Interface Focus

rsfs.royalsocietypublishing.org

Research

Cite this article: Jefferys BR, Nwankwo I, Neri E, Chang DCW, Shamardin L, Hänold S, Graf N, Forgó N, Coveney P. 2013 Navigating legal constraints in clinical data warehousing: a case study in personalized medicine. Interface Focus 3: 20120088.
http://dx.doi.org/10.1098/rsfs.2012.0088

One contribution of 25 to a Theme Issue 'The virtual physiological human: integrative approaches to computational biomedicine'.

Subject Areas:
computational biology

Keywords:
clinical data warehouse, personalized medicine, legal and ethical constraints

Author for correspondence:
Benjamin R. Jefferys
e-mail: b.jefferys@ucl.ac.uk

Navigating legal constraints in clinical data warehousing: a case study in personalized medicine

Benjamin R. Jefferys[1], Iheanyi Nwankwo[2], Elias Neri[3], David C. W. Chang[1], Lev Shamardin[1], Stefanie Hänold[2], Norbert Graf[4], Nikolaus Forgó[2] and Peter Coveney[1]

[1]Centre for Computational Science, Department of Chemistry, University College London, London WC1H 0AJ, UK
[2]Institut für Rechtsinformatik (IRI), Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, Germany
[3]Custodix, Kortrijkseteenweg, 9830 Sint-Martens-Latem, Belgium
[4]Saarland University Hospital Faculty of Medicine, Campus Homburg Building no. 9, 66421 Homburg, Germany

Personalized medicine relies in part upon comprehensive data on patient treatment and outcomes, both for analysis leading to improved models that provide the basis for enhanced treatment, and for direct use in clinical decision-making. A data warehouse is an information technology for combining and standardizing multiple databases. Data warehousing of clinical data is constrained by many legal and ethical considerations, owing to the sensitive nature of the data being stored. We describe an unconstrained clinical data warehousing architecture, some of the legal constraints that have led us to reconsider this architecture, and the legal and technical solutions to these constraints developed for the clinical data warehouse in the personalized medicine project p-medicine. We also propose some changes to the legal constraints that will further enable clinical research.

1. Introduction

Personalized medicine promises a revolution in healthcare by moving from a focus upon the disease to a focus upon the individual patient. This vision was inspired by the sequencing of the human genome, and fuelled by the promise of high-throughput sequencing allowing treatment to be tailored according to an individual's genetic make-up. The concept has now extended beyond this to include personalization based upon multiple factors, including details of a person's precise disease phenotype, and even cultural and psychological considerations. Progress towards this goal has been slow, with the lack of data-sharing technology blamed as one of the hindrances [1].

Increased personalization depends intrinsically upon insight from experimentation, data and models. The p-medicine project ([2] and www.p-medicine.eu) is developing infrastructure to bring together all of these elements to feed into clinical decision support systems. At the core of this project is a warehouse to collect and standardize data from clinical trials and patient management systems, for the purpose of meta-analysis, to derive statistics to feed into computational models and to provide source data for decision support. Such data are generally considered to be sensitive, the subject being the health status of individuals who may not want that shared beyond their treating clinician—particularly if there is a possibility that they may be identified personally.

As a result of the sensitive nature of the data involved, an adequate framework must be put in place in order to gain the trust of the data subjects as well as to comply with regulatory requirements. Thus, as well as the constraints of interoperability and standardization of the various data sources in building a data warehouse, legal constraints also profoundly affect a clinical data

warehousing architecture. These can be broadly divided into data protection, data security and ethical constraints.

1.1. Data protection

Health-related data are seen as a special category of personal data, requiring higher protection than other personal data. While the relationship between the patient and the treating physician is seen as fiduciary and generally covered by the consent of the patient to be treated by the physician, a legal basis is required for the further processing of the data generated from this transaction. This will be the case where the data are further used for research or transferred to a third party for marketing purposes. Integrating such clinical data into a data warehousing infrastructure that contains other source data and information technology tools that allow for their mining, for example, may introduce legal issues. Thus, in most cases, the first hurdle is to find a legal basis for additional processing of medical data, especially where the data to be processed are still in a form that could be regarded as personal data. Although consent for this additional processing may suffice, this may be invalid as a result of the 'specific purpose rule' in data protection law that frowns upon collecting data for a future unspecified purpose—see, for example, Article 6 of Directive 95/46 EC (Data Protection Directive). Anonymization may seem to be a solution, but there are legal uncertainties as to how to do it, as well as the fact that the process of anonymization is a form of data processing, which also requires a legal basis [3]. The EU prescribes no uniform criteria relating to anonymization of personal data, and this has resulted in divergent anonymization rules among the member states. For example, the UK seems more liberal regarding the secondary use of health data when anonymized, and the data protection authority has issued a guideline on this [4]. In some other states, there is no clear position, and even anonymized data may be seen as personal data if there is any possibility of linking back to the original data subject [5,6]. This cloud of legal uncertainty is made more complex where a data warehouse is to include data from various EU countries, meaning that a substantial body of regulatory requirements have to be integrated into a single data warehousing infrastructure.

