms for cancer registry, Methods Inf. Med. 35 (1996) 112–121.
[70] D. Kalra, P. Singleton, D. Ingram, J. Milan, J. MacKay, D. Detmer, A. Rector, Security and confidentiality approach for the Clinical E-Science Framework (CLEF), in Second UK E-Science All Hands Meeting, Nottingham, UK, 2003.
[71] N. Zhang, A. Rector, I. Buchan, Q. Shi, D. Kalra, J. Rogers, C. Goble, S. Walker, D. Ingram, P. Singleton, A linkable identity privacy algorithm for healthgrid, in: T. Solomonides, R. McClatchey, V. Breton, Y. Legré, S. Nørager (Eds.), Proceedings of Healthgrid 2005, Studies in Health Technology and Informatics, vol. 112, 2005.
[72] M. Clarke, Systematic review of reviews of risk factors for intracranial aneurysms, Neuroradiology 50 (August (8)) (2008) 653–664.
[73] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (September–October (5)) (2007) 550–563.

Bernice Elger (MD, PhD, MA), after having studied at the universities of Essen, Bochum, Bethel, Heidelberg (Germany) and Houston (Texas), has been teaching health law and bioethics at the University of Geneva for the past 14 years. In 2004–2005 she was visiting scholar in the University of Pennsylvania School of Medicine Center for Bioethics, Kennedy Institute of Ethics (Georgetown University) and the Department of Clinical Bioethics (NIH, Bethesda). Her research areas include genetic testing and tissue banks and she is member of the Swiss federal commission of experts concerning the Swiss law on genetic testing as well as member of the subcommission of the commission responsible for the Swiss Academy of Medical Sciences' guidelines on the ethical use of human material. She has published widely on issues related to health law and ethics and has received several awards for her academic work, among which was the Bizot Award of the University of Geneva for her PhD thesis (habilitation) on ethical and legal issues related to medical research involving biobanks.

Jimison Iavindrasana studied mathematics and computer science and received his master degree in "network, multimedia and Internet" from the University of Reunion Island, France in 2000. After 4 years in industry, he decided to work in the medical field and obtained a masters degree in medical informatics from the University Paris VI, France. He is currently a PhD student at the University of Geneva, Switzerland.

Luigi Lo Iacono studied computer sciences with a major in systems and security engineering and received his PhD from the University of Siegen (Germany) in 2005. He has previously worked in academic and industry research laboratories and currently holds the position of senior researcher in the Distributed IT Services and Security group with NEC Laboratories Europe.

Henning Müller received his master degree in medical informatics from the University of Heidelberg, Germany, in 1997, specializing in medical signal and image processing. After a scholarship for a stay at Daimler-Benz Research and Technology North America in Portland, OR, USA, he started his PhD in content-based image retrieval at the University of Geneva Switzerland in 1998. After a research stay at Monash University, Melbourne, Australia in 2001 he finished his PhD in 2002. In the same year he started as a postdoctoral research fellow at the medical informatics service of the University and Hospitals of Geneva, Switzerland, where he started the MedGIFT project on medical image retrieval. Since 2007 he has been named professor for business information systems at the University of Applied Sciences Western Switzerland, Sierre, Switzerland.

Nicolas Roduit studied geology and computer sciences at the University of Geneva, Switzerland. In 2007, he obtained an interdisciplinary PhD after specializing in image processing and analysis. In particular, he developed a software application for the specific needs of a major oil company. He then joined the digital imaging team at the Geneva University Hospitals.

Paul Summers studied physics and mathematics prior to obtaining a PhD in medical physics from King's College London. Having supported clinical research in university and hospital departments, he is now focussed on imaging investigations of the spinal cord and CSF.

Jessica Wright is a specialist in research ethics, data protection, genetic databases and biobanks after working for over 6 years on EC-funded projects PRIVIREAL, PRIVILEGED and @neurIST. She obtained an MA in biotechnological law and ethics from the University of Sheffield School of Law, where she currently holds a position as Research Coordinator.

Please cite this article in press as: B.S. Elger, et al., Strategies for health data exchange for secondary, cross-institutional clinical research, Comput. Methods Programs Biomed. (2010), doi:10.1016/j.cmpb.2009.12.001