Deliverable No. 5.2

Report on Legal and Ethical Issues Regarding Data Warehouse, Data Mining and Intellectual Property Issues

Grant Agreement No.:	270089
Deliverable No.:	D5.2
Deliverable Name:	Legal and ethical issues regarding data warehouse, data mining and intellectual property
Contractual Submission Date:	31/07/2012
Actual Submission Date:	31/07/2012

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission	

	Services)	

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	***p-medicine***
Project Full Name:	From data sharing and integration via VPH models to personalized medicine
Deliverable No.:	D5.2
Document name:	Legal and ethical issues regarding data warehouse, data mining and intellectual property
Nature (R, P, D, O)[1]	R
Dissemination Level (PU, PP, RE, CO)[2]	PU
Version:	2
Actual Submission Date:	31/07/2012
Editor: Institution: E-Mail:	Prof. Dr. Nikolaus Forgó Leibniz Universität Hannover, Institut für Rechtsinformatik forgo@iri.uni-hannover.de

ABSTRACT:

This deliverable contains an analysis of the relevant legal and ethical requirements for the establishment and mining of the p-medicine research data warehouse. It also looks at the intellectual property rights accruable from the use of the database.

At a high level, it provides a brief overview of the legal and ethical preconditions for the establishment of a database containing sensitive data, in particular health-related data, for medical research on a European level. These conditions also include those from international GCP guidelines such as the Helsinki Declaration.

An evaluation of the architecture of the p-medicine data warehouse is also made, including the incorporation of privacy enhancing data mining techniques and the integration of security framework in the structure.

Rules guiding the background and foreground material used in EU P7F were also analyzed in conjunction with the Consortium Agreement that specifies the intellectual property rights of participating partners. Issues relating to the legal protection of the p-medicine database under the Database Directive were equally x-rayed.

KEYWORD LIST: data protection, data warehouse, data mining, intellectual property, data security, anonymisation, pseudonymisation, medical research, good clinical

[1] **R**=Report, **P**=Prototype, **D**=Demonstrator, **O**=Other

[2] **PU**=Public, **PP**=Restricted to other programme participants (including the Commission Services), **RE**=Restricted to a group specified by the consortium (including the Commission Services), **CO**=Confidential, only for members of the consortium (including the Commission Services)

practice.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270089.

The author is solely responsible for its content, it does not represent the opinion of the European Community and the Community is not responsible for any use that might be made of data appearing therein.

MODIFICATION CONTROL			
Version	Date	Status	Author
1.0	15/04/2012	Draft	Nikolaus Forgó, Iheanyi Nwankwo, Tina Krügel, Stefanie Hänold, Stefan Rueping, Henrik Großkreutz, Elias Neri
2.0	29/07/2012	Final	Nikolaus Forgó, Iheanyi Nwankwo, Stefanie Hänold, Stefan Rueping, Henrik Großkreutz, Elias Neri

List of contributors

- Nikolaus Forgó, LUH
- Iheanyi Nwankwo, LUH
- Stefanie Hänold, LUH
- Tina Krügel, LUH
- Elias Neri, Custodix
- Henrik Großkreutz, FhG-IAIS
- Stefan Rueping, FhG-IAIS

Abbreviations and acronyms

Acronym	Definition
CAT	Custodix Anonymisation Tool
CATS	Custodix Anonymisation Services
CA	Consortium Agreement
CDP	Center for Data Protection
D	Deliverable
DoW	Description of Work
DPA	Data Protection Authority
EC	European Commission
ECJ	European Court of Justice
EEA	European Economic Area
ENISA	European Network and Information Security Agency
ECHR	European Court of Human Rights
EU	European Union
EMEA	European Medicines Agency
GCP	Good Clinical Practice
ICO	The UK Information Commission Officer
ICT	Information and Communication Technology
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
IP	Infrastructure Provider
ISO	International Organisation for Standardisation
IdP	Identity Provider
IT	Information Technology
PET	Privacy Enhancing Technologies
P-MEDICINE	From Data Sharing and Integration via VPH Models to Personalized Medicine
REST	REpresentational State Transfer
SP	Service Provider
STS	Secure Token Service
SLA	Service Level Agreement
TRIPS	Agreement on Trade-Related Aspects of Intellectual Property Rights

WMA	World Medical Association
WIPO	World Intellectual Property Organization

Contents

1 Executive Summary

The major aim of this report is to assess the nature and scope of issues that may arise in the establishment and mining of the p-medicine data warehouse, and to make recommendations and provide guidance on good practices. Mainly, this Deliverable addresses the contextual and strategic legal and ethical issues raised by establishing and mining of the p-medicine research data warehouse, as well as intellectual property rights that may exist therein. In order to address these issues, a review of existing laws and guidelines affecting a medical research database was carried out.

While there are a number of legal and ethical documents within the EU relating to clinical trials, there is none that is published specifically for the integration of medical research data warehouses that harmonises the different positions within the Member States on the ethical, legal and operational issues raised by such database. The guidance that this Deliverable gives would make a positive impact in facilitating translational medical research through ensuring that data warehousing and mining comply with legal and ethical requirements within the EU and respect patients' privacy. In this respect, a privacy enhancing data mining technique has been introduced in the database architecture.

Because a medical research database will ordinarily contain personal data, one of the legal instruments considered in this report is the Data Protection Directive which stipulates the basic conditions for the processing of personal data. Other directives relevant in this regard include the Clinical Trial Directive, the Good Clinical Practice Directive and the Database Directive concerning the intellectual property residing in the database. Furthermore, provisions of relevant international ethical documents such as the Declaration of Helsinki and the WMA Declaration on Ethical Considerations regarding Health Databases were also looked at.

In all, the following issues were considered in this report in regard to setting up a medical research data warehouse under a pan-European framework:

i. Informed consent of the data subjects;
ii. Protection of the sensitive data used in the research;
iii. Privacy enhancing technique for data mining;
iv. Security of the data warehouse;
v. Access policy for the use of data warehouse.

At the end of this report, a policy guideline is annexed to ensure the continued protection and security of the data. The guideline is complementary to the obligations already in existence by virtue of the contracts within the p-medicine data protection framework and gives a fine-grained implementation of the data protection framework.

Finally, intellectual property rights inherent in the data warehouse were analysed in line with the consortium agreement which distinguishes between background information/knowledge and foreground results. In line with this, background information/knowledge is retained by the beneficiary who holds such knowledge prior to joining the consortium, while foreground information, materials and knowledge shall be the property of the beneficiary carrying out the work generating that foreground. However, where several beneficiaries have jointly carried

out the work generating the foreground, and where their respective share of the work cannot be ascertained, the foreground shall be jointly owned by such beneficiaries. The beneficiaries in this case can also have a joint agreement on how to apportion the foreground which in p-medicine is reflected the Consortium Agreement.

2 Introduction

A data warehouse is a repository for securely storing and maintaining data from diverse sources, integrated semantically to enable reporting and analysis.[3] Establishing such a data warehouse for integrating large amount of data needed for a translational medical research such as the p-medicine project seems essential and could serve as a foundation for a knowledge discovery system.[4] Such a database[5] is more beneficial when heterogeneous data from various sources are seamlessly integrated so as to produce a predictive result when mined appropriately. While the use of information technology in setting up data warehouse and mining capabilities in health-related research can be very advantageous, it is nevertheless without attendant legal and ethical concerns. Especially, regarding the privacy of the patients involved in the research, as well as their close relatives. Not only is it a legal prerequisite that such a database should contain accurate data, the security of the pooled data resources also has to be guaranteed in order to maintain the confidentiality, integrity and availability of the data.

While it is of great public interest that medical researches are carried out, in order to expand knowledge in health care delivery, various EU legislations as well as professional guidelines have set conditions for the collection and establishment of sensitive health-related databases. The prominent rationale for these conditions is the need to protect the privacy of the patients, and thus, as far as possible, personal identifying information of patients participating in medical research should be coded, anonymised or removed.[6] However, complete anonymisation without any possibility of re-linking the data to anyone may not serve the purposes of the research in most cases. Even when personal data are strongly de-identified, the advancements in data mining technologies could make re-identification of the persons (sometimes from other publicly available databases) possible, if appropriate mechanisms, including privacy enhancing mining techniques are not put in place. In this regard, the aim of this report is to outline the legal and ethical issues to be considered prior to setting up the p-medicine data warehouse that will contain sensitive health-related data, as well as those issues that may arise in its mining. In the end, data warehousing and mining guidelines for the p-medicine project will be developed to tackle the identified issues. These guidelines, in addition to the end user agreement that is already in existence, will regulate the use of the data warehouse and will be published on the web portal for all end users to read and adhere to.

On the flip side of these issues is the ownership of the intellectual property rights that may exist in the p-medicine data warehouse. Who really owns the data in the database and what legal protection exists for it? Current discussions in the field of ownership of data in medical research databases containing patients' information are still unsettled. Analyses have shown however, that possible interested parties in such ownership and intellectual property rights may include: the patients, the hospitals, the public, the sponsors and the

[3] Deliverable D7.1: Report on overall system design including VPH-Share D2.2 and indicating its impact, p. 6.

[4] Nevena Stolba, Marko Banek and A Min Tjoa 2006, 'The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine', available at: http://wit.at/people/stolba/documents/paper-Stolba-Banek-Tjoa_000.pdf (accessed 25 April 2012).

[5] In this report, data warehouse is used synonymous with database.

[6] Council of Europe, Committee of Ministers, Recommendation No. R (97) 5 on the Protection of Medical Data (Feb. 13, 1997).

researchers/participants in the research.[7] In the case of p-medicine, one other party can be identified – the Center for Data Protection (CDP), which assumes the duty of the central data controller of the database. Although there may be interesting arguments favouring ownership in each of these parties mentioned above, the p-medicine Consortium Agreement as well as the EC Grant Agreement shed some lights in resolving these intellectual property issues. The EC Grant Agreement provides that background material used in EU project belongs to the beneficiary who has it prior to accession to the project. However the foreground property, which is the result obtained from the processing of the background information jointly belongs to the beneficiaries that produced them if individual contributions cannot be ascertained.[8] The EC Grant Agreement gives room for the beneficiaries to establish a joint ownership agreement regarding the allocation and terms of exercising such joint foreground. In p-medicine however, it is provided in the Consortium Agreement that "each of the joint owners shall be entitled to use their jointly owned foreground as it sees fit, and to grant non-exclusive licences, without obtaining any consent from, paying compensation to, or otherwise accounting to any other joint owner, unless otherwise agreed between the joint owners".[9]

From a legal perspective, three legal regimes may protect a medical research database. First, where a medical database contains copyright protected material such as medical images, copyright protection is available for such intellectual creation. Database rights may also govern aggregated sources of medical information, and patent law may protect derivative technologies.[10] In a final analysis, the European Database Directive provides a two-fold method of protection of databases. The first scheme of protection is as an intellectual creation through copyright, while the second scheme is the *sui generis* right. In accordance with the former, databases which by reason of the selection or arrangement of their contents constitute the author's own intellectual creation shall be protected as such by copyright. The latter focuses on the protection of the investment spent on the creation of databases as a compilation of data.[11] To qualify for the *sui generis* database protection, it is required that the elements in the database are individually accessible by electronic or other means. Secondly, the creator of the database must show that there has been qualitatively and/or quantitatively, a substantial investment in either the obtaining, verification or presentation of the contents.[12]. Much controversy have been generated in the interpretation of what amounts to a substantial investment, and whether an investment could be seen as substantial if the investment was not directly directed towards the creation of the database. However, courts in several countries have held that even if the database is a "spin-off" of some other activity, the creator may hold a database right if that other activity required a substantial investment.[13] Although there are divergent opinions on these issues from national

[7] Chris Hinds et. Al, 'Ownership of Intellectual Property Rights in Medical Data in Collaborative Computing Environment', available at: http://www.ncess.ac.uk/events/conference/2005/papers/papers/ncess2005_paper_Hinds.pdf (accessed 10 June 2012).

[8] FP7 Grant Agreement - Annex II General Conditions, available at: ftp://ftp.cordis.europa.eu/pub/fp7/docs/fp7-ga-annex2-v6_en.pdf. (accessed 10 June 2012). See section 6 below for detailed discussion.

