BioMedBridges

Deliverable 8.1

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	Process specification for secure sharing of and access to PM data	
WP No.	8	
Lead Beneficiary:	7: Klinikum rechts der Isar der Technischen Universitaet Muenchen (TUM-MED)	
WP Title	Use case: Personalized Medicine	
Contractual delivery date:	30 June 2014	
Actual delivery date:	30 June 2014	
WP leader:	16: Imre Västrik	16: UH
Contributing partner(s):	1: EMBL, 3: KI, 5: UDUS, 7: TUM-MED, 16: UH	

Authors: Florian Kohlmayer, Raffael Bild, Imre Västrik, Klaus Kuhn, Benedicto Rodriguez, Sabine Brunner, Ashish Lamichhane, Christian Ohmann

Contents

1 Executive summary

The aim of this deliverable is to present a process specification for secure sharing of and access to personalized medicine (PM) data. The intention is that a producer of data can share and the user of the data can gain access to personalized medicine (PM) data in a secure and legal, yet easiest possible manner.

For the specification described in this deliverable, close cooperation with the Secure Access Work Package (WP) 5 has been of high relevance. Previous work in WP5 started with the specification of a usage scenario for PM, and the identification of regulations, privacy and security requirements, which were presented by deliverable D5.1 [1]. Deliverable D5.2 further elaborated the work of D5.1 and published templates of relevant forms under http://www.biomedbridges.eu/deliverables/52-0. Next, a security architecture and framework has been developed in WP5 and described in deliverable D5.3. Secure access to and sharing of PM data is one of the most relevant use cases for this architecture. Deliverable D8.1 on its part will massively build upon D5.3.

As a follow-up, a proof of concept is planned, which will be covered by a forthcoming deliverable, D8.3. Cooperation with the Technical Integration Work Package 4 will be sought for this step.

Deliverable D8.1 relies on the security and privacy architecture which has been developed and put forward in deliverable D5.3 of the Secure Access Work Package 5. This architecture has been developed to support the security and privacy requirements of all the Use Case (UC) WPs, i.e., WP6-10, including WP8 the use case of personalized medicine.

Deliverable D8.1 revisits the generic security and privacy architecture presented in D5.3 to address the data management challenges of the BioMedBridges (BMB) project as a whole. It builds upon Usage Scenarios described in D5.1 and on the Data Flow Diagrams (DFDs) described in D5.3. Alltogether, D8.1 can be perceived as a particular “instantiation” of the general security architecture of BMB, with a specific focus on PM.

Deliverable D8.1 is structured as follows:

Section 3 provides an overview of the background of personalized medicine. Section 4 describes the methodology applied, which essentially follows the approach described in D5.3. Section 5 elaborates on the process specification conducted as a basis of a threat and risk analysis that is described in Section 6. Section 7 then explains the design of the security framework derived from the threat and risk analysis results. Section 8 puts forward processes for secure sharing of and access to personalized medicine data based on work carried out in WP5.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Develop a process for secure sharing of and access to PM data	x	
2	Define types of PM data and mapping between them		x
3	Develop a PM informatics pipeline		x

3 Background

“Personalized Medicine” refers to the tailoring of medical treatment to the individual characteristics of each patient to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment. As yet the molecular profiling technologies involved in characterization of the patient (or his/her disease) have not quite made it to the healthcare institutions but are available in research institutes. Furthermore, classifying patients into subgroups and tailoring of the treatment is still very much a topic of research. Hence, personalised medicine based on genomic data can (at present) only be applied at the interface of healthcare and research – two areas with different approaches and requirements regarding privacy protection and data sharing. Alas, personalised medicine

also involves many different categories of people using and/or producing the data – physicians, lab personnel, researchers contracted to interpret the data for the benefit of the physician and patient, and “independent” researchers. Much of the effort in the task has been dedicated to understanding the needs of different groups of personnel in different roles as well as related security requirements.