It should be noted that absolute anonymization without any possibility of re-linking the data to a particular person may be impossible in some cases (for example, genetic data) or may prevent useful research (for example, where a new successful treatment is proved, and it may be desirable to contact the patient(s) and/or monitor patient(s) reactions to the treatment). Furthermore, even when personal data are strongly de-identified, advancements in data mining technologies could make re-identification of the patient possible. What may be regarded currently as 'irrevocably anonymized' data may not be seen as such in the future in the light of new technologies [7].

Another legal basis for processing of health data relies upon national exemptions for scientific research, as provided in most EU states. However, a number of countries impose varying additional obligations such as specific security measures and obtaining approval from the data protection authority [8]. Building a unified structure that will comply with and balance these conflicting requirements seems to be a Herculean task, yet it is paramount for a successful clinical data warehousing infrastructure.

1.2. Data security

Apart from the initial legal challenge of processing personal data in compliance with data protection regulations, it is also part of the legal requirements to implement technical and organizational measures to safeguard clinical data when integrated into a unified warehousing system. Security in medical data warehousing is necessary for a number of reasons: to protect the integrity and confidentiality of healthcare data; to ensure availability of resources; and for accountability [9]. Although there is no one-size-fits-all security mechanism under the Data Protection Directive, data controllers and processors are expected to implement a state-of-the-art system that will ensure the confidentiality, availability and integrity of data within their control. This broad provision again gives raise to various interpretations and fragmentation of requirements among Member States. The French data protection authority, for example, has published a guide on security of personal data that recommends avoiding outsourcing services offering cloud computing functions in the absence of any guarantee regarding the effective geographical location of the data [10]. Strict adherence to this recommendation may mean losing the benefits of cloud computing in the health sector. However, the Council of Europe Recommendation no. R (97) 5 on the Protection of Medical Data (13 February, 1997) and the WMA Declaration of Helsinki have recommended some security models for medical databases. These include

— access control;
— management system for the database;
— secure transmission;
— audit or log system;
— anonymization or pseudonymization of data;
— constant review of the security mechanism; and
— conservation of data.

These in effect mean that clinical data warehousing must be robust enough to integrate security mechanisms that not only take care of national regulations, but also international recommendations from medical professional bodies.

1.3. Ethical constraints

The accelerated development in information technology currently enables the collection and evaluation of huge numbers of datasets including those containing sensitive health information. Ethically, the use of such databases demands caution because unauthorized disclosure of sensitive data could have serious negative impacts on the life of the concerned data subject when they become identified. For instance, they can be refused health or life insurance or lose their job as a result of their health status. Even when the data subjects are participating in a clinical trial, it is also important to respect their autonomous decisions as secured by their informed consent. A major challenge, therefore, in integrating medical databases is how to reflect each and every subject's wishes, including their withdrawal of consent and the right to be forgotten [11]. In view of the above, two principles have to be observed:

— the data subjects must be able to make a free decision about providing their data; and