[9] See section 8.1 of the Consortium Agreement.

[10] Chris Hinds, op. cit.

[11] Michele Oliva and Marcelo Corrales, Law Meets Biology: Are Our Databases Eligible for Legal Protection? 2011, *Scripted*, Vol. 8, Issues 3, p. 227.

[12] See Article 7 of the Directive 96/9/EC (Database Directive).

[13] See for example in France France Telecom vs. MA Editions (Tribunal de commerce de Paris, 18 June 1999), in Germany Tele-Info-CD (Bundesgerichtshof, 6 May 1999), or in the Netherlands KPN v. Denda (Gerechtshof Arnhem, 15 April 1997, follow-up in Rechtbank Almelo, 6 December 2000); See generally: http://www.iusmentis.com/databases/crashcourse/requirements/ (accessed 12 June 2012).

courts,[14] various decisions from the ECJ have also throne some uncertainty on legal protection of databases under the Directive.[15] The p-medicine database in our assessment is protected by *sui generis* right, independent of any other copyright that may accrue to the content of the database. This basically is on the premise that a substantial investment has been made in obtaining and establishing the database, as well as its accessibility via electronic means as envisaged by the Database Directive. Detailed discussions on these issues will be made in the subsequent sections of this report.

[14] For example *NVM vs. De Telegraaf* (21 December 2000) from the Dutch Court of Appeal in The Hague.

[15] See for instance, The British Horseracing Board Ltd and Others v William Hill Organization Ltd, (2004) EUECJ C-203/02; *Fixtures Marketing Ltd v OY Veikkaus Ab* (2004) EUECJ C C-46/02; Fixtures Marketing Ltd v Svenska Spel AB, (2004) EUECJ C-338/02.

3 Structure

The document is divided into four main parts.

The first part (chapter 4) gives an overview of the legal and ethical requirements for establishing a data warehouse for a translational medical research. Various legal instruments and ethical codes were considered, especially those that are relevant for the protection and security of the sensitive data of the trial participants, as well as for good clinical practices.

In the second part (chapter 5) the architecture of the p-medicine data warehouse and mining is x-rayed.

The third part (chapter 6) looks at the issues of the intellectual property rights that exit in the p-medicine database, while the final part (chapter 7) contains the guidelines for the use of the p-medicine data warehouse. Concise guidelines are annexed at the end of the chapters.

4 Legal and ethical issues relating to establishment of a data warehouse in translational medical research

4.1 Introduction

A Data warehouse that facilitates translational medical research such as p-medicine involves a huge amount of sensitive data. The legal pre-conditions for establishing such a database within the EU stem from diverse sources, ranging from directives and national laws to medical practice regulations and guidelines. Health-related personal data are in the first instance part of the general personal data protected under the Data Protection Directive[16] which is the basic document regulating personal data processing at EU level. Hence, any data controller processing personal identifiable health information must comply with the rules and principles established in the Directive as transposed into national law. These principles have been enunciated in Deliverable D5.1 which in a nutshell are that data controllers shall: limit data use to the purpose for which it was collected (purpose principle), ensure data quality (relevancy and accuracy principle), limit data retention (and not further process the data for incompatible purposes), provide individuals with data collection information and access to the information collected (with rights of correction), and provide appropriate data security measures.[17] Where sensitive data are to be processed, additional requirements such as obtaining explicit and informed consent of the data subject are also imposed.

Other Directives such as the Directive 2001/20/EC,[18] the Directive 2005/28/EC,[19] as well as internationally recognised rules such as the Declaration of Helsinki, the ICH Guidelines for Good Clinical Research and the Council of Europe Recommendation No. R (97) 5 also have some impact on the issue of establishing a medical research database as a result of their specific application to the medical field. In a nutshell, it could be deduced from all these instruments that processing of data relating to a person's health is particularly sensitive and therefore requires special protection.[20] Thus, it is very important to comply with all the prerequisites of the laws and regulations mentioned above, as such will go a long way in reducing the risk associated with establishing a data warehouse containing sensitive data as seen in p-medicine.

A federated data warehouse system in which data are stored in physically separated datasets and integrated through the use of information technology has been adopted in p-

[16] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive).

[17] See Deliverable No. D5.1 - Setting up of the data protection and data security framework for p-medicine.

[18] Directive 2001/20/EC of the European Parliament and of the Council on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use (Clinical Trials Directive).

[19] Directive 2005/28/EC laying down principles and detailed guidelines for good clinical practice as regards investigational medicinal products for human use, as well as the requirements for authorisation of the manufacturing or importation of such products" (Good Clinical Practice Directive).

[20] Article 29 Working Party, Working Document on the processing of personal data relating to health in electronic health records (EHR), WP 131, p. 7.

medicine.[21] This provides direct access to a compilation of medical data from different sources. From the technical perspective, multiple access points over an open network like the Internet will be used to access these data. The use of an open network however may increase the possibility of patient data interception as indicated by the Article 29 Working Party.[22] This poses a significant challenge in ensuring that not only authorised persons gain access to the data warehouse, but also that the mining results will respect the privacy of the patients. As such, it will require implementing special safeguards for the security of the data warehouse, as well as incorporating privacy enhancing techniques into the architecture of the warehouse. The legal sources establishing these preconditions as well as the safeguards necessary for protecting a medical research database will be discussed in the following sections.

4.2 Legal requirements for the establishment and use of a data warehouse in medical research

As mentioned earlier, legal preconditions for setting up a medical research database within the EU do not only stem from directives and national laws, but also from international medical practice regulations and guidelines. While there may be plethora of these rules, key reference points for our purposes in p-medicine will be the relevant ones that will impact the database such as the Directive 95/46/EC, Directive 2001/20/EC, Directive 2005/28/EC and other relevant documents that provide context driven ethical and good practice guidelines. They are discussed below.

4.2.1 The Data Protection Directive (Directive 95/46/EC)

Although the Data Protection Directive did (of course) not specifically focus its attention on the establishment of a translational medical data warehouse, its general principles apply to such databases by virtue of the sensitive personal data involved. To that extent, the conditions in the Directive for the processing of personal data will apply, if any category of data involved in any medical research reveals an identifiable person. Deliverable D5.1 which extensively dealt with the data protection and security framework for p-medicine in general, also relates by extension to the p-medicine database. Nevertheless, it has been established that the p-medicine data warehouse will contain only *defacto* anonymous data, which removes data processing in the data warehouse from the regulation of the Directive. In view of this, we will streamline our analysis here on the principles and guidelines that will safeguard sensitive data *albeit* not personal, in the warehouse, as well as generate the trust of the patient participants. It is worthy of note to point out here that in addition to observing the basic principles outlined in section 4.1 above, Article 8 of the Directive on informed consent was equally observed during the initial collection of data to be used for this research. In this respect, a comprehensive informed consent form was developed for the project, taking into account the relevant information that the trial participants should know before consenting to the trial.[23]

A safety net has been established for the project, consisting of a tripartite legal structure for data processing that took into account the preconditions for obtaining informed consent of the trial participants as stipulated in the Directive and other relevant instruments.[24] This

[21] Deliverable D7.1: Report on overall system design including VPH-Share D2.2 and indicating its impact.
[22] Ibid.
[23] See Annex B of Deliverable D5.1, op. cit.
[24] Not only is informed consent a legal requirement but also an ethical requirement. See Deliverable 5.5 for the report on legal and ethical issues for p-medicine tools used for international GCP trials.

tripartite structure include: the obtaining of the informed consent of patients participating in the project; *defacto* anonymisation of the patients' data and the reliance on national exception for processing of health data.[25] However, one issue that may arise with the first structure in relation to the data warehouse may be where retrospective data are involved, of which consent has not been specifically obtained for the transfer to the p-medicine data warehouse. This has the effect of watering down the use of consent as a legal basis for the data transfer to the p-medicine data warehouse. In this case however, reliance will be placed on the other safety net components of anonymisation and national exception. These are relevant because national laws within the EU permit the processing of sensitive personal data for scientific research, and in most cases, anonymisation of such data is a basic condition if the research purposes allow for this, or should take place as early as possible.[26]

A related issue to be furthermore considered as a result of the international nature of the p-medicine project involving a partner from Japan, is whether access to or transfer of data to such a third country can be permitted through the p-medicine data warehouse.[27] This is relevant because under the data protection directive, transfer of data to a third country is prohibited except such a country provides adequate level of personal data protection or any of the legal exceptions permits such transfer.[28] Japan has not been assessed by the EC as having adequate level of personal data protection, but nevertheless, data can be transferred to such country based on other legal bases.[29] The Article 29 Working Party has suggested that data in this case should be transferred "in anonymised or at least pseudonymised form" because of the sensitive nature of the data to be processed.[30] The p-medicine warehouse is *defacto* anonymous and is in line with this opinion. This affords an avenue for further safeguards and easy transfer of data for research purposes. In addition to *defacto* anonymisation, contractual safeguard, such as the standard contracts issued by the EC was also adopted to bind the parties on the appropriate use of the data.[31]

Other relevant precondition for the establishment of a medical research data warehouse relevant to p-medicine at this stage will include the institutionalisation of a technical and organisational measure for the security of the database.[32] This is to ensure the confidentiality, integrity and availability of the data in the warehouse, especially that no unauthorised persons gain access to it. The overall security architecture of the p-medicine project has previously been analysed in D5.1. We will subsequently look more specifically at the security framework for the integrated data warehouse.

[25] See Deliverable D5.1, op. cit.

[26] See for example the UK, Germany and Belgium data protection laws. In addition to this, it is also a p-medicine requirement to obtain a declaration from the data transferor indicating that it is permitted to do such transfer under relevant laws applicable to it.

[27] See Articles 25 and 26 of the Directive 95/46/EC (Data Protection Directive).

[28] See EC, Commission decisions on the adequacy of the protection of personal data in third countries, available at: http://ec.europa.eu/justice/policies/privacy/thridcountries/index_en.htm.

[29] Apart from the adequacy assessment transfer of personal data to a third country that does not afford adequate level of data protection is possible if the data controller could adduce adequate safeguard for the transfer in the form of appropriate contractual clauses and binding corporate rules. See, Christopher Kuner, *European Data Protection Law - Corporate Compliance and Regulation*, 2007.

[30] Article 29 Working Party, Working Document on the processing of personal data relating to health in electronic health records (EHR), WP 131, p. 19.

[31] By having defacto anonymous data removes obligation to use EC approved contractual clauses, but because of the third party benefits accruable to the patients if such clauses are use, we find it desirable to use them for patients empowerment.

[32] See Article 17 of the Directive 95/46/EC (Data Protection Directive).

4.2.1 The Clinical Trial Directive (Directive 2001/20/EC)

The Clinical Trial Directive is another legal instrument to be considered when establishing a database for medical research. This is because it lays a framework for safeguarding clinical trial subjects. The scope of the Directive borders on good clinical practice, defined as a set of internationally recognised ethical and scientific quality requirements that must be observed for designing, conducting, recording and reporting clinical trials that involve the participation of human subjects.[33] It envisages that compliance with this good practice will provide assurance that the rights, safety and well-being of trial subjects are protected, and that the results of the clinical trials are credible.[34]

While the Directive did not specifically elaborate on the establishment of databases used in clinical trials, it however provides that clinical trial may be undertaken only if *inter alia*, the rights of the trial subjects to physical and mental integrity, to privacy and to the protection of the data concerning them in accordance with Directive 95/46/EC are safeguarded.[35] It can be deduced from this that the Clinical Trial Directive recognises the pre-eminence of the Data Protection Directive in safeguarding personal data, including when processed in clinical trials, and as such, it did not elaborate further on such. In view of this, the relationship between the p-medicine database and the Data Protection Directive has been considered above and remains relevant for this provision.