This deliverable describes the process specification for secure sharing of and access to PM data. It is based on the work performed in WP5 (Secure access) and represents an instantiation of the security framework described in D5.3: “Report describing the security architecture and framework”. In addition, work performed in WP4 (Technical Integration) was and even more will be highly influential.

4 Methods

The definition of the process for secure sharing of and the access to personalized medicine data has to start with determining the requirements of the use case, followed by a threat analysis. The discovered threats have to be seen in the context of their likelihood and their impact, leading to an identification of risks. Most underlying work has been performed in deliverable D5.3, including the STRIDE [2] and LINDDUN [3] analyses with their corresponding Data Flow Diagrams (DFDs) [4]. The PM use case has been explicitly addressed by WP5 and D5.3.

Highly influential to deliverable D5.1, D5.3, and, in the end, to D8.1 were two surveys performed by WP5. The first survey focussed on the Research Infrastructures (see Section 12 Annex I of [1]), whereas the second survey was focused towards the use cases (see Section 6 of deliverable D5.3). In these surveys the requirements were collected. For modelling the use case with the focus on the threat analysis, DFDs were used as suggested by STRIDE and LINDDUN. The model was used to identify security and privacy threats. This was followed by a risk analysis which was based on [5]. Results from the literature were included in this step, for example [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17]. Using the results from the previous steps, the most

important countermeasures were selected and a generic security framework was developed.

Based on this generic framework a more concrete instantiation for the specific use case of personalized medicine was developed. It contains an overview of the processes modelled.

The Usage Scenario Personalised Medicine of D5.1 (Section 10.2.5.) was considered for subsequent work. The DFD presented in D5.3 is shown here as **Figure** 2. Figure 1 presents a further development of this DFD, while Figures 5-10 show activity diagrams developed for D8.1. Figures 3 and 4 are taken out of an interim report of WP8.

Definitions of terms used in this document and more details on the methodology can be found in deliverable D5.3.

5 Process specification: sharing of and access to personalized medicine data

The WP8 use case focusses on the integration of complex Personalized Medicine (PM) data sets, aiming at a better understanding of disease pathogenesis, at improving biomarkers, and at selecting optimal treatment. The PM data bridges (where the term “data bridge” is used with the same meaning as in deliverable D5.3 throughout this document) will provide access to integrated, often heterogeneous and distributed patient-related data. The goal is to enable better treatment decisions for individual patients. As the data is patient-related, both data security and data protection are of high relevance. In each case, it has to be clear whether data have been collected for health care or for research purposes. We will address this in Section 5.4.

As described by the title of the deliverable and by the corresponding Work Task (WT) 1 of WP8, the focus of this deliverable is on secure sharing of and access to PM data.

5.1 Activity diagrams

The work put forward builds on a process specification covering the needs of users and consumers of data bridges in the specific context of PM, and combines it with work carried out in the secure access work package 5. We start with the process specification which has been developed by WP8 in cooperation with WP5. The process specification started with the usage scenarios already presented in D5.1. At this point, we refer to the activity diagram on page 75 in Section 9.3.8 of D5.1.