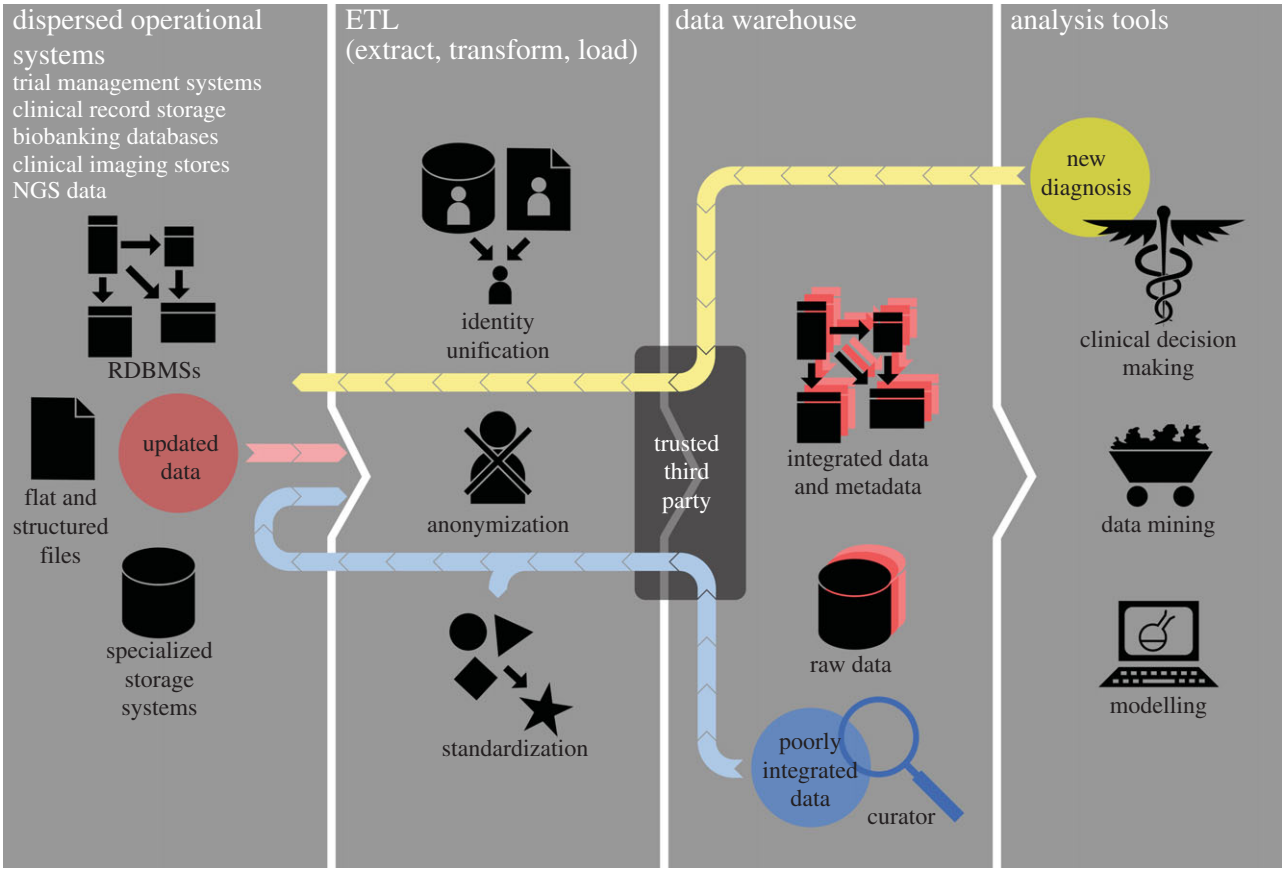

Figure 1. Clinical data warehousing without legal constraints (monochrome) and the alterations needed to allow for legal and ethical constraints (yellow, red and blue). Source data are taken from dispersed operational systems. The extract/transform/load (ETL) stage, these data are extracted from these systems, identities of patients from different databases are unified and anonymized, the data are integrated to a common schema and metadata are extracted. The anonymized raw data, integrated data and metadata are stored in the data warehouse, which can be queried by modelling, data mining and clinical decision-making tools. In yellow, a reversed data flow is shown in the event of a new diagnosis of a disease in a patient, determined during analysis. In red, a resubmission of previously submitted data is shown, resulting in new versions of data already stored in the warehouse. In blue, a data flow is highlighted where erroneous or poor integration of new data with existing data are discovered by a curator. The curator can redo the standardization themselves, or ask the original contributor to address the problem.

— the researchers must implement measures to protect the data from any disclosure to unauthorized persons, as well as have capability to delete any withdrawn data.

1.4. Overview

This paper addresses these constraints and shows how they have been navigated in the course of building the p-medicine data warehouse. This paper is divided into three parts. Section 2 presents a typical clinical data warehouse, with the p-medicine warehouse as a specific instance. Three elements which are affected by legal constraints are described in further detail. In §3, we present the legal constraints that affect these elements. In §4, we describe the technical solutions that we have developed to bend to these constraints. In the conclusions, we additionally describe how legal constraints might be changed to make clinical research simpler, while maintaining the confidence of the general public that their data will not be misused.

2. Clinical data warehousing without legal constraints

Data warehousing in general deals with the gathering of dispersed databases, both live and legacy, into a single resource for analysis and reporting. This is intended to allow more powerful and far-reaching analyses than would be possible upon each database separately.

The monochrome parts of figure 1 show a typical data warehousing workflow for clinical data. Source data are taken from dispersed operational systems such as hospitals, clinics and universities. In the extract/transform/load (ETL) stage, this is extracted, identities of patients from different databases are unified and anonymized, the data are integrated to a common schema and metadata are extracted. The anonymized raw data, integrated data and metadata are stored in the data warehouse, which can be queried by modelling, data mining and clinical decision-making tools.

2.1. Background: the p-medicine data warehouse

A core goal of p-medicine is to collect data on clinical trials of treatment for three different cancers: acute lymphoblastic leukaemia, nephroblastoma and breast cancer. These data are to be contributed from previous trials (retrospective data) and future trials (prospective data) at clinics across Europe. These data are to be collected and standardized such that they can be fed into data mining, biophysical and biomolecular modelling, and clinical decision support systems.