4.2.2 The Good Clinical Practice Directive (Directive 2005/28/EC)

Apart from the Directive 2001/20/EC, there are other legislative and non-legislative documents that may impact a medical research data warehouse such as the Directive 2005/28/EC and the EMEA Note for Guidance on Good Clinical Practices.[36] Directive 2005/28/EC lays down principles and detailed guidelines for good clinical practice in respect to investigational medicinal products for human use.[37] It supplements the Directive 2001/20/EC. While this Directive did not also deal on specific requirements for the establishment of a data warehouse for medical research, it provides in its Article 5 that: “All clinical trial information shall be recorded, handled, and stored in such a way that it can be accurately reported, interpreted and verified, while the confidentiality of records of the trial subjects remains protected.” This indicates that the handling and storage of data in a medical research data warehouse should protect the confidentiality of the data. This for example requires technical and organisational measures such as encryption of the data at rest. The EMEA Notes on the other hand stem from the ICH guidelines and will be discussed in the next section.

4.2.3 The International Conference on Hamonisation (ICH) Topic E6 (R1) Guideline for Good Clinical Practice Step 5

ICH GCP Guideline is an international ethical and scientific quality standard for designing, conducting, recording and reporting trials that involve the participation of human subjects. The objective of the ICH GCP Guideline is to provide a unified standard for the European Union (EU), Japan and the United States to facilitate the mutual acceptance of clinical data by the regulatory authorities in these jurisdictions.[38] While this

[33] Article 1 of the EC Directive 2001/20/EC (Clinical Trial Directive).

[34] Ibid.

[35] Art 3(2)(c) of the EC Directive 2001/20/EC (Clinical Trial Directive).

[36] ICH Topic E 6 (R1) Guideline for Good Clinical Practice, CPMP/ICH/135/95, July 2002.

[37] This Directive applies where the database is used for investigational medicinal products for human use.

[38] ICH Topic E 6 (R1) Guideline for Good Clinical Practice, CPMP/ICH/135/95, July 2002, p. 5.

document stipulates guidelines for generating clinical trial data as well as scientific quality standard for designing, conducting, recording, storing and reporting trials that involve the participation of human subjects, it did not directly address issues of integrating these trial data into a data warehouse. However, its impact is felt in its overall attempt to safeguard the rights of the clinical trial participant. It specifically provides that all clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification.[39] Furthermore, it states that the confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).[40] Generally, this guideline has been reflected in the Clinical Trial Directive.

4.2.4 The Council of Europe, Committee of Ministers, Recommendation No. R (97) 5 on the Protection of Medical Data (Feb. 13, 1997)

The Council of Europe's Committee of Ministers has adopted several recommendations on the basis of Convention No. 108[41] which aim at ensuring that the collection and processing of personal data including sensitive data relating to health are always carried out in accordance with the principles of the Convention. In 1981, it adopted Recommendation No. R (81) 1 on the regulation of automated medical data banks which was replaced in 1997 by the present Recommendation No. R (97) 5. The recommendation recognises that it is desirable to regulate the collection and processing of medical data, as well as to safeguard the confidentiality and security of personal data regarding health. It also aims at ensuring that such data are used subject to the rights and fundamental freedoms of the individuals involved, and in particular, their right to privacy. This is basically in view of the increasing use of automatic systems in processing of medical data, not only for medical care and research, but also for non-healthcare purposes. While most of the principal provisions of this recommendation have been enshrined in Directive 95/46/EC, it however sheds more light on the security mechanism relating to a medical data bank as envisaged in the p-medicine data warehouse.

In order to ensure in particular the confidentiality, integrity and accuracy of processed data, as well as the protection of patients, the recommendation provides that appropriate measures should be taken:[42]

a. to prevent any unauthorised person from having access to installations used for processing medical data (control of the entrance to installations);

b. to prevent data media from being read, copied, altered or removed by unauthorised persons (control of data media);

c. to prevent the unauthorised entry of data into the information system, and any unauthorised consultation, modification or deletion of processed medical data (memory control);

d. to prevent automated data processing systems from being used by unauthorised persons by means of data transmission equipment (control of utilisation);

e. with a view to, on the one hand, selective access to data and, on the other hand, the security of the medical data, to ensure that the processing as a general rule is so designed as to enable the separation of:

   - identifiers and data relating to the identity of persons;

[39] Ibid, Clause 2.10.

[40] Ibid, Clause 2.11.

[41] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data.

[42] See Principle 9 of the Council of Europe, Committee of Ministers, Recommendation No. R (97) 5 on the Protection of Medical Data (Feb. 13, 1997).

- administrative data;
- medical data;
- social data;
- genetic data (access control);

f. to guarantee the possibility of checking and ascertaining to which persons or bodies personal data can be communicated by data transmission equipment (control of communication);

g. to guarantee that it is possible to check and establish a posteriori who has had access to the system and what personal data have been introduced into the information system, when and by whom (control of data introduction);

h. to prevent the unauthorised reading, copying, alteration or deletion of personal data during the communication of personal data and the transport of data media (control of transport);

i. to safeguard data by making security copies (availability control).

The recommendation further provides that where necessary, controllers of files processing medical data should draw up appropriate internal regulations for such processing, as well as appoint an independent person responsible for security of the information systems and data protection. Above all, whenever possible, medical data used for scientific research purposes should be anonymous.[43] These recommendations are important to the p-medicine data warehouse and will be reflected in the granular guidelines for its use. It should be noted however, that most of the recommendations have been implemented in the general data protection and security framework of the project such as the appointment of a central data protection authority – CDP.[44]

4.3 Ethical issues surrounding medical research data warehousing and mining

The digital revolution is rapidly changing how medical care and research are conducted in the present time. IT has provided the capabilities of using information systems to pool large data relating to patients from different sources. This has the effect of rapidly changing the landscape of patients' privacy protection as well as the traditional mechanisms used to protect research participants such as consent and anonymisation of datasets.[45] Stolba (2006) reflects that the main ethical concern of federated data warehouses for evidence-based medical purposes is to provide mechanisms and policies to preserve patient privacy while delivering a huge decision support system for research purposes and health care support.[46] It is important to realise that even with linked or unlinked anonymous data, there may still be the potential to deduce individual's identities through combinations of information, held either by people handling the research data or by those who see the results. This problem of connectivity applies where one piece of data on its own may not provide identification, but may do so when used together with another.[47] Thus it is important to consider ethical guidelines when setting up a medical research database.

[43] See Principle 12 of the Recommendation.

[44] See Deliverable D5.1

[45] Jane Kaye et al, 'Data Sharing in Genomics – Re-shaping Scientific Practice', Nat Rev Genet, 2009, 10(5): pp. 331 – 335.

[46] Nevena Stolba, op cit, p.3.

[47] Mags McGeever, 'Sharing Medical Data', 2006, available at: http://www.dcc.ac.uk/resources/briefing-papers/legal-watch-papers/sharing-medical-data#5 (accessed 25 April 2012).

4.3.1 The Declaration of Helsinki, 2008

The World Medical Association (WMA) General Assembly has extensively considered ethical issues relating to health databases. It developed the Declaration of Helsinki[48] as a statement of ethical principles for medical research involving human subjects, including research on identifiable human material and data. One basic condition for participating in a clinical trial under the Declaration is that the voluntary and informed consent of the trial participants must be sought and obtained. It provides that: "For medical research using identifiable human material or data, physicians must normally seek consent for the collection, analysis, storage and/or reuse of the data."[49] Where consent would be impossible or impractical to obtain for such research, approval of a research ethics committee must be obtained. This is the ethical corner stone for integrating any medical research database. The Declaration also imposes a duty on physicians who participate in medical research to protect the life, health, dignity, integrity, right to self-determination, privacy, and confidentiality of personal information of research subjects.[50] They shall also take every precaution to protect the privacy of research subjects and the confidentiality of their personal information and to minimize the impact of the study on their physical, mental and social integrity.[51]

4.3.2 WMA Declaration on Ethical Considerations regarding Health Databases, 2002

In the 2002 WMA Declaration on Ethical Considerations regarding Health Databases, the Association reiterated the advantages of maintaining an accurate health database, especially as a source of secondary use of health information.[52] In order to safeguard this valuable source of information, the WMA developed principles for health database controllers. Generally, these principles mirror the data protection and security considerations discussed above, but also include salient ethical issues that are specific to health databases. These considerations should be taken into account by anyone establishing such a health-related database, and broadly include:

i. Ensuring that patients have the right of access to their information in the health database, as well as the right to decide to delete such information;

ii. Creating appropriate security mechanism to safeguard the health database including appointing a guardian for such database;

iii. Informing the patients that their information will be stored in a database, as well as obtaining their consent if such information will be disclosed to a third party;

iv. Obtaining approval from an ethical review committee if patients' data are to be used for research, and using such data for only the purpose it was approved;

v. De-identifying or coding health data used for secondary purposes;

vi. Ensuring the accuracy and up-to-date of the data held in the database;

[48] WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects, 59th WMA General Assembly, Seoul, October 2008.
[49] Ibid, Clause 25.
[50] Ibid, Clause 11.
[51] Ibid, Clause 23.
[52] WMA Declaration on Ethical Considerations regarding Health Databases, 53rd WMA General Assembly, Washington, DC, October 2002.

vii. Documentation of events transpiring in the database, including maintaining an audit system;

viii. Establishing a procedure for addressing inquiries and complaints relating to the database.

These ethical concerns were considered in building the general data protection structure of the p-medicine project, however, combining databases can raise other important practical, scientific, and ethical concerns that may be unrelated to the original consent process. Kerp *et al* (2008) have reviewed these complications, pointing out for instance that where retrospective data are stored in a medical research data warehouse, such data may have been collected without authorisation that meets today's standards for informed consent.[53] Research participants may for example not have consented to participation in genetics research in general, to inclusion in genetics databases specifically, or to use of their samples in genetic analyses that were unanticipated, unknown, or nonexistent at the time samples were collected. In the same vein, participants who consented to the processing of their data for a particular study, or inclusion in a particular database, may not have consented to the "secondary uses" of those data for unrelated research, or for use by third parties.[54] These raise issues such as the future use of research databases, and the risk of re-identification of not only individual participants, but also their families and groups. A similar issue to be considered in this case will be how to effectively erase a participant's data when consent is withdrawn, not only from the data warehouse, but also from other places where copies may have been made.[55]

The supplementary guidelines that will be introduced in this report will give a fine-grained protection to the participant through best practices to be observed by the researchers when using the database. In the first place, a prospective consent form has been produced which informs the participants of the use of the data warehouse. Data in the p-medicine data warehouse will be destroyed after the conclusion of the research which removes the concern of using the database for purposes which the patients have not consented for.

Data in the p-medicine warehouse are by default anonymous, however, privacy enhancing data mining technology will be introduced in the architecture of the database to ensure that sensitive data that could link to an identifiable individual should not be published in the data mining result. Furthermore, end users of the data warehouse are by contract not allowed to embark on re-identification process, with penalty for violation. This serves as another protective factor in the p-medicine safety net. However, a salient issue envisaged in this scenario is the practical possibility of deleting all traces of a patient's data from the warehouse as well as other places where copies may be held, in order to respect the wishes of the patient who has withdrawn consent. While this is possible from the data warehouse, it remains for each researcher holding such data to comply with the deletion not only by virtue of the end user contract, but also an audit check can be performed to ensure that the request has been enforced.

4.4 Data security in medical research data warehouse

An often profound issue to be addressed by data controllers is the security of the data they are processing. This is important so as not to breach the Data Protection Directive, as well as to maintain public trust and confidence. Directive 95/46/EC states in its Art. 17 that Member States shall provide that the controller/processor "must implement appropriate technical and

[53] David Karp et al, 'Ethical and practical issues associated with aggregating databases', PLoS Medicine, 2008, vol. 5, Issue 9, p. 1333.
[54] Ibid.
[55] Ibid

organisational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access, in particular where the processing involves the transmission of data over a network, and against all other unlawful forms of processing". While D5.1 has elaborated on these measures, technical measures generally deal with the practical use of the methods implemented to secure the data being processed such as access control, the use of encryption, secure connections, firewalls or access by biometric identification or similar methods. Organisational measures on the other hand refer to a set of rules to enable data security by regulating authorisation and authentication procedures, such as access policies and identity management for the IT system processing the data.[56]

In addition to the above, the security model recommended in Recommendation No. R (97) 5 and the WMA Declaration mentioned earlier are also very relevant for the data warehouse and mining because they specifically apply to a medical database. These documents show in a nutshell that certain security measures must be in place in a health database such as:

a) Access control
b) Management system for the database
c) Secure transmission
d) Audit or log system
e) Anonymisation or psudonymisation of data
f) Constant review of the security mechanism
g) Conservation of data.