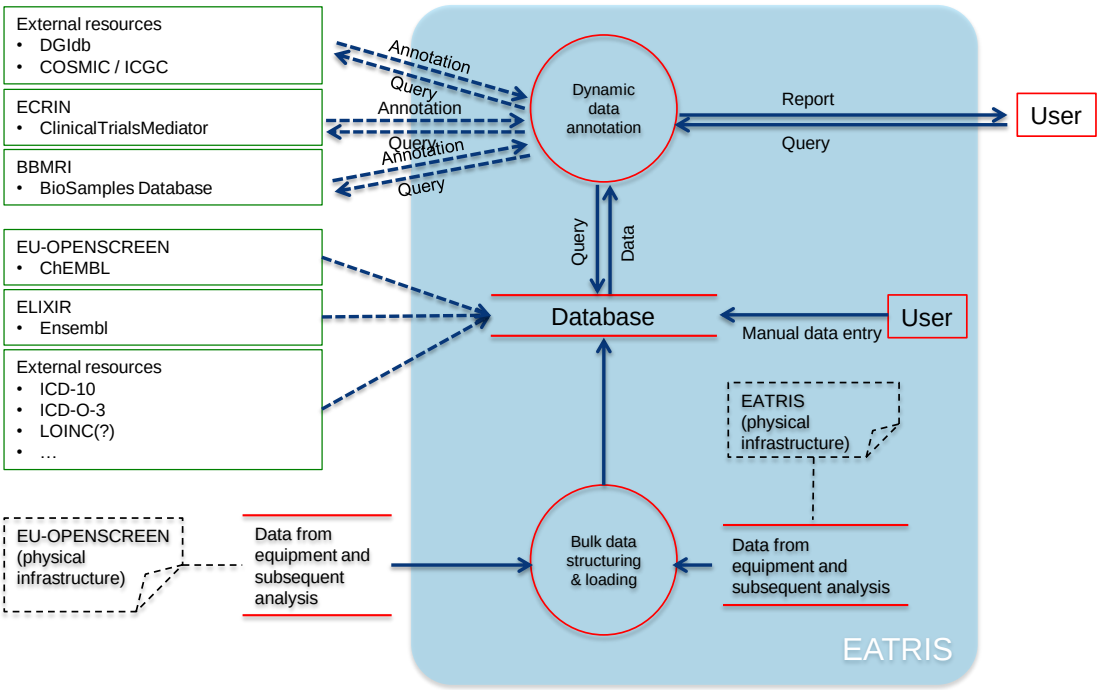
5.2 Data flow diagram

WP8 will need bridges between the research infrastructures ECRIN, BBMRI, EU-OPENSCREEN, ELIXIR and EATRIS. The open, unrestricted databases of these infrastructures are Cosmic[1], ICGC[2] (in parts), DGIdb[3], ClinicalTrialsMediator (CTIM)[4], BBMRI Biosample Database[5], ChEMBL[6], and Ensembl[7]. In addition, several terminologies like ICD-10, ICD-O-3 and LOINC will be integrated. Based on the surveys, no dedicated security and privacy measures are necessary for this integration. Nevertheless, the terms and conditions for accessing this data have to be respected. Finally, restricted data from EU-OPENSCREEN should also be integrated in the PM use case.

As WP8 integrates data from various sources, both open and restricted, and enriches this data with patient-related data from EATRIS and EU-OPENSCREEN, access to this data has to be well guarded. Figure 1 shows a DFD which is a revised version of the DFD that was originally populated with input from the second survey and additionally incorporates feedback on interim progress in the context of D5.3.

[1] http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/
[2] https://icgc.org/
[3] http://dgidb.genome.wustl.edu/
[4] http://www.ecrin.org/
[5] http://bbmri.eu/
[6] https://www.ebi.ac.uk/chembl/
[7] http://www.ensembl.org/

Figure 1 Data Flow Diagram of WP8: Personalised Medicine



5.3 Data Bridges

The following data bridges have been identified for WP8:

Bridge 1: ECRIN → EATRIS: Open data from the ClinicalTrialsMediator is used during the dynamic data annotation process.

Bridge 2: BBMRI → EATRIS: Open data from BBMRI catalogues is used during the dynamic data annotation process.

Bridge 3: EU-OPENSCREEN → EATRIS: Open data from the ChEMBL Database is used during the dynamic data annotation process.

Bridge 4: EU-OPENSCREEN → EATRIS: Restricted data from equipment and subsequent analysis. If restricted data are used, the use of templates as provided by D5.2 and of additional security measures (e.g. restricted access also on the side of the "consumer") will apply.

Bridge 5: ELIXIR → EATRIS: Open data from the Ensembl Database is used during the dynamic data annotation process.

Bridge 6: EATRIS → researchers: Three access layers are needed for secure data access and a spectrum of security measures is needed for data release.