The contributed datasets have schemas that are not currently available. Few standards are adhered to in this realm, and each database has its own, often makeshift, schema. Therefore, the structure of the destination database must be very flexible and be able to accommodate new schemas as

datasets are added, and deal with the missing data fields that often result. Our chosen solution borrows from semantic annotation of web resources, storing data as subject–verb–object triples in a triplestore, with terms taken from an ontology developed within our project, which is built from several standard ontologies. We are using the OWLIM triplestore from Ontotext (www.ontotext.com/owlim).

2.2. Anonymization

Clinical data in the raw form in the originating databases may contain information that can be used to identify a specific individual, either directly (such as name and date of birth) or indirectly (such as a hospital or trial identifier). This information is usually not needed for analysis, and its collection into a central warehouse presents obvious problems of data protection. Therefore, it may be removed or generalized into categories or summaries in order to create anonymous data, which may be distributed more widely than the more sensitive source data.

2.3. Standardization

A common stage in gathering data is to standardize and merge the various data sources. This is to allow data analysts to query data in a standardized manner, without having to be overly aware that they have originated in many places without adherence to a standard data schema. Standardization involves mapping tables and fields to equivalents in a common schema in the destination data storage system, and mapping data values so that data types match, the same terms are used for the same concepts, unit systems are equivalent, relative values are relative to the same base, and so on.

The standardization process is best performed by the people who collect or manage the data. These are the clinicians or data managers of a particular treatment unit or clinical trial. However, the process is far from trivial, and they may make mistakes in this process that leads to incomplete or erroneous integration with the other data in the warehouse. It cannot be reasonably expected that busy clinicians and data managers are trained to completely understand the mapping process such that they rarely or never make mistakes. Therefore, it is useful for the contributors to be able to see how their data have been integrated, and work with curators familiar with the database to correct errors or make integration more complete.

2.4. Updating

Data warehouses are not intended to store live data—that is, data currently being queried and updated to support day-to-day tasks, such as patient care. However, in the clinical context, it may be necessary to add to an existing dataset as a clinical trial or patient treatment proceeds. It may later be found that there are errors in source data that should be corrected, or sensitive data have been released that should not have left the originating clinical domain. A clinical data warehouse should be capable of accepting new versions of data, replacing the erroneous or sensitive data already stored.

3. Legal constraints which affect clinical data warehousing

Three key principles, derived from the legal and ethical perspectives described in the introduction, have constrained the development of the data warehousing infrastructure for p-medicine:

— if a new diagnosis is made during analysis of data in the warehouse, there is a moral duty to communicate this to the patient via their doctor;
— the data contributor cannot know which data they have submitted: owing to merging with other data sources, this may allow unauthorized identification of an individual; and
— clinical decisions must be justifiable and auditable.

In order to allow for these constraints, while still enabling useful research and analysis, consent must be obtained to allow the addition of prospective data to a data warehouse, and to permit further processing of that data. To protect the data in p-medicine, it is pushed into the warehouse in anonymized form, applying both technical and contractual measures to keep the data anonymous (at least, unreasonable effort of time, cost and labour would be required to attach the data to a certain individual). In addition, a specific data protection authority was set up to take care of the data protection framework: the Centre for Data Protection (CDP) (cdp.custodix.com).

3.1. Moral duty to communicate new diagnoses

The debate as to whether trial participants should positively benefit from the research by communicating new diagnoses to them have tilted towards a feedback approach at least as a moral duty in many cases [12]. With more international endorsement of this approach, it is forward-looking to incorporate this mechanism into the warehousing infrastructure. The simplest way to allow communication of new diagnoses to patients would be to keep a link to them. Complete anonymization will prevent this, which means that other legal grounds for storing and processing patient data have to be found.

In constructing a data warehouse for clinical information, such as that being developed for p-medicine, there are two key sources of data. The first is retrospective data: that is, data generated in trials or care before the data warehouse existed or before the researchers were aware of the desire to collect data into the warehouse. The second is prospective trial data: data generated in trials after the warehouse existed and with reference to the desire to collect data into the warehouse. Prospective data may be collected with the consent of the patients, explicitly permitting their transfer to the warehouse.

Retrospective data give a little more trouble. There may not be the specific consent that covers their transfer to the data warehouse, which means that a re-consent is needed from the data subjects, except other national exemptions under Art. 8 (4) of the Data Protection Directive apply. Obtaining re-consent may be impossible because patients have died or clinicians have lost contact with them. Even where it is possible, it can be costly and time-consuming. As indicated earlier, reliance on national exceptions seems preferable, but this introduces its own difficulties.