Another area that also deserves consideration is the security of the cloud network used for the storage of the database. While it is still debated whether critical information should be stored in the cloud, it is beyond the scope of this analysis to delve into such debates. However, the security level in cloud architecture should be considered, especially, when a public cloud is used for the storage. Issues such as loss of data control, jurisdiction, security breaches and vendor lock-in have in most cases been seen as militating factors in using the cloud.[57] It is recommended in this respect to use a private cloud infrastructure where data control will not be completely relinquished to a third party.

4.4.1 Security framework of the P-medicine data warehouse

The p-medicine storage system will be properly secured to provide reliable data preservation functionality resistant to any potential failures and data theft. In the first instance, data are stored in a fully distributed manner such as in multiple, geographically dispersed Storage Nodes.[58] Security measures, such as access control, secure transmission and auditing are addressed by the p-medicine security framework as defined in Deliverable 5.1 and 3.4. Deliverable 3.4 also explains how the different services, such as the data warehouse, within p-medicine can be integrated into the security framework.

[56] See details in Deliverable D5.1.

[57] See for example: Sebastien Lapointe, 'Legal Cloud Computing: Concepts and Ramifications', Legal IT 2010; ENISA, Cloud computing: benefits, risks and recommendations for information security, 2009.

[58] See Deliverable D8.1.3.

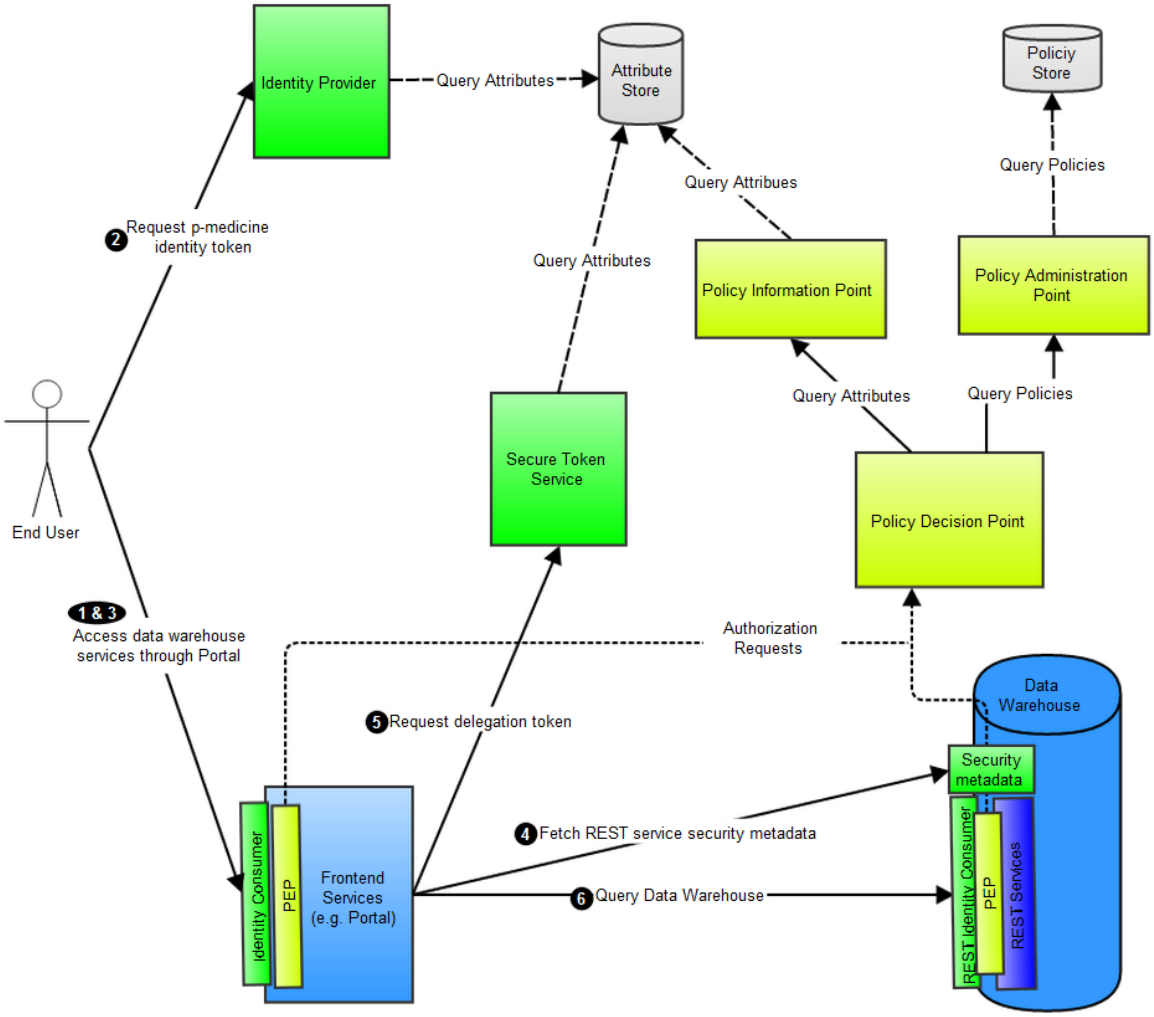

Figure 1: Overview of security framework with integrated data warehouse

The data warehouse integrates with the following authentication components: an identity provider (IdP), a secure token service (STS), web and REST identity consumers and security metadata. The IdP is a service provider within p-medicine responsible for authentication. It provides identity tokens to other service providers by which end users are identified. When an end user visits a service provider (SP), such as the portal (1), through a web browser, the browser is redirected to the IdP (2) if there is no active security context on the SP. Once authenticated, or if there is an active SSO session, the IdP will issue an identity token and redirect the browser to the SP (3) passing through the issued token.

An identity consumer is a software component integrated in a SP that consumes the tokens provided by the IdP. It will verify the received token and pass it to the application layer if valid. The SP can then set up a security context for the authenticated end user.

The STS is responsible for issuing identity and delegation tokens for REST clients. Before calling a REST service, a REST client will download (4) the REST service's metadata file

(e.g. WSDL or WADL[59]) which specifies the REST interface. This metadata file is annotated with security policies defining the type of security token accepted by the REST service. The policies also define which STS the client needs to call in order to fetch such a token. The REST client then requests a token (5) from the STS and sends that token together with the REST call embedded in a HTTP authorisation header to the REST service (6).

An end user typically does not access the data warehouse directly, but instead queries the data warehouse through portlets on the p-medicine portal (Liferay).

- (1, 2 & 3) An end user thus initiates the flow by accessing the p-medicine Liferay portal. An identity token is then fetched from the IdP by redirecting the browser as described above.
- The end user can query the data warehouse through portlets installed on the p-medicine portal. Such a portlet would then call the data warehouse acting in name of the end user.
- After fetching the REST services metadata (4), the portlet requests from the STS (5) a delegation token passing through the end user's identity token.
- The portlet finally calls the data warehouse REST service (6) passing through the delegation token. This token states that the portlet is calling the REST services in name of the end user.

Through the delegation token, which contains the identity of the end user, the data warehouse can take decisions on whether a given user is allowed to query the data warehouse. Next to that, whenever sensitive pseudonymised patient data is accessed, the data warehouse should query the p-medicine central policy decision point (PDP) to find out whether the given end user is allowed access to this sensitive information.

The PDP will initially not take fine grained access control decisions. A user has access to sensitive pseudonymised data when he has signed the Contract on Data Protection and Data Security within p-medicine. As such the PDP enforces that only users who have signed these contracts are able to access sensitive pseudonymised data.

In a later stage though access control should be more fine-grained so that users can only query the patients on which they have access. For example, if data sources only want their data to be used for a specific purpose, this data should only be queryable for that particular purpose. If a query returns patients as result, all patients which should not be accessible can be filtered out. Securing queries that use the whole dataset to return a specific result, e.g. the medium age of all patients, is more challenging. In such a case the query needs to be modified so that the result is only calculated based on the data the user has access to. Further research is needed on how this can be achieved with RDF as query language.

[59] See Deliverable D3.4 for detailed information.

5 Overview of the architecture of the P-medicine data warehouse and data mining

5.1 Introduction

A data warehouse is a repository for securely storing and maintaining data from diverse sources integrated semantically to enable reporting and analysis.[60] Such a warehouse with a distributed, highly accessible environment has been adopted as the central repository of p-medicine's data. The p-medicine data warehouse is made up of input data collected from clinical trials, the patient information system, experimental works and other tools meant for data management and mining purposes to enable a personalised medical treatment. Precisely, the objectives of having a p-medicine data warehouse are described as follows:[61]

1. To develop federated data warehouse infrastructure components to store the multiple different types of medical data produced in this project
2. To develop storage services for large data objects
3. To develop mechanisms for ensuring reliability and auditability of the data
4. To develop and deploy suitable programmatic interfaces, to allow the different types of data to be uploaded to the warehouse via tools developed in other work packages
5. To develop a service integrated into the web portal to allow users to search for and view available data
6. To develop federated capability-based secure access mechanisms compliant with the legal and ethical framework of the project
7. To integrate security- and role-based access, based on the legal and ethical framework developed in the project
8. To integrate the disparate types of data produced and available in the project using appropriate ontological tools.

5.2 Architecture of the p-medicine data warehouse

The p-medicine data warehouse is built on a federated database architecture, where data are stored in physically separated datasets that use information technology to provide a virtual common dataset.[62] The main concept is to make multiple storage systems to cooperate as a team to store data in the aggregated storage resources (LUNs or volumes) of all federated members.[63] This is similar to the system defined by Sheth and Larson (1990) as a collection of cooperating database systems that are autonomous and possibly heterogeneous.[64] It is a functional data warehouse, where heterogeneous data warehouses are integrated into a single unit from the conceptual point of view, using a unique common conceptual model.[65] Federation addresses the limitations presented by a single storage system, such as capacity, performance and maintenance availability. Thus, in p-medicine, data will be pushed from exporting databases into a main p-medicine data warehouse, with data stored across many systems hosted by participating partners for optimisation purposes such as load balancing and exploitation of additional resources, among others. Deliverable D8.1.3 has outlined the capabilities of federation including: storage expansion, storage

[60] Deliverable D7.1: Report on overall system design including VPH-Share D2.2 and indicating its impact, p. 6.

[61] Ibid, pp. 6-7. See also the Description of Work.

[62] Heather Piwowar *et al*,, 'Towards a data sharing culture: recommendations for leadership from academic health centers', PLoS Medicine, 2008, vol. 5, Issue 9, p. 1317.

[63] P-medicine Deliverable D8.1.3, Report on federated Storage Services.

[64] Amit Sheth and James Larson, 'Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases', ACM Computing Surveys, 1990, Vol. 22, No. 3, pp. 183-185, available at: http://knoesis.wright.edu/library/download/SL90.pdf (accessed 7 June 2012).