5.4 Data provenance

FIMM as a prototype personalised medicine ***data producer*** operates a molecular profiling service which is used primarily by researchers but increasingly also by well-informed physicians. Details about the existing PM data types and mappings between them can be found in the report on deliverable D8.2. In D8.1, the focus is on security aspects. The "raw" data produced are typically exome sequences from tumour and normal tissues, RNA sequences from tumour tissues and ex vivo drug sensitivity measurements of tumour samples. These are further processed into germline genotype calls, somatic mutation calls, gene copy number estimates, gene expression levels and differential (compared to the sensitivity of a normal tissue) drug sensitivity scores. As such, the data can originate from several different scenarios:

Research projects affiliated with FIMM. These projects have permissions from ethics boards. Patients have consented to the use of their samples and data for specific research projects. Most of the leukaemia data produced at FIMM belongs to this category and has been produced in the "Molecular pathogenesis, risk factors and individualised treatment of hematologic malignancies" project. We note here that further use of these data by external researchers has to comply with any legal and ethics approval requirments and be covered by appropriate informed consent. Moreover, security measures like anonymity (typically) and pseudonymity (if required and justified) are needed.

Biobanking. Biobanks have permission from the responsible ethics boards. They are based on informed consent, and they typically seek for broad consent allowing utilization of data for multiple purposes as long as these are congruent with the aims (as stated in the statutes of the biobank) of the biobank (and hence with the consent of the donors). For example, the aim of the Finnish Hematology Registry and Biobank (FHRB) is to promote the

registration, treatment and diagnostics of haematological diseases. It is the FHRB scientific board which decides on the use of the biobanked samples in research. It is not planned here that biosamples leave FIMM.

Physicians wanting in-depth analysis of their patients. In this case, molecular profiling is similar to any other (diagnostic) procedure performed as part of patient care. This data cannot be shared with or shown to anyone else except the physician commissioning their production and FIMM personnel involved in their analysis and reporting. FIMM specialists are working as consultants co-treating patients. The security architecture may be helpful, but health care is not the focus of this deliverable. In order to enrich research data bases, data may be imported from external sources in strict compliance with use agreements, ethics committee approval and informed consent. If data is to be transferred from the health care context to a research environment, the complete spectrum of measures is needed which includes informed consent, approval of ethics committees, as well as security measures (restricted access, anonymity).

We also foresee **patients donating their samples directly** with the help of their physician. These samples are always to be accompanied with broad consent. In those cases it would be especially desirable if we had a means of sharing the data with the patient as well as with anyone the patient wants it to be shared with, i.e. a physician or a researcher. In the context of BioMedBridges, secure access (see access layers) and secure release (typically: anonymity) are an option.

6 Threat and risk analysis of the processes

Based on the information collected, a threat and risk analysis based on STRIDE and LINDDUN were performed. The results and appropriate countermeasures are documented in Table 1.

Table 1. Detailed threats, risks and countermeasures (LoT: Likelihood of Threat, LoI: Level of Impact, L: Low, M: Medium, H: High)

A row in light blue background color indicates that the threat event is addressed by **STRIDE** or **LINDDUN** and applies only to the **"data flow"** element type of the DFD under evaluation.
A row in light red background color indicates that the threat is addressed by **STRIDE** only and applies to the one of the rest element types of the DFD under evaluation (i.e. **"data store", "process", "external entity"**).
A row in light green background color indicates that the threat is addressed by **LINDDUN** only and applies to the one of the rest element types of the DFD under evaluation (i.e. **"data store", "process", "external entity"**).

Threat Event	**Threat Sources**	**Vulnerabilities and Predisposing Conditions**	**LoT**	**LoI**	**Risk**	**Counter-measures (Elements of security architecture)**
Spoofing as user to get access to the PM database	External	Weak authentication	L	H	M	Authentication system
Spoofing: Pharming 'PM Analysis tool'	Processing	Weak Authentication system/ Weak configuration management	L	H	M	Authentication system/ Configuration management
Tampering: Modify/ delete data in the PM database	Internal	Weak access control of database (accidental)	M	M	M	Authorization
Tampering: Modify data flow from external entity	External	Insecure data transfer	L	L	L	Secure data communication
Tampering of analysis tool (input validation failure)	Processing	Missing Input validation	M	H	H	Input validation practices/ Configuration management
Tampering: Overcapacity failure of PM database	Storage	Missing handling of overcapacity failures	L	M	L	Configuration Management