3.2. Data contributor must not know which data they submitted

Within the EU, different rules apply for anonymized data as opposed to non-anonymous data. It is not clear what the

legal effect of a link back to the identifiable data for the purpose of communicating new diagnoses will be in most cases. There is no uniform opinion laid down in European law as to what 'anonymous' means in a legal sense, and how anonymization is to be achieved. Some Member States take a very strict view and require the irrevocable breach of the link to the person. Others regard key coded data as anonymized, as long the researcher or sponsor has no access to the key. Germany, for instance, has adopted the 'disproportionate effort' approach, where data are regarded as anonymized, if a re-link to the data subject is impossible, or if an unreasonable effort of time, cost and labour would be required to attach the data to a certain individual (§3(6) German Data Protection Act). Ireland demands an irrevocable deletion of the link for data to be regarded as anonymous.

In the strictest sense, if data are to be regarded as anonymous, even the data exporters should not know the exact data they submitted (in spite of the fact that the exporter is the data controller of the original data, and has all the personal identifiers). This is required because part of the ETL phase of data warehousing involves data from multiple sources on a single patient being linked, such that it refers to the same identifier or pseudonym. If a particular contributor knew which data they contributed, this linking process would mean they may gain new information about a patient they can personally identify— information that was originally given to a different clinic. This is a breach of the data protection laws.

3.3. Clinical decisions must be justified and auditable

Doctors and hospitals are bound by legal agreements to minimize potential harm to patients during treatment. By extension, other aspects of a healthcare system are subject to legal scrutiny, including drugs and equipment. A data warehousing infrastructure may initially be used only for research purposes, but once it is used as part of a clinical decision support system, it becomes part of the healthcare infrastructure. This transition from research to healthcare is mirrored in human genetics work, where there have been calls for clinical standards to be applied to such research [13]. An initial step in this transition would be to ensure all advice offered to a clinician can be backed up by an audit trail, from the decision support system, through the modelling and mining methods, to the source data and the data contributors. Such an audit trail also conveniently supports the reproduction of results in published research.

4. Clinical data warehousing with legal constraints

The effect of these constraints upon our clinical data warehousing system are highlighted in red, yellow and blue (figure 1).

4.1. Reversible pseudonymization

Anonymization removes all personally identifiable data. An alternative is pseudonymization, which replaces personally identifiable data with a warehouse patient identifier that allows an analyst to see what data relate to a particular patient, but not know exactly which patient outside the context of the warehouse.

As previously stated, it may be desirable to identify individuals referred to in the warehouse. Our solution is to have a trusted third party (TTP) that maintains secret tables relating pseudonyms to entities outside the realm of the warehouse, such as patients. If, during the course of analysis, a possible undiagnosed condition is identified, an alert can be sent to the data contributors via the TTP and CDP. This data flow is highlighted in yellow in figure 1. This is a reverse of the normal flow of data in data warehousing. It is an exceptional circumstance unique to clinical data warehousing, owing to the moral and ethical duty to communicate a new diagnosis.

Researchers will not have access to the original non-anonymous data, and are under some contractual obligations not to identify individuals described in the data. Where there is a need to recontact the data subject, a procedure is followed whereby a request is made to the CDP, detailing the reason for such recontact. A request for the key is then made by the CDP to the TTP for de-identification.

Once authorized, messages about a specific patient can be sent back to the sources through a computational interface made available by the TTP. A message is sent to the treating clinician using the warehouse patient identifier (the pseudonym issued by the TTP for the patient). After approval of the re-identification by a privacy manager, the TTP will forward the message to each source and replace the pseudonym by the corresponding patient identifier (e.g patient hospital number) for the given source. The treating physician is contacted, who has access to the hospital pseudonymization key and who can then contact the patient, assuming they have indicated that this is their wish.

4.2. Curated standardization

The legal constraints mean that data contributors cannot know which data in the warehouse were contributed by them. This means it is impossible for them to know if the data has been properly integrated, and also impossible for them to correct any mistakes.

Our proposed solution to this problem is twofold. Warehouse curators, who are familiar both with the schema of the warehouse and the apparent schema of the contributed data, can check for and perform proper standardization. While this is a reasonable approach, it may fail where the curator is not sufficiently familiar with the contributed data schema to correct problems, and may in fact introduce new ones.

The second solution is to allow data contributors to revise the mapping from the source data schema to the data schema used by the data warehouse, in response to advice from curators. Clearly, this will involve a curator being able to indicate to the contributor which dataset has a problematic mapping from the source schema to the warehouse schema; and contributors must be able to identify which dataset's mapping is being altered.

The process allowing this involved the TTP providing the data contributor with a contributor dataset identifier (CDI) upon submission of data. When data is added to the warehouse, it is associated with a warehouse dataset identifier (WDI). The TTP maintains a secret table mapping from CDIs to WDIs. If a curator sees a problem with how a dataset has been standardized, they can either attempt to correct it themselves, or follow this procedure:

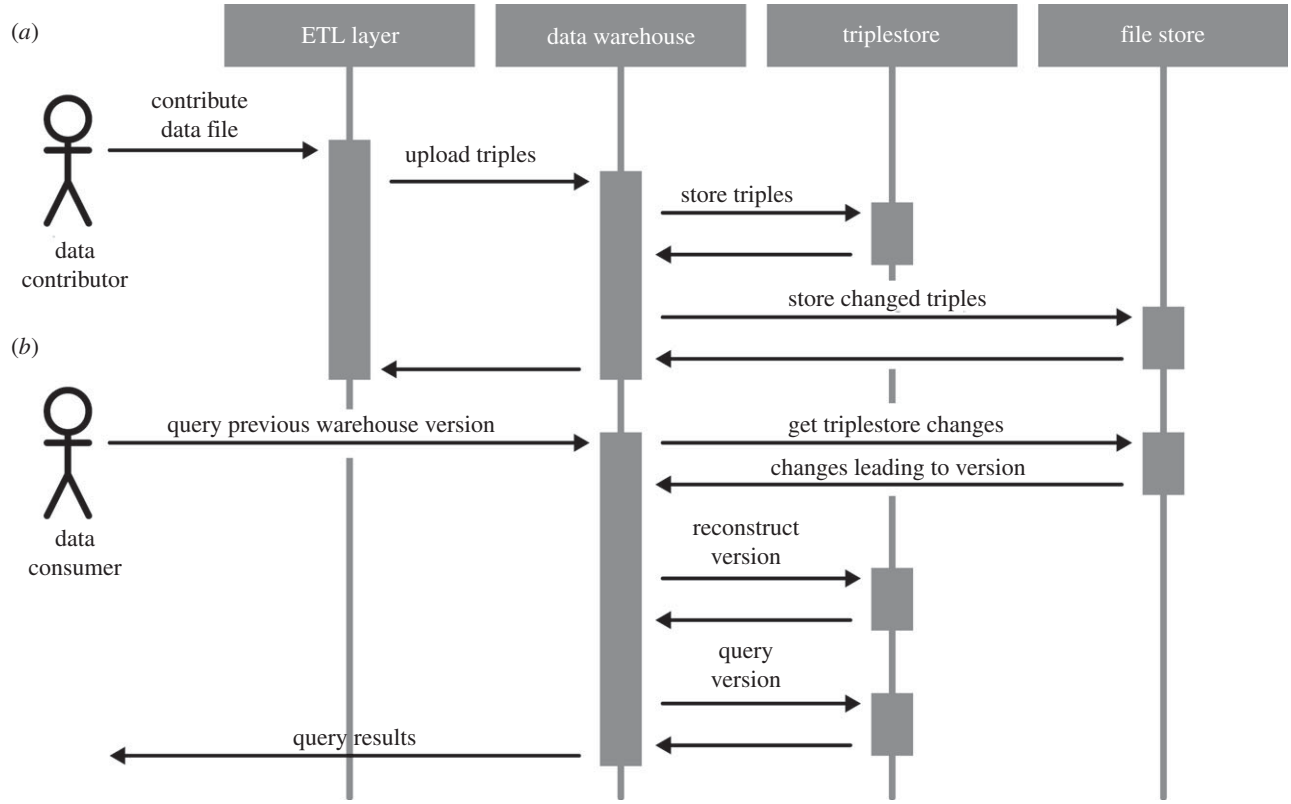

Figure 2. Interaction diagram illustrating the process of submitting new data to the warehouse and (*a*) recording the associated changes, and (*b*) later performing a query upon a specific version of the data warehouse.

— contact the TTP with a WDI indicating which dataset has not been properly integrated and an indication of the problem;
— the TTP knows who submitted what data, and contacts them with a CDI derived from the WDI using its mapping table; and
— the contributor can then alter the mapping and submit it, via the TTP, to the warehouse, using the CDI as an identifier.

This process is highlighted in blue in figure 1. A key disadvantage of this process is that, because the data contributor is not allowed to see which data they have contributed in the context of the data warehouse, they cannot see if there is a problem with the way the dataset has been standardized and integrated with the existing data, and they cannot see the precise nature of the problem—although the curator may attempt to explain it.

4.3. Versioned updates

Making updates to a warehouse changes the content, which will clearly change the results of queries upon the warehouse. This may make it impossible to justify a clinical decision made on the basis of supporting evidence from the warehouse. Likewise, journals require that published results must be repeatable. If these results rely upon queries to the data warehouse, and the data have subsequently changed or been added to, the queries will not be repeatable.

Therefore, apart from cases where very sensitive data have been accidentally released, updates to the warehouse should be subject to a versioning system. When a query is performed, the result is given with a data warehouse version identifier, which can be later used to rerun the query upon the same version of the warehouse, giving the same results.

While there are many methods for implementing a versioned warehouse system, ours was guided by the following assumptions:

— most queries will be performed upon the latest version of the warehouse, and this process must be as fast as possible because operational systems may rely upon the results;
— execution of queries upon previous versions of the warehouse is a rare occurrence and part of a specific, isolated task that does not require constant maintenance of previous versions;
— updates typically involve large quantities of data, and the process of ensuring previous versions are kept should not significantly slow down an update; and
— storage is cheap.

Recall that our warehousing solution stores structured data in a triplestore, which is, loosely speaking, a single table with three columns: subject, verb and object. Each row is known as a triple. Upon every update, we store a list of subject–verb–object triples that are added or removed in that transaction. A transaction results in an increment in the warehouse version number. The first attempt to query any version of the warehouse other than the current one results in the reconstruction of that version in a separate triplestore, from the stored transaction records. This results in a potentially long start-up time, but once this task is complete, subsequent queries are fast. Reconstructed historic triplestore versions are not kept forever: they are deleted according to last access and storage space rules, configured according to local requirements. This process is outlined in red in figure 1, and explained in greater detail in a sequence diagram in figure 2.

5. Conclusions

Clinical data warehousing is a growing area with several solutions available, most notably commercial. Commercial solutions developed by IDBS (www.idbs.com) and Aridhia (www.aridhia.com) have so far focused upon warehousing within the context of the organization providing the data, and only providing subsets of data to people with a specific research goal upon application. p-Medicine differs in taking and merging data from many institutions in many countries, thereby taking data outside their realm, both institute and country, with the associated legal consequences. These data are provided as a complete resource to analysts who have signed the appropriate agreements, without them needing to provide a specific research agenda—meaning that the consent forms need to be very general, presenting another legal challenge. The diverse sources of data mean that standardization is difficult, and cannot be characterized as a simple 'one-way' process: it is collaborative and must be subject to curation, with the constraint that data contributors cannot know which data they have contributed. These factors present legal problems which other efforts do not face.

5.1. Changes to data protection law to aid clinical trial analysis

In this paper we have presented technical solutions that we have adapted to the legal constraints presented by EU data protection law. However, problems remain. In particular, curation of the standardization of data to a common schema in the warehouse is very difficult. We are likely to encounter further problems in the future. Technical solutions are not the only course of action, and in parallel we might usefully suggest changes to these constraints to simplify attempts to analyse clinical data, particularly meta-analysis of multiple datasets from many sources.

A UK study [14] has indicated that the general public are not as concerned with the use of personally identifiable data for clinical research as has been previously assumed, although it has yet to be seen what the general reaction would be to a loosening data protection legislation for this purpose. A general willingness to share data is particularly emphasized by private sector projects that collect information freely offered by people in return for some service to them:

— 23andMe (www.23andme.com) provides genetic testing;
— ancestry.com (www.ancestry.com) constructs family trees and links people with their ancestors;
— patientsLikeMe (www.patientslikeme.com), HealthUnlocked (www.healthunlocked.com), I Had Cancer (www.ihadcancer.com) and others provide disease-specific social networks for people undergoing treatment; and
— Nike+ (http://nikeplus.nike.com/plus/) collects data on physical activity for personal training.

All of these collect personally identifiable data that are freely given up by data subjects, and the legal agreements associated with using these websites allow very liberal use of the data. These examples all provide a service as an incentive for providing data: by contrast, the Personal Genome Project only offers the chance to help biological research as reward. If public sector efforts, such as p-medicine, are hampered by cumbersome legislation, then large-scale collection of data by the private sector will dominate—losing possible public health benefits of free access to that data for research.

In addition to the harmonization of data protection law within the EU that is underway with the proposed Data Protection Regulation [15], which will, to a large extent, remove the fragmentation of data protection rules as seen in the transposed legislations within the Member States, we propose two possible changes which will enable clinical research.

It would be useful to have a clear guide on the secondary use of medical data for research. So far, p-medicine has had to negotiate its own terms for such usage, and this has been extremely time-consuming. Currently legislation is focused upon use of data for traditional hypothesis-driven research. The use of clinical data for general data mining and analysis to feed into live clinical decision support systems is made difficult by this emphasis. Legislation must be changed to allow very general use of data for research which has not been specified in advance. Clearly there are concerns, in granting access to large quanitites of data to researchers with no clear agenda, that data may be misused. Legal agreements must be formed between the data consumers and the data controllers that associate penalties with such misuse, and data controllers must track who is accessing what data for auditing and to detect possible abuse of data. This paper has described how we have done this in p-medicine, but clear advice on standards for legal agreements and auditing procedures would be useful.

It would also be useful to have a better and more consistent definition of the terms anonymous and pseudonymous in the proposed data protection regulation, stating their legal effects for medical research. This is currently defined in detail in the USA by the HIPAA Privacy Rule [16], which recognizes the potential utility of health information even when it is not individually identifiable, (§164.502(d) of the Privacy Rule) and thus permits a covered entity or its business associate to create information that is not individually identifiable by following the de-identification standard and implementation specifications in §164.514(a)–(b). The Privacy Rule provides two de-identification methods:

— a formal determination by a qualified expert; or
— the removal of specified individual identifiers as well as absence of actual knowledge by the covered entity that the remaining information could be used alone or in combination with other information to identify the individual.

With either method, the entity is allowed to use and disclose information without the HIPAA restrictions. Although the risk of identification cannot be completely eliminated, the rule recognizes that health information is anonymized if there is shown to be a very small risk of it (perhaps in combination with other accessible information) being used to identify the subject of the data by a person well-versed in the statistical or scientific principles required to do so, or if a detailed list of identifiers of the subject or people related to the subject are removed. It would be useful for the EU to proceed in this direction, by having harmonized and clear rules that state specifically how to anonymize, and the legal implications of the use of anonymized, pseudonymized and encrypted data—especially where re-identification is not possible either through restrictive contractual obligations, or through technical means (for example, keys for reidentifying data are not accessible). The

EU Data Protection Directive 95/46/EC defines anonymous data as a form in which data cannot be used to identify the data subject, and gives no further detail on the matter, leaving the decision as to what constitutes personal data to individual Member States. This is not sufficient guidance for infrastructure projects that source and share data across the whole EU. In this regard, the UK anonymization code of practice should be emulated.

The European Integrated Project p-medicine is supported through Coordination Theme 3 (Information and Communication Technologies) of the European Community's 7th Framework Programme, grant agreement no. 270089.

References

1. Editorial. 2012 What happened to personalized medicine? *Nat. Biotechnol.* **30**, 1. (doi:10.1038/nbt.2096)
2. Rossi S, Christ-Neumann ML, Rüping S, Buffa FM, Wegener D, McVie G, Coveney PV, Graf N, Delorenzi M. 2011 p-Medicine: from data sharing and integration via VPH models to personalized medicine. *E-Cancer* **5**, 218. (doi:10.3332/ecancer.2011.218)
3. The Federation of European Academies of Medicine (FEAM). 2012 Data protection regulation: a FEAM statement. See http://www.feam-site.eu/cms/docs/publications/FEAMDataProtectionStatementJune2012.pdf.
4. Information Commisioner's Office. 2012 Anonymisation: managing data protection risk code of practice. See http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx.
5. Quathem KV. 2007 *Controling personal data in clinical trials*, pp. 1–11. Washington, DC: Covington & Burling.
6. Quathem KV. 2005 *Controling personal data: the case of clinical trials*. Washington, DC: Covington & Burling.
7. Ohm P. 2010 Broken promise of privacy: responding to the surprising failure of anonymisation. *UCLA Law Rev.* **57** 1701.
8. Retzer K *et al*. 2011 Navigating the sea of data protection law in European clinical research. See www.scripclinicalresearch.com.
9. Hamilton D. 1992 Identification and evaluation of the security requirements in medical applications. In *IEEE Symp. on Computer-Based Medical Systems, Durham, NC, 14–17 June 1992*. New York, NY: IEEE.
10. Commission Nationale de l'Informatique et des Libertes. 2010 The CNIL's guides: security of personal data. See http://www.cnil.fr/fileadmin/documents/en/Guide_Security_of_Personal_Data-2010.pdf.
11. Karp D *et al*. 2008 Ethical and practical issues associated with aggregating databases. *PLoS Med.* **5**, 1333–1337. (doi:10.1371/journal.pmed.0050190)
12. Shalowitz DI, Miller FG. 2008 Communicating the results of clinical research to participants: attitudes, practices, and future directions. *PLoS Clin. Trials* **5**, e91. (doi:10.1371/journal.pmed.0050091)
13. Lyon GJ. 2012 Personalized medicine: bring clinical standards to human-genetics research. *Nature* **482**, 300–301. (doi:10.1038/482300a)
14. Barrett G, Cassell JA, Peacock JL, Coleman MP. 2006 National survey of British public's views on use of identifiable medical data by the National Cancer Registry. *BMJ* **332**, 1068. (doi:10.1136/bmj.38805.473738.7C)
15. European Commission. 2012 Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). See http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.
16. U.S. Department of Health & Human Services. 2012 Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. See http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.