[65] Ibid.

migration, safe system upgrades, load balancing and storage tiering.[66] The OpenStack cloud computing interface will be used in the p-medicine data warehouse infrastructure for easy integration, storage and other processing capabilities.[67]

While other formats of data warehousing exist such as centralised formats, the federated architecture has been adopted for this project because of its scalability in storing large amount of data, as well as distribution of server load. Local copies of remotely-held data can also be maintained to further ensure quality of service.[68] Stolba, Banek and Tjoa (2006) suggest that federated data warehouse is best used when several independent organisations share their data for mutual purposes, for example when highly confidential healthcare records are involved.[69] Article 29 Working Party also appears to favour this storage format in their opinion on the processing of personal data relating to health in electronic health records.[70] In sum, the purpose of the federation is to unite the data assets of local data warehouses in order to gain a broader base for knowledge discovery and data mining, while keeping their physical separation.[71]

The p-medicine data sources are diverse, and made up of heterogeneous data of which some of the datasets contain personal data. Generally, the p-medicine data warehouse sources include:

(a) Clinical trials
(b) Patient information systems
(c) Experimental work
(d) p-medicine tools
(e) Obtima clinical trial management system
(f) Modelling and data mining tools
(g) Patient empowerment tools.[72]

Datasets coming from (a) to (c) above usually contain personal data, and the data subject envisaged here are not only patients, but also include doctors, relatives of patients and persons involved in the data mining and development of the toolkit. Attributes of data from the above sources may generally include personal identifiers such as: name, date of birth, gender, profession, address (street, postal code and city), etc. However, these datasets have been *defacto* anonymised to remove personal identifiers before being pushed into the p-medicine data warehouse. In the first instance, data are pseudonymised at the hospital database. Another set of pseudonymisation is performed on the data using the CATS before they are pushed to the p-medicine data warehouse, resulting to having only a defacto anonymous data in the data warehouse.[73]

The above system eliminates the risks of identifying the patients. Although complete anonymisation has not being adopted in this project because of instances where there may be need to re-identify patients in order to give them the best treatment in the event that the research within p-medicine reveals that a certain therapy is highly effective for a certain

[66] P-medicine Deliverable D8.1.3, Report on federated Storage Services, p 7.
[67] Ibid.
[68] Deliverable D7.1, p. 9.
[69] Nevana Stolba, op. cit, p. 2.
[70] However, the Working Party is also of the opinion that a centralised format is better than where patients manage the data. See, Article 29 Working Party, Working Document on the processing of personal data relating to health in electronic health records (EHR), WP 131, adopted in February 2007, p. 17.
[71] Nevana Stolba, op. cit, p.6.
[72] Deliverable D7.1, p. 7.
[73] For more details see Deliverable D 5.1.

disease.[74] In these instances, a procedure for such re-identification including the use of a Trusted Third Party has been put in place so that researchers do not directly contact the patients.[75]

5.3 Data mining patterns within P-medicine

The data mining tools developed for p-medicine aim at translating data mining techniques into practical solutions for the analysis and prediction of patient data. In order to facilitate a seamless transition of methods from data mining research to medical research, a pattern-based approach will be followed. This primary goal of having data mining patterns is to enable the re-usability and standardisation of generic data mining scenarios in the analysis of clinical data, while ensuring that all specialisations of these scenarios to specific problem instances can be executed in the architectural framework of p-medicine.[76] In p-medicine, data mining patterns will be employed to apply research on privacy-preserving data mining that incorporates guarantees about privacy and data security directly into the algorithm. The p-medicine architecture will allow researchers to perform data mining tasks directly. The goal of this task is to use data mining to make a collaborative research environment such as the p-medicine system more flexible and more adaptive to the user's needs, for decision support, data set selection and patient empowerment. There is also a literature mining capability in predicting clinical outcomes.

5.3.1 Privacy enhancing data mining within P-medicine

The combination of data warehousing and mining in medical research have shown great innovation in the application of ICT in the health sector. Techniques recently developed for evaluation of stored clinical data have lead to discovery of trends and patterns hidden within these data that could significantly enhance the understanding of disease progression and management.[77] Data mining technology has the capabilities of extracting implicit, previously unknown, and potentially useful information from large datasets.[78]

Data in the p-medicine warehouse will be mined through the p-medicine portal using a DM Webapp.[79] While valuable information could be extracted in the p-medicine data warehousing through mining, it should be safeguarded so that sensitive data that could link to an identifiable individual should not be published in the data mining result. One way of achieving this is through implementing a privacy preserving measures during query processing in the federated data warehouse architecture. The EC considers that "…the use of appropriate technological measures is an essential complement to legal means and should be an integral part in any efforts to achieve a sufficient level of privacy protection…:" The use of PETs should result in making breaches of certain data protection rules more difficult and/or helping to detect them.[80] Before describing the details and the benefits of the

[74] Ibid, p. 69.

[75] Ibid.

[76] P-medicine Deliverable D8.1.1, Specification of the interaction with the VPH toolkit, p. 28.

[77] Jonathan Prather, 'Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse', *Proc AMIA Annu Fall Symp.*, 1997, p. 101.

[78] Nevena Stolba and A Min Tjoa, ' Relevance of Data Warehousing and Data Mining in the Field of Evidence-based Medicine to Support Health Decision Making' *World Academy of Science, Engineering and Technology* 2005, p. 192.

[79] See P-medicine Deliverable D11.1.

[80] Communication from the Commission to the European Parliament and the Council on Promoting Data Protection by Privacy Enhancing Technologies (PETs), 2007, p.3, availabe at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2007:0228:FIN:EN:PDF (accessed 23 May 2012).

privacy-preserving data mining services, we will briefly review two central notions from the area of privacy-preserving data analysis: *k*-anonymity and *l*-diversity

5.3.2 The notions of k-anonymity and l-diversity

The idea of k-anonymity was introduced by Samarati and Sweeney (Samarati&Sweeney, 1998, Samarati 2001, Sweeney 2002). The general approach assumes tabular data containing sensitive information about individuals. The basic idea is that for every combination of attributes and values, either none of the records must match this combination, or at least *k* different records must match. Such combinations of attributes and values are called quasi-identifiers. We will now illustrate the concept of quasi-identifiers and *k*-anonymity using following table:

Date_of_Birth	***Marital Status***	***Health_Problem***
2.1.1984	Single	Hypertension
12.12.1984	Single	Cancer
2.12.1979	Divorced	Cancer
2.12.1979	Divorced	Chest Pain
30.8.1984	Divorced	Cancer
6.4.1984	Maried	Hypertension
6.4.1984	Maried	Hypertension

Table 1: A table showing a dataset containing some personal identifiers.

In this example, the quasi-identifier “Date_of_Birth=2.1.1984 and Marital Status = Single” is only matched by the first record. Hence, it uniquely characterizes this record. This represents a threat to privacy, because an attacker might de-anonymise the record by linking the dataset with other data source. Potential data source that could be linked include data from public authorities, phone directories, social media, data from the web or the like.

If, however, *k* different records cannot be distinguished by any combination of attributes and values, then no linking attack can succeed in de-anonymising these records. In the above table, for example the last two records cannot be distinguished. The central idea of *k*-anonymity is to ensure that every quasi-identifier is matched by at least *k* records (or by zero records).

While *k*-anonymity is a very useful concept, it still suffers from one drawback: even if a quasi-identifier does match multiple records, these records might all have the same value for one sensitive attribute. In the example, even though the last two records are indistinguishable with respect to the first two attributes, they share the same value for the sensitive attribute “Health_Problem”. This can allow an attacker to draw conclusions about the sensitive attributes of a particular individual if the *k*-anonymous data is linked with other data via the non-sensitive attributes – even if the attacker cannot uniquely associate the individual with one particular record.

The problem described above was first pointed out by Machanavajjhala et al, in 2006. As a remedy, the authors have proposed a new concept termed "*l*-diversity". The idea is that "All tuples that share the same value of their quasi-identifiers should have diverse values for their sensitive attributes" That is, every quasi-identifier block is required to contain at least *l* "well-represented" values for the sensitive attributes. Although there are several ways to make the notion of "well-represented" precise, arguably the most intuitive and simple way is to require that the entropy of the distribution of values of every quasi-identifier block is no lower than *log(l).* This definition, which is also called "Entropy *l*-Diversity", implies that every block contains at least *l* different values. In the above example, the block corresponding to the quasi-identifier "Date_of_Birth=6.4.1984 and Marital Status = Maried" is 2-anonymous but not 2-diverse, while the block for "Date_of_Birth=2.12.1979 and Marital Status = Divorced" is 2-diverse (and 2-anonymous).

Different algorithms have been proposed to transform a given dataset into a *k*-anonymous or *l*-diverse dataset. Although the technical details differ, all these approaches share the same idea: to either generalize the values in the table, or suppress individual records, until the resulting table satisfies the anonymity constraints. In the above example, one option for generalization would be to replace the exact date of birth by the year of birth – this would result in same values for the first two and the last three records. Similarly, the marital status could be generalized in a way that only "never_maried" and "been_maried" are distinguished. This would allow generalizing the original table to the following 2-diverse table.

Year_of_Birth	***Marital Status***	***HealthProblem***
1984	Single	Hypertension
1984	Single	Cancer
1979	Been_Maried	Cancer
1979	Been_Maried	Chest Pain
1984	Been_Maried	Cancer
1984	Been_Maried	Hypertension
1984	Been_Maried	Hypertension

Table 2: A table showing a transformed dataset.

5.3.3 Data-Mining under k-anonymity and l-diversity-Constraints

While the above-described concepts of *k*-anonymity and *l*-diversity were developed in the context of tabular data, they also concern data-mining, and in particular the outcome of a data mining process. In fact, a data-mining algorithm like a rule learner can, when applied to the above dataset, come up with rules like the following:

Date_of_Birth="2.1.1984" and Marital_Status="Single"

→ Health_Problem = Hypertension" (support = 1, precision = 100%)

It is obvious that the above rule, together with the associated statistics about its precision and support, represents a threat to privacy. While the above privacy-threat is obvious, for

more complex data mining outcomes (like sets of rules, or complex decision trees), checking whether the set of patterns represents a threat to privacy can become much more complex (see Atzori et al, 2008).

In general, two different approaches exist to ensure that the outcome of data-mining will satisfy privacy-constraints:

- Anonymise-and-mine:
  Here, in a first step the tabular data is turned into a *k*-anonymous or *l*-diverse dataset. In a second step, a standard data-mining algorithm is applied. The result is guaranteed to satisfy the privacy guarantees due to the fact that the data algorithm only had access to anonymised data.

- Privacy-Preserving Data-Mining Algorithms:
  Here, the Data-Mining is performed directly on the original data. The algorithms, however, guarantee that the outcome satisfies *k*-anonymity and/or *l*-diversity constraints. Several privacy-preserving algorithms have been proposed, which allow applying a wide range of data-mining approaches under privacy constraints (e.g. [Friedman et al 2006]). The big advantage of this (more complex) approach is that compared to the earlier option the patterns tend to be of higher quality. The reason is that the algorithm has access to the full information in the original data; it is only in case that the patterns present a threat to privacy that some specific parts of the patterns need to be modified (e.g. by local generalization of the problematic part of the pattern).

5.3.4 Proposed Approach

In the scope of p-medicine, we plan to provide *l*-anonymization services to support the publication of data mining results obtained from the p-medicine data. Say a researcher is doing work in p-medicine under p-medicine's legal framework and that at some point, she wants to publish her findings in a journal. This means she exports some information that is derived from the p-medicine data outside of the P-medicine legal framework. If this information takes the form of complex data mining models, e.g. rules, it may not be obvious that they contain no critical information. *L*-diverse datasets and *l*-diverse patterns are a way to provide formal guarantees about the amount of information that can be derived from the publication.

Actually, we plan to realize two different services, which are based on the two approaches sketched above:

- First, we plan to incorporate some algorithms into the data warehouse which would turn the data into an *l*-diverse dataset. The resulting data could be made public with the formal guarantee that it is *l*-diverse. Moreover, we plan to provide access to standard data mining algorithms on the anonymised data. Here, we will focus on rule-based approaches like subgroup discovery, rule learners or decision tree, which produce patterns that can directly be interpreted by a human expert. Like the *l*-diverse datasets, any data mining results obtained this way could also be made public with the *l*-diversity guarantee.
- Second, we plan to incorporate privacy-preserving algorithms which mine for *l*-diverse patterns. Again, the outcome of a ***single*** data mining analysis could be made public with the same guarantees as above. However, unlike in the earlier option, if patterns found in a sequence of independent analyses are put together, then *l*-diversity would no longer be guaranteed.

The reason why we head for both approaches is that both have their advantages. The second approach is likely to produce higher-quality patterns, but this comes at a price: if a single analysis is done, then the resulting set of patterns is guaranteed to satisfy the privacy constraints; however, if several different analyses are done, then the guarantee of the second approach does to carry over to the union of the outcomes. These must hence be kept separately, which makes the publication of several results problematic. A single publication of results is still possible, but if several publications are planned, then the legal consequences must be considered carefully. Alternatively, the first approach can be taken, which provides the same guarantees for unions of outcomes as for single outcomes.