Repudiation: Changes in PM database cannot be traced	Processing	No/weak audit trail	L	M	L	Auditing and logging
Repudiation: PM analysis tool activities (e.g. connection to databases) cannot be traced	Processing	No/weak audit trail	L	L	L	Auditing and logging
Repudiation: Version of external data sources used in the analysis not logged	Processing	Weak logging/ Missing audit trail	M	M	M	Auditing and logging
Information disclosure of patient data in PM application	Internal	PM application is not secure, weak access control	H	H	H	Authorization
Information disclosure of query / query results (data flow, process)	External	Insecure data transfer	M	M	M	Secure data communication
Information disclosure of annotated patient data (data flow, process, data store)	External	Insecure data transfer / insufficient access control of data store	M	H	H	Secure data communication/ Authorization
Denial of Service of PM database (input validation, lack of resources)	Processing, Storage	Missing input validation/ insufficient resources handling	L	L	L	Input validation practices/ Configuration management
Denial of Service of PM analysis tool (input validation, lack of resources)	Processing, Storage	Missing input validation/ insufficient resources handling	L	M	L	Input validation practices/ Configuration management
Elevation of Privilege: A user has access to patient data that s/he is not allowed to.	Processing, Storage	Insufficient access control	M	H	H	Authorization
Elevation of Privilege: Unauthorized users can edit/delete patient data.	Processing, Storage	Insufficient access control	H	H	H	Authorization
Linkability of query results for particular patient	External	Query results are in some way connected	L	H	M	Secure data communication or Encryption query result

Linkability of entry in “Pseudonymized /anonymized datastore” to patient	External	Insufficient anonymization/ Pseudo-nymization	L	M	L	Anonymization/ Pseudo-nymization
Identifiability of patient with the help of queries/ query results	External	Queries/ results not protected.	L	H	M	Secure data communication
Identifiability of the patient based on visit pattern	External	Insufficient anonymization	L	L	L	Anonymization/ Pseudo-nymization
Identifiability of patient based on diagnosis codes	External	Insufficient anonymization	M	H	H	Anonymization/ Pseudonymization
Identifiability of patient based on genomic data (Gene expression, SNPs, DNA sequence data)	External	Insufficient anonymization	M	H	H	Anonymization/ Pseudo-nymization
Identifiability of researcher using analysis tool	External	Researcher is logged	M	M	M	Anonymization/ Authorization
Non-repudiation of patient data in database	External	Data can be somehow linked to patient	L	L	L	Anonymization/ Pseudo-nymization
Information disclosure: Attribute disclosure -> inferring phenotype from genotype	External	Insufficient anonymization	M	M	M	Data protection techniques may not exist / Data to be shared with trusted party (Data Access committee)
Content Unawareness: Patient does not know what data s/he provides and how it is processed	Organizational	Patient is not informed well enough	M	H	H	Standard operating procedure
Policy and consent non-compliance: Insufficient consent; must cover storing the data in the PM database, annotating/processing the data, research and publishing results	Organizational	Insufficient consent*	H	H	H	Consent management

7 Design of the security framework

In this use case, data from open as well as restricted databases should be integrated. **Table** 2 reports the security requirements derived from the survey questionnaire referring to the DFD shown in **Figure** 2. The update of the DFD to the more recent one shown in Figure 1 led to no changes affecting the security requirements of the use case.

The main security requirements identified for WP8 are:

- The (internal) one Data Store (PM analysis data base) and the one Process (PM analysis pipeline) are subject to restricted access (see below, DAC-controlled access). They require the pseudonymization of PM data also internally.
- The WP utilizes several external data source entities. The majority are available via open access, while the data from the EU-OPENSCREEN and EATRIS infrastructures are restricted and provide highly sensitive data.

Table 2. Summary of answers to survey questionnaire (DFD = Data Flow Diagram)

WP8 DFD Data Store Elements				
Name	**Type of Data**	**Individual Level Data**	**Access Mode**	**Security Measures**
PM analysis database	Patient data, Drug data and Lab measurements, Gene expression, DNA Sequence, mutation, diagnosis, etc.	Patient demographic, clinical data and genetic data	Open as well as Restricted	User authentication and authorization system (*) Pseudonymous identity management for PM data (*)

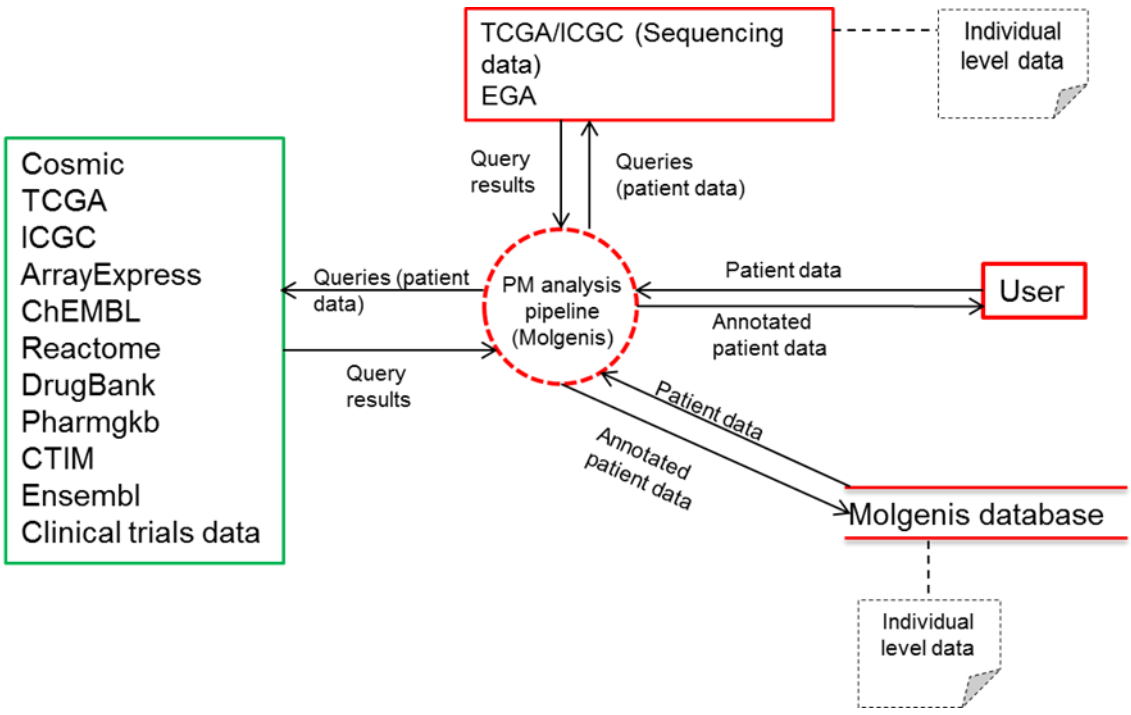
WP8 DFD Process Elements			
Name	**Input Data**	**Output Data**	**Security Measures**
PM analysis pipeline	Patient data (e.g. Gene expression, DNA sequence, mutation data, diagnosis, etc.)	Annotation about input data	User authentication and authorization system (*) Pseudonymous identity management for PM data (*)

WP8 DFD Data Flow Elements			
Name	**Source**	**Destination**	**Security Measures**
Patient data	User	PM analysis tool	Needs to be secured because it is patient data
Annotated patient data	PM analysis tool	User	

WP8 DFD Entity (External) Elements	
Name	**Access Mode (**)**
Cosmic	Open
ICGC	Open
TCGA	Open
ICGC/TCGA (Sequencing data) (***)	Restricted
EGA(***)	Restricted
Array Express	Open
Clinical Trial Information Mediator (CTIM)	Open
Drugbank, ChEMBL, Pharmagkb	Open

Notes:
(*) refers to security measures identified **to be implemented** in the future within the use case.
(**) the access mode indicates: “Open” if the data access is free; or “Restricted” if the data access is granted only after user registration with the external data source.
(***) WP8 currently does not want to access the restricted data from entity namely (EGA, ICGC and TCGA) for the analysis pipeline.