5.4 Access to the data warehouse

The data warehouse is accessed through the p-medicine portal which is a secure unified access point to the tools, services and data shared in the p-medicine environment in the form of a web-based user interface. A user has access to the portal if he presents a valid p-medicine identity token. Such a token can be obtained from the p-medicine identity provider once authenticated. Access to portlets that query the data warehouse will be limited to specific p-medicine roles (to be specified).

Authorisation is enforced on the data warehouse. The data warehouse is responsible to send an authorisation request to the p-medicine central Policy Decision Point (PDP) for each received data access query. The PDP will allow or deny access according to access rules defined in the p-medicine policy files. The PDP will for example only give partners who have signed the contracts as defined in D5.1 access to the sensitive patient data.

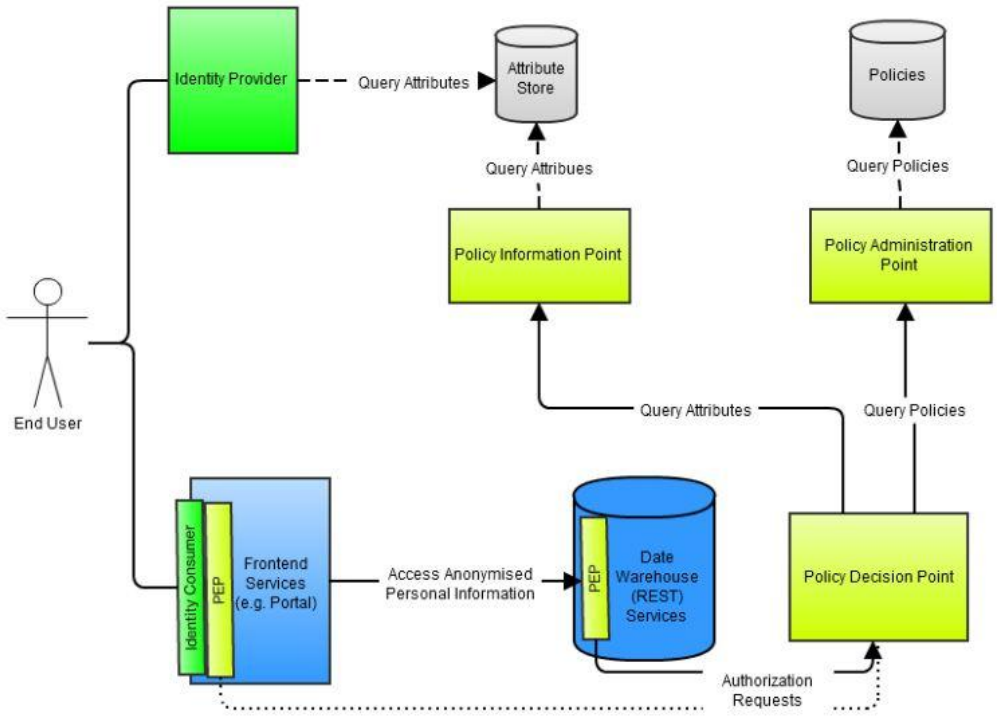

Figure 2: p-medicine access policy

6 Intellectual property rights in the P-medicine database

6.1 Introduction

Ownership of the intellectual property in a medical research database and associated rights therein have been subject of controversy over the years. Many hospitals consider the records in their systems to be their property, whereas many patients argue that their medical information is their own. This scenario becomes more complicated when many institutions are involved in a medical research, with multiple interests, such as the funders and researchers in the project. No less controversial in some instances is where the content of the database has been assimilated from other databases.[81] While a clear resolution of the issues raised here can be made by contract involving the parties, policies or regulations on these issues may differ substantially in different states.[82]

In this section, we will specifically look at the issues of intellectual property rights in the background and foreground materials in the p-medicine database, as well as the legal protection of such database under EU law.

6.2 Who has the intellectual property rights in the p-medicine data warehouse?

Interesting arguments have been going over the years as to who should posses the intellectual property rights of data used in clinical trials – patients, hospitals, sponsors of the trial, researchers/participants in the project, the public, etc. In deciding this issue, a lot of factors would have to be considered, including for example the fundamentals of trial design, protocol management and regulatory oversight.[83] Equally relevant here will be how to control these data if too many persons are involved in their management. While there may be divergent opinions as to who should have proprietary rights in medical research databases, it is settled that patients should have the right of access to their medical records, even when used for research purposes.[84] By virtue of Article 12 of the Data Protection Directive (subject to the derogations in Article 13), data subjects have the right to access to their data. This right has equally been associated with the right to private and family life under Article 8 of the ECHR as indicated in *Roche v UK*[85] and *KH v Slovakia*.[86] In these cases, the court ruled that there is a positive obligation on the hospital to make available to the patients copies of their data file.

The notion that hospitals should own medical databases has on the other hand received some legal and legislative backing.[87] In *R. v. Department of Health ex parte Source*

[81] Mags Mc Geever, 'IPR in Databases', 2006, available at: http://www.dcc.ac.uk/resources/briefing-papers/legal-watch-papers/ipr-databases#6 (accessed 11 June 2012).

[82] Roy Schoenberg and Charles Safran, 'Internet based repository of medical records that retains patient confidentiality, BMJ, 2000, vol. 321.

[83] Sharon Terry and Patrick Terry, 'Power to the People: Participant Ownership of Clinical Trial Data', *Sci. Transl. Med.* 2011, 3, 69cm3, p.1.

[84] In p-medicine this has been expressly granted to the trial participants. See the p-medicine general terms in Deliverable D.5.1, pp. 107-8.

[85] Roche v. The United Kingdom [2005] ECHR 32555/96.

[86] K.H. and others v. Slovakia [2009] 32881/04.

[87] For instance, see the 2nd sentence of the Article 16 of the act no. 277/1994 Coll. on health care (Slovak) "Medical records are property of the medical institution". The Explanatory Memorandum also

Informatics Ltd, the English appeal court rules that a patient had no proprietary claim to the prescription form or to the information it contained and had no right to control the way the information was used provided only that his privacy was not put at risk.[88] Rodwin has argued against such private ownership of patients data, insisting that it would preclude downstream invention and benefit for individual owners and the society at large; for instance, where patent is granted to a private entity for genetic discoveries.[89]

Although according ownership right of medical research databases to patients will not usually be problematic on its face value, provided that the data are managed and processed by researchers to sooth the research purposes, most opinions on this issue have tilted towards giving proprietary rights to the sponsors of the trial.[90] This is in view of the efforts and financial investment they made in the collection of the data and the trial as a whole.[91] It follows then that where clinical trials are funded with public money, the data generated from such trials should be public property. Rodwin (2009) reiterated this argument, insisting that "core values of medical professionalism – the promotion of patients' interests, medical knowledge, and public health also support public ownership."[92] Suggesting that sponsors will likely own the data in clinical trials, Drazen (2002) pointed out that ownership could be specified in the informed consent form, where patients would agree to give up ownership of data to sponsors, even when such may be used for commercial purposes.[93] He however adds that such right should not be exclusive. It should permit dissemination of data by participating investigators for non-commercial uses such as re-analysis of findings and publication in per-reviewed journals.[94] But it is contentious whether data should be in public domain when sponsored by public authorities in view of the risk of violating patient's privacy where they could be identified from such data. Although Drazen's analysis did not consider ownership rights between sponsors and investigators or participants in the clinical trial, it does suggest that such ownership rights could be spelt out contractually.[95] We shall look at how this is approached in p-medicine below.

6.2.1 Background and Foreground rights in p-medicine

In p-medicine, the consortium agreement[96] provides for the intellectual property rights in the "background" and "foreground" materials which include the data in the data warehouse and the results obtained from processing such data.[97] In the first place, it should be recalled that the project is funded under the EU Seventh Framework Programme (FP7) which makes it subject to the regulations guiding the framework.[98] In this respect, the EC Guide to

states that "Medical records remain the property of the medical institution concerned." See also *R. v. Department of Health ex parte Source Informatics Ltd* [2000] 1 All ER 786.

[88] Ibid.

[89] Marc Rodwin, The case for public ownership of patient data, *JAMA*, 2009, vol. 302, no. 1, p. 87.

[90] Ibid.

[91] Ibid, p.3.

[92] Marc Rodwin, The case for public ownership of patient data, *JAMA*, vol. 302, no. 1, p. 87.

[93] Jeffrey Drazen, 'Who owns the data in a clinical trial?', *Science and engineering ethics, 2002,* vol. 8, Issue 1, p. 409.

[94] Ibid, p. 410.

[95] See also Jill Burrington-Brown, Beth Hjort and Lydia Washington, 'Health data access, use and control', *Journal of AHIMA* 2007*, 78*, no. 5.

[96] See section 8 of the Consortium Agreement.

[97] See Article II.1 of the FP7 EC Grant Agreement Annex II.

[98] See Regulation (EC) No 1906/2006 of the European Parliament and of the Council of 18 December 2006.

Intellectual Property Rules for FP7 Projects[99] and the EC Grant Agreement Annex II[100] are relevant in the apportioning of the intellectual property rights in the project. From the definitions relating to intellectual property rights in the FP7 Regulation (EC) No. 1906/2006, and the EC Grant Agreement, "*background"* means "information which is held by participants prior to their accession to the grant agreement, as well as copyrights or other intellectual property rights pertaining to such information, the application for which has been filed before their accession to the grant agreement, and which is needed for carrying out the indirect action or for using the results of the indirect action".[101] "*Foreground*" on the other hand means the results, including information, whether or not they can be protected, which are generated by the indirect action concerned. Such results include rights related to copyright, design rights, patent rights, plant variety rights or similar forms of protection.[102] Thus, foreground includes the tangible and intangible intellectual property results of a project.[103]

An interpretation of the above provision in relation to p-medicine indicates that individual medical or other databases (including any copyright or database rights that may attach to them) belongs to individual p-medicine partners who generated such a database. This is what is regarded as the background property in the definition above. This remains so even when it has been integrated into the p-medicine data warehouse or access to such a database has been granted to other partners via the p-medicine data warehouse. However, the resultant foreground generated from the processing of this background information either in isolated units or in conjunction with other information in the data warehouse may have a different outcome. In the first place, the FP7 Regulation (EC) No. 1906/2006 provide that "foreground" arising from work carried out under indirect actions (other than those referred to in paragraph 3[104]) shall be the property of the participant carrying out the work generating that foreground.[105] The regulation goes further to state: "Where several participants have jointly carried out work generating foreground and where their respective share of the work cannot be ascertained, they shall have joint ownership of such foreground."[106]

The later provision is particularly relevant in a collaborative scenario where each partner's contribution is relevant for the outcome of the project, even though such contribution can only be indirectly felt in most aspects. For instance, a research group in the p-medicine project develops a state of the art algorithm for mining. Who owns the intellectual property of this algorithm? The p-medicine consortium or the research group who developed it? Many arguments may crop up in such instances because, even though it was a group that developed the algorithm, other groups have somehow contributed. For instance, the semantic team may argue that without their oncology the mining would not work. The clinicians on their part may argue that without their data, the group would not be able to test

[99] European Commission, Guide to Intellectual Property Rules for FP7 projects, available at: ftp://ftp.cordis.europa.eu/pub/fp7/docs/ipr_en.pdf (accessed 11 June 2012).

[100] FP7 EC Grant Agreement Annex II , available at: ftp://ftp.cordis.europa.eu/pub/fp7/docs/fp7-ga-annex2-v6_en.pdf (accessed 11 June 2012).

[101] See Article 2 of Regulation (EC) No 1906/2006 of the European Parliament and of the Council of 18 December 2006.

[102] Ibid.

[103] European Commission, Guide to Intellectual Property Rules for FP7 projects, available at: ftp://ftp.cordis.europa.eu/pub/fp7/docs/ipr_en.pdf (accessed 11 June 2012).

[104] Paragraph 3 here refers to the foreground which shall be the property of the Community such as: (a) coordination and support actions consisting in a purchase of goods or services subject to the rules on public procurement set out in the Financial Regulation; (b) coordination and support actions relating to independent experts. See Article 39 of Regulation (EC) No 1906/2006 of the European Parliament and of the Council of 18 December 2006.

[105] Ibid.

[106] Article 40 of Regulation (EC) No 1906/2006 of the European Parliament and of the Council of 18 December 2006.

the algorithm, the legal team on the other hand, may argue that without the contract framework, no group would be able to have access to data. And the argument may continue unending.

To find a common position regarding the foreground (eg, the algorithm), it is our considered opinion that joint ownership of all partners will be most favourable in deciding the intellectual property rights, irrespective of the specific group(s) which has produced it. This in our view is the intent of incorporating Article II. 26 of the EC Grant Agreement into section 8 of the Consortium Agreement. Furthermore, it may be impossible to ascertain all direct and indirect contributions of each group in the project when analysing a foreground. However, there is also a further opportunity to develop a joint ownership agreement in addition to the Consortium Agreement, if the partners so decide, in the absence of which the default rule in the Grant Agreement shall apply.[107]

At the moment, it is provided in the Consortium Agreement that in the case of joint ownership, each of the joint owners shall be entitled to use their jointly owned foreground as it sees fit, and to grant non-exclusive licences, without obtaining any consent from, paying compensation to, or otherwise accounting to any other joint owner, unless otherwise agreed between the joint owners. Issues relating to transfer, publication and access right to the foreground are also dealt with in the Consortium Agreement.[108]

However, it should be noted that the involvement of the CDP as a central data controller for the project does not accord any intellectual property right to it. Rather, the CDP serves the purposes of ensuring compliance with the data protection and security framework established for the project, including the data warehouse.

6.3 Legal protection of databases under the EU Database Directive

Protection of databases has had a chequered legal history in Europe. Historically, international attempts at protecting the intellectual property in databases could be traced to the Berne Convention, even though the database protection accorded by the Convention only extends to a collection of works and not of mere data.[109] The TRIPS[110] and WIPO Copyright Treaty[111] however extended this provision by incorporating 'compilation of data' in such databases protection. However, one short coming of the above instruments is that they are applicable to international situations among signatory states. On the home front within the EU, databases are seldom protected by copyright in the strict sense, as they rarely proved to be "original" in their "arrangement" or "selection".[112] Prior to the Database Directive, copyright protection for databases in the Member States could be seen to have been divided into two general groups, depending on the threshold for protection. On the one hand, the Anglo-Irish common law systems applied a "sweat of the brow" doctrine for a database to

[107] See Article II. 26 of the EC Grant Agreement. The default rules states that where no joint ownership agreement has yet been concluded, each of the joint owners shall be entitled to grant non-exclusive licences to third parties, without any right to sub-licence, subject to the following conditions:

a) at least 45 days prior notice must be given to the other joint owner(s); and

b) fair and reasonable compensation must be provided to the other joint owner(s).

[108] See sections 8 and 9 of the Consortium Agreement.

[109] See Article 2 (5) of the Berne Convention for the Protection of Literary and Artistic Works of September 1886 (as amended in 1948 by the Act of Brussels).

[110] See Article 10(2) of the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPs), 1994.

[111] Article 5 of the WIPO Copy Right Treaty 1996.

[112] European Commission, The Implementation and Application of Directive 96/9/EC on the Legal Protection of Databases, Study – Contract ETD/2001/B5-3001/E/72, Plan, 2001, p. 7.

qualify for protection which is of a lower threshold. On the other hand, in the continental legal systems, copyright imposed a fairly high threshold of originality for any database to qualify for protection. This is in keeping with the "author's right" approach that prevails throughout most of Continental Europe, where originality is defined as an expression of the author's individual personality.[113] In series of cases, copyright protection was denied to databases, thus leaving them to contractual protection just between the parties to the contract, or under unfair competition.[114]

This tide however changed in 1996 when the EU passed the Database Directive. Although the Directive adopted a higher standard of originality than that required under the common law system, by requiring an "intellectual creation", it however recognised that there is a need to protect the financial investment of makers of databases, in view of the fact that a large market is being built around such intellectual creation, even though they may lack originality. To that extent, database right subsists in a database if there has been a substantial investment in obtaining, verifying or presenting the contents of the database (even if the contents and/or structure of the database are not original and therefore do not attract copyright). This is known as the *sui generis* right. Investment is construed widely here, and covers financial, human and technical resources.[115]

The Database Directive provides that the first owner of the database right is the 'maker' of a database. The maker is the person who takes the initiative in obtaining, verifying or presenting the contents of a database and assumes the risk of investing in the same. The database right enables the owner to prevent others from extracting and/or re-utilising all or a substantial part of the contents of their database, even though what constitutes a 'substantial part' is still unclear. Database right subsists for 15 years from the creation of the database, but if the database is published within this time, then the term is 15 years from publication. Although this term is much shorter than its copyright equivalent there is some uncertainty surrounding it. This is based on the fact that there is a renewal of the term of the right each time there is a "substantial change" to the contents of the database. This means that if a database is continually changed and updated, the right could last indefinitely.[116]

In general, an intellectual property right that protects the *expression* of ideas or information in a database can generally be seen in two ways, depending on the form of the expression: either by copyright or *sui generis* right or even both. There is often confusion around the subsistence of copyright in a database, but nevertheless, a database may attract copyright protection in certain limited circumstances, depending on the *structure* of the database and the *content.* In relation to the structure, a database may be protected by copyright if, by reason of the selection or arrangement of the contents, it constitutes the author's own intellectual creation.[117] Secondly, depending on the content of the database, copyright may also exist independently in the contents of the database (for example, a database of images where each of the images would attract its own copyright protection as an artistic work).[118]

[113] Mark Davison, *The Legal Protection of Databases*, Cambridge University Press, United Kingdom, 2003, pp. 10-24.

[114] European Commission, The Implementation and Application of Directive 96/9/EC on the Legal Protection of Databases, op. cit.

[115] Mags McGeever, op.cit.

[116] See Article 10(3) of the Directive 96/9/EC (Database Directive).

[117] See Article 3 (1) of the Directive 96/9/EC (Database Directive).

[118] Ibid. See also Mags McGeever, op.cit.

On the other hand, a database right can be claimed as a *sui generis* form of intellectual property developed exclusively to protect databases.[119] This scheme provides a protection for non-original databases where there has been a substantial investment in their creation.[120] Copyright or database right of an intellectual creation arises automatically once the creation assumes a material form without any procedure for registration. The Directive permits member states to adopt exceptions from the *sui generis* right for lawful users in three specific categories: (a) extraction for private purposes of the contents of a non-electronic database; (b) "extraction for the purposes of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved"; and (c) "extraction and/or re- utilization for the purposes of public security or an administrative or judicial procedure."

6.3.1 The Database Directive and the p-medicine data warehouse

To trigger the protection of the Database Directive, a compilation of data must fit within the notion of database as defined in the Directive.[121] In this regard, a pertinent question in relation to the p-medicine database will be whether it qualifies for database protection under the Directive. A look at the definition of the term "database" under the Directive gives an insight on how to answer the above raised question. The Directive defines a database as a "collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means."[122] As could be seen from the above definition, the scope of what amounts to a database is very broad and includes literary, artistic, musical or other collections of works or materials such as texts, sounds, images, numbers, facts, and data. Once these are systematically or methodically arranged, and can be individually accessed by electronic or other means, they qualify to be protected under the Directive. Based on this broad definitional approach of the Directive, the p-medicine data warehouse is *prima facia* covered as a database, and qualifies for legal protection in its own right.

Firstly, the nature of data contained in the p-medicine data warehouse could be located within the description of Recital 17 of the Directive as a combination of facts, images, figures and data.[123] These data are separable from one another without their contents being affected and therefore can be regarded to be independent materials because they are pooled from various trial centres. They are also systematically or methodically arranged and individually accessible by electronic means as required in Art. 1 (2) of the Database Directive.[124] Secondly, substantial investment has been made in collecting and integrating the data in the database from multiple partners involved in the project. Although what amounts to a substantial investment was not explicitly defined in the Directive and will be up to the court to decide on a case by case basis, however, Recitals 7 and 40 of the Directive provide further guidance regarding the sort of investment that can be regarded as substantial. Financial, technical and human resources are factors to be considered in this regard, and it will not be hard in our opinion to find that the project has made a quantitative and qualitative investment in building the data warehouse. The ECJ has however, ruled that the investment protected under the Directive must be directed as obtaining (that is, collecting the data) and not to the

[119] European Commission, The Implementation and Application of Directive 96/9/EC on the Legal Protection of Databases, op. cit.

[120] See Article 7 of the Directive 96/9/EC (Database Directive).

[121] Jasper Bovenberg, *Property Rights in Blood, Genes and Data Naturally Your?* 2008, Martinus Nijhoff Publishers, 2008, Netherlands, pp. 76-80.

[122] See Article 1 of the Directive 96/9/EC (Database Directive).

[123] See section 5 above.

[124] See OPTIMIS Cloud Legal Guidelines Deliverable D7.2.1.1.

"creation" of data.[125] Thus, in the light of the ECJ decision, the *sui generis* right under the Directive can be applied to protect the p-medicine data warehouse if it can be shown that the investment in establishing the data warehouse is directed at obtaining its content. This in our opinion is the case in p-medicine.

The rights accorded to databases under the Directive belong to the maker of the database, which in this case is the p-medicine consortium. These include the right to prevent extraction or re-utilisation of the whole or a substantial part of the contents of the database.[126] Finally, the period of protection of databases is 15 years after the completion of the database and is rolled over following any substantial modification to the database that requires a substantial investment.[127] As indicated earlier, the database protection accruable to the p-medicine database under the Directive does not affect any of the other intellectual property rights including patent and copyright subsisting on the elements making up the content of the database.

[125] *Fixtures Marketing Ltd v. AB Svenska Spel,* op. cit.

[126] See Article 7(1) of the Directive 96/9/EC (Database Directive).

[127] Jasper Bovenberg, op. cit.

7 Guidelines for the establishment and mining of the p-medicine data warehouse

7.1 Background

The importance of data warehousing and data mining support tools in the creation of evidence based and personalised medicine cannot be ignored. While these platforms provide tools for knowledge discovery and pattern recognition, there are legal and ethical concerns that they may be used to breach individual privacy, most especially of the patients. Therefore, we have developed a set of rules that will guide the establishment and use of the p-medicine warehouse to forestall such occurrence. All persons having access to the p-medicine database shall be guided by these guidelines for its use as set out in Annex 1 of this report.

These guidelines take into account the legal, ethical and societal concerns in managing a translational research data warehouse, as well as the need of balancing these concerns in order to achieve the project objectives. These guidelines complement the contractual obligations of participating partners in the project, and should be observed in conjunction with those obligations. This in effect means that those who will have access to the data warehouse are persons who had completed the Annex C of the Contract on Data Protection and Data Security within p-medicine. In addition to the defacto anonymisation mechanism used in the database, privacy enhancing data mining techniques and strict access policy are also put in place. There will also be an annual review of these guidelines as experience is gained and lessons learned.

The guidelines will be published in the p-medicine access portal and end users are to read and adhere to them. These guidelines are meant to give a clear and documented instruction to all authorised personnel on how to properly use the p-medicine data warehouse systems, and how to avoid security risks and breaches. To enable a better understanding of the intent of the guidelines, a short background summary of their content will be given below.

First of all, it is aimed that the data must only be used for the objectives of the p-medicine project. This is in accordance with Article 6 (b) of the Data Protection Directive which states that personal data must be collected for specified and legitimate reasons, and must only be processed in a way compatible with those defined purposes. Although we deal with de-facto anonymous data here, we still have to ensure a high level of data protection aimed at enhancing the trust of the trial participants regarding the safety of their data. In addition, Article 2 of the Good Clinical Practice Directive also sets out that the safety and rights of the trial subject shall prevail over the interests of science and society, which indicates that the protection of the participants has paramount priority. All these aim at ensuring confidentiality and integrity of the trial.

Article 3 of the Clinical Trial Directive also requires that the privacy and the data relating to the participants have to be safeguarded. It also has to be kept in mind that re-identification is not absolutely excluded in this project. Thus to ensure the patients' privacy, a laid down procedure for any re-identification has to be strictly followed in achieving the p-medicine purposes when using the data from the data warehouse.

To this end, only those entities that have consented to the p-medicine framework will be allowed to access the data. This must not be bypassed through the transfer of participants' data to third parties by a lawful end user. Exemptions have to be allowed by the CDP. To

control this provision, a comprehensive logging of access and documentation of all processing activities is needed and will be kept.

The protection of the participants' privacy also requires high technical and organisational security standards, and prohibits the publishing of the data in a way that enables others to re-identify the concerned subjects. It is also absolutely vital that the consortium partners who have access to the participants' data will not try to re-identify them. To secure the integrity of the data, as well as the quality of the trial as envisaged under Art. 2 of the Directive 2001/20/EC, manipulation of data by the end users is forbidden. The integrity of data is of high value and any alteration must be traceable and accounted for.

Annex 1: Guidelines on the Use of the P-medicine Data Warehouse

Background

The establishment of a translational medical research database as envisaged in this p-medicine data warehouse is to integrate large amount of data needed for knowledge discovery system. When mined appropriately, predictive results that would assist researchers and care givers in achieving personalised medicine can be produced form such a database. At the same time, ensuring the protection and security of the trial participants' data will lead to maximum benefit from these resources. In order to ensure a high level of compatibility and flexibility between the p-medicine research objectives and the respect for the participants' privacy, the following guidelines have been developed to further serve as good practice in the data warehousing and mining. They also give a fine grained implementation structure of the data protection and security framework of the project, and will operate together with the p-medicine end user agreement concluded by each partner who will be accessing this data warehouse.

Purpose

The purpose of the establishment and use of the p-medicine data warehouse shall be for conducting medical research and for the development of information technology tools to facilitate personalised medicine. The use of data for other purposes is not allowed.

Access

Access to the p-medicine data warehouse will only be given to the persons who have completed the Annex C to the Contract on Data Protection and Data Security within p-medicine, and whose institution is a partner of the p-medicine project. Authorised users shall not give access to third parties. CDP is responsible for access to the p-medicine data warehouse.

Non-transferability of data to third parties

Disclosure or transfer of data from the p-medicine data warehouse to a third party, that is, any party who has not completed the Contract on Data Protection and Data Security within p-medicine, as well as the Annex C to the contract, is not allowed. Any such disclosure or transfer must be expressly authorised by the CDP.

Downloads

Any person having access to the data warehouse shall ensure that when p-medicine data are stored by him/her, they are technically and organisationally separated from other data.

Access Control and Logs

There shall be a comprehensive logging of accesses and documentation of all processing steps which have taken place within the p-medicine data warehouse system. Authentication procedures shall be put in place and all data manipulation shall be logged for later analysis where necessary.

Publication of Data

Data used for this research shall not be published in a form that enables the data subjects to be identified. The users (researchers) are not allowed to publish the data or to transmit or disclose data received via the p-medicine database to any third person outside of the consortium. This does not affect the publication of articles in journals, so far as the identity of any trial participant is not mentioned and cannot be revealed from the publication.

Non-identification

The end users, that is, researcher using data received from the p-medicine data warehouse, shall not attempt to re-identify the patients in order to address them directly. Such re-identification is not allowed. Re-identification within p-medicine shall only be in accordance with the data protection framework as established in Deliverable 5.1. Any need for re-identification shall only be addressed to the CDP in accordance with p-medicine privacy policy, so that the patient can be identified by the CDP with the help of the Trusted Third Party that holds the key to the link of the pseudonymised data set of the patient concerned.

Alteration of Data

End users are not authorised to alter any data obtained from the p-medicine data warehouse.

Data Security

Data obtained from the p-medicine data warehouse are to be stored in encrypted form. Appropriate organisational and technical measures as indicated in the end user agreement shall be fully implemented by end users to ensure the confidentiality, integrity and accuracy of processed data. The end user shall prevent unauthorised access to the database and shall maintain an audit trail as well as the following controls:

- control of the entrance to installations
- control of data media
- memory control
- control of data transmission
- access control
- control of communication
- control of data introduction
- separation of data
- availability control

Data Breach Notification

End user shall report any breach of confidentiality of data obtained from the p-medicine data warehouse to the CDP as soon as the breach is noticed.

End Users' Responsibility

The end user of the p-medicne data warehouse is responsible for the security and use of research data in accordance with the p-medicine data protection and security framework.

Backups

Effective backup and recovery mechanisms shall be put in place in order to secure the content of the database.

Intellectual Property

The database right in the data warehouse as a compilation belongs to the p-medicine consortium.

Audit

The CDP reserves the right to inspect the processing facilities, audit data files and documentations of the end users to ascertain compliance with these guidelines.

Data retention

Upon on the completion of the p-medicine project, the end user shall return all the data and the copies thereof obtained from the data warehouse to the CDP or shall destroy all the data unless legislation prevents such an end user from doing so. In that case, the end user shall continue to maintain the confidentiality of the data and shall not further process the data stored for any other purpose.

Violations

The CDP reserves the right to restrict the access of any end user who violates these guidelines.

References:

A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, 'l-Diversity: Privacy Beyond k-Anonymity', *International Conference on Data Engineering*, 2006

A. Friedman, A. Schuster and R. Wolff: 'k-Anonymous Decision Tree Induction', European Conference on Principles and Practice of Knowledge Discovery in Databases, 2006.

Amit Sheth and James Larson, 'Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases'*, ACM Computing Surveys*, 1990, Vol. 22, No. 3.

Article 29 Working Party, Working Document on the processing of personal data relating to health in electronic health records (EHR), WP 131.

Act no. 277/1994 Coll. on Health Care (Slovak).

Berne Convention for the Protection of Literary and Artistic Works of September 1886 (as amended in 1948 by the Act of Brussels).

Council of Europe, Committee of Ministers, Recommendation No. R (97) 5 on the Protection of Medical Data (Feb. 13, 1997).

Chris Hinds et. Al, 'Ownership of Intellectual Property Rights in Medical Data in Collaborative Computing Environment'.

Christopher Kuner, *European Data Protection Law - Corporate Compliance and Regulation*, 2007.

Communication from the Commission to the European Parliament and the Council on Promoting Data Protection by Privacy Enhancing Technologies (PETs), 2007.

Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data.

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive).

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (Data Protection Directive).

Directive 2001/20/EC of the European Parliament and of the Council on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use (Clinical Trials Directive).

Directive 2005/28/EC laying down principles and detailed guidelines for good clinical practice as regards investigational medicinal products for human use, as well as the requirements for authorisation of the manufacturing or importation of such products" (Good Clinical Practice Directive).

David Karp et al, 'Ethical and practical issues associated with aggregating databases', PLoS Medicine, 2008, vol. 5, Issue 9.

EC, Commission decisions on the adequacy of the protection of personal data in third countries, http://ec.europa.eu/justice/policies/privacy/thridcountries/index_en.htm.

ENISA, Cloud computing: benefits, risks and recommendations for information security, 2009.

European Commission, Guide to Intellectual Property Rules for FP7 projects, available at: ftp://ftp.cordis.europa.eu/pub/fp7/docs/ipr_en.pdf.

European Commission, The Implementation and Application of Directive 96/9/EC on the Legal Protection of Databases, Study – Contract ETD/2001/B5-3001/E/72, Plan, 2001.

European Commission, The Implementation and Application of Directive 96/9/EC on the Legal Protection of Databases.

France Telecom vs. MA Editions (Tribunal de commerce de Paris, 18 June 1999).

Fixtures Marketing Ltd v OY Veikkaus Ab (2004) EUECJ C C-46/02.

Fixtures Marketing Ltd v Svenska Spel AB, (2004) EUECJ C-338/02.

FP7 Grant Agreement - Annex II General Conditions.

FP7 EC Grant Agreement Annex II , available at: ftp://ftp.cordis.europa.eu/pub/fp7/docs/fp7-ga-annex2-v6_en.pdf.

Heather Piwowar *et al*,, 'Towards a data sharing culture: recommendations for leadership from academic health centers', PLoS Medicine, 2008, vol. 5, Issue 9.

ICH Topic E 6 (R1) Guideline for Good Clinical Practice, CPMP/ICH/135/95, July 2002.

Jane Kaye et al, 'Data Sharing in Genomics – Re-shaping Scientific Practice', Nat Rev Genet, 2009, 10(5).

Jonathan Prather, 'Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse', *Proc AMIA Annu Fall Symp.*, 1997.

Jeffrey Drazen, 'Who owns the data in a clinical trial?', *Science and engineering ethics, 2002,* vol. 8, Issue 1.

Jill Burrington-Brown, Beth Hjort and Lydia Washington, 'Health data access, use and control', *Journal of AHIMA* 2007*, 78*, no. 5.

Jasper Bovenberg, *Property Rights in Blood, Genes and Data Naturally Your?* 2008, Martinus Nijhoff Publishers, 2008, Netherlands.

KPN v. Denda (Gerechtshof Arnhem, 15 April 1997, follow-up in Rechtbank Almelo, 6 December 2000).

K.H. and others v. Slovakia [2009] 32881/04.

Latanya Sweeney: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002.

M. Atzori, F. Bonchi, F. Giannotti and D. Pedreschi: Anonymity preserving pattern discovery, The VLDB Journal, 2008.

Michele Oliva and Marcelo Corrales, Law Meets Biology: Are Our Databases Eligible for Legal Protection? 2011, *Scripted*, Vol. 8, Issues 3.

Mark Davison, *The Legal Protection of Databases*, Cambridge University Press, United Kingdom, 2003, pp. 10-24.

Mags McGeever, 'Sharing Medical Data', 2006, available at: http://www.dcc.ac.uk/resources/briefing-papers/legal-watch-papers/sharing-medical-data#5.

Mags Mc Geever, 'IPR in Databases', 2006, available at: http://www.dcc.ac.uk/resources/briefing-papers/legal-watch-papers/ipr-databases#6.

Marc Rodwin, The case for public ownership of patient data, *JAMA*, 2009, vol. 302, no. 1.

NVM vs. De Telegraaf (21 December 2000) from the Dutch Court of Appeal in The Hague.

Nevena Stolba, Marko Banek and A Min Tjoa 2006, 'The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine'.

Nevena Stolba and A Min Tjoa, ' Relevance of Data Warehousing and Data Mining in the Field of Evidence-based Medicine to Support Health Decision Making' *World Academy of Science, Engineering and Technology* 2005, p. 192.

OPTIMIS Cloud Legal Guidelines Deliverable D7.2.1.1.

P. Samarati and L. Sweeney, 'Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression', Technical report, 1998.

P. Samarati, 'Protecting Respondents' Identities in Microdata Release', IEEE Transactions on Knowledge and Data Engineering, 2001.

P-medicine Deliverable D8.1.3, Report on federated storage services.

P-medicine Deliverable D8.1.1, Specification of the interaction with the VPH toolkit.

P-medicine Deliverable 5.5 for the report on legal and ethical issues for p-medicine tools used for international GCP trials.

P-medicine Deliverable D3.4 on service integration guidelines

P-medicine Deliverable D7.1: Report on overall system design including VPH-Share D2.2 and indicating its impact

P-medicine Deliverable No. D5.1 - Setting up of the data protection and data security framework for p-medicine.

Roy Schoenberg and Charles Safran, 'Internet based repository of medical records that retains patient confidentiality, BMJ, 2000, vol. 321.

Roche v. The United Kingdom [2005] ECHR 32555/96.

R. v. Department of Health ex parte Source Informatics Ltd [2000] 1 All ER 786.

Regulation (EC) No 1906/2006 of the European Parliament and of the Council of 18 December 2006.

Sebastien Lapointe, 'Legal Cloud Computing: Concepts and Ramifications', Legal IT 2010.

Sharon Terry and Patrick Terry, 'Power to the People: Participant Ownership of Clinical Trial Data', *Sci. Transl. Med.* 2011, 3, 69cm3.

The British Horseracing Board Ltd and Others v William Hill Organization Ltd, (2004) EUECJ C-203/02.

Tele-Info-CD (Bundesgerichtshof, 6 May 1999).

The p-medicine Description of Work.

The p-medicine Consortium Agreement.

The Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPs), 1994.

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

WMA Declaration on Ethical Considerations regarding Health Databases, 53rd WMA General Assembly, Washington, DC, October 2002.

WIPO Copy Right Treaty 1996.