Figure 2. Data Flow Diagram used for the second survey



7.1 User roles

Roles have to be seen in the context of different phases of the research process: collecting data, providing access to other researchers, and integrating data from other resources.

Preparatory work has been done in WP5.1. In collaboration with WP8, a Usage Scenario for the personalised medicine Data Bridge was developed which constitutes the basis for the development of the personalised medicine use case in WP8. The Usage Scenario consists of descriptions of the intended data bridge, aim and motivation, overview of data sources (e.g. Cosmic, ICGC, TCGA, Ensembl, Reactome, Pharmgkb, CTIM, DrugBank), actors involved, requirements/prerequisites, description of processes, events and actions, and a graphic description of the data flow in the form of a diagram. A survey, which was completed by four potential users consisting of three research scientists and one medical practitioner, was conducted to investigate the relevance of the usage scenario for further research in the area. One result was that clinical trials are of relevance in clinical research for personalised medicine. For this reason, the integration of the Clinical Trials Mediator will allow the efficient querying to get better insights into the effect of drugs or genes on patient's health based on clinical trials registers and publications databases (see Section 9.3, D5.1).

The basic data of the personalised medicine use case comprises a set of measurements performed on an individual patient. These data typically include one or more of the following: a detailed clinical history of the patient, the results of various laboratory tests (typically measurements of one or more clinical markers), data on somatic mutations, data on germline mutations, gene expression data, DNA copy number data, protein abundance or phosphorylation status data and drug sensitivity data from ex vivo screening. The nature of the Personalised Medicine Data Bridge is data enrichment, i.e. enhancing the available patient data with various other relevant data and information. We note again that, in this use case, data used for research internally by the data provider are pseudonymous whereas only anonymised data are transferred to external requestors.

Data collection. The purpose and therefore the legal and ethical context of data collection have to be considered. While collection of personalised

medicine data in research is based on informed consent and approval by responsible ethics committees, additional measures are needed if the primary purpose of data collection is healthcare. Secondary use of health care data is addressed by other projects like EHR4CR. We state here that re-use in research projects will typically require informed consent, ethics committee approval and security measures (pseudonymisation or anonymisation of data).

Providing access to data. External access to data is based on the three access layers described in 7.2.

Physician (and other medical personnel treating the patient): Internally, a physician has to be able to insert and update a patient's clinical and demographic data and to view the (interpreted) results of molecular profiling data derived from a patient's samples. Each physician should be able to access only his/her patient's data. If the purpose of data collection is research, pseudonymisation will be used. If healthcare data is collected, only healthcare professionals involved in the care process are allowed to see the data of their patients. If data is reused, additional measures are needed (see above: informed consent, ethics committee approval, pseudonymity or anonymity).

Lab personnel: This internal activity will follow typical rules: if the analysis is done in the research lab, the samples have usually been coded or pseudonymised already and hence the lab personnel just needs to be able to recall data fields relevant for the analysis performed by the pseudonym of the sample code or patient pseudonym. If secondary use of healthcare data is planned, the situation is as described above.

Researcher contracted to interpret the data for the benefit of the physician and patient: From the security perspective, there is no difference to the descriptions above. Research data have to be pseudonymous, whereas data used for healthcare are only available to healthcare professionals involved in the diagnosis and treatment of the disease. This may include bioinformaticians who perform some sort of analysis or operations which help to interpret the data and make it more meaningful and actionable for the clinician. The three cases healthcare, research, and secondary use of

healthcare data for research can also occur under this category and are to be treated as described above.

***Independent researcher*:** Typically, users in this category want to access the data of a cohort of patients matching certain criteria in order to test a hypothesis or for data mining. For these purposes anonymised data suffice, i.e. the users should just be able to see the data fields relevant for their analysis from patients fulfilling the inclusion criteria.

7.2 Access layers

As personalised medicine data has high distinguishability, it has to be secured carefully against re-identification attacks. For access, three layers are needed: In order to get access to individual or fine grained aggregate data, the user needs to specify the data required, the scientific justification and the inclusion criteria for the patients. As described in 5.4 above, all data considered here are research data. They are accompanied by informed consent and identifiers have been replaced by pseudonyms. Patients' personalised medicine data (e.g. germline variants) as well as pseudonymous clinical data comprise sensitive information. Before they can be released they must be anonymised and the release process has to be accompanied by appropriate legal and organizational measures.

The access concept is compliant with what is described in the report to deliverable D5.3, which we refer to for details.

- *Open/*public: no authentication/authorization needed.

For the data collection to be attractive to the users, summary statistics will be made publicly available. On this layer, only high level aggregates will be presented. They are anonymous and based on pre-defined queries.

- *Restricted*: authentication/authorization is required, agreement to terms and conditions will typically be requested. We also refer to this tier as tier 2.

In order to let independent researchers know if the data collection contains anything relevant to them (and is hence worth the bother of applying for access), "existence" queries and access to anonymous and aggregate data

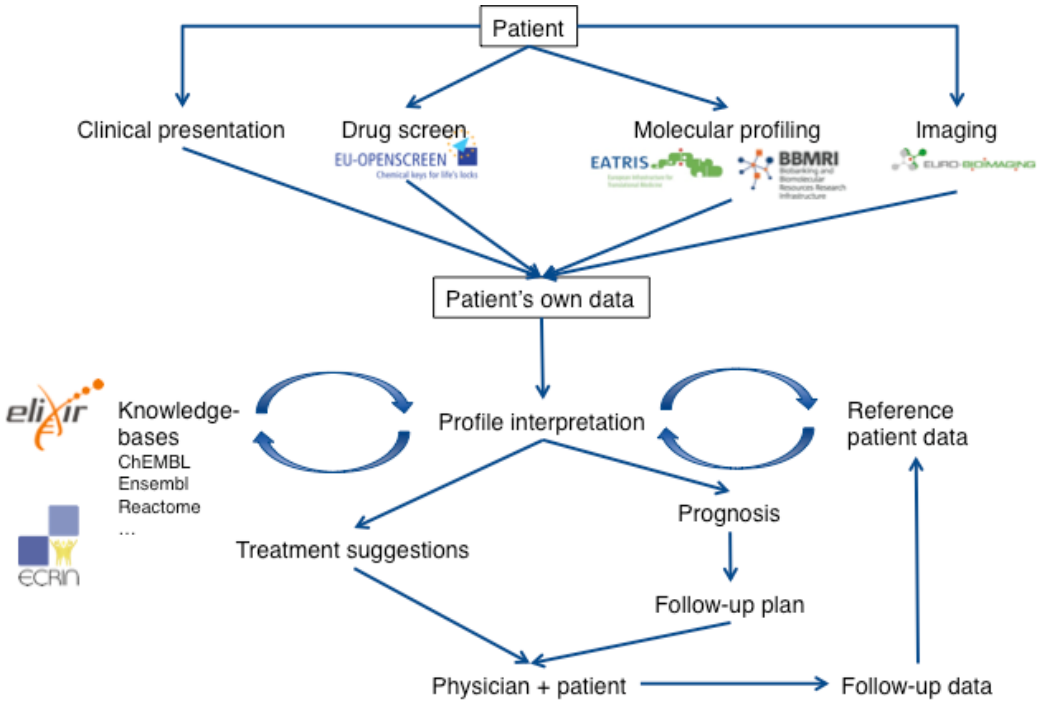
are needed. Queries will return the number of hits (>k) as well as other aggregate data, but no individual level data. Only if the user sees that the data collection contains enough data for his/her purposes will he/she get into contact with the data provider (see below: contact to responsible data controller) and apply for individual-level access. Agreement to terms and conditions will be requested.

- *Committee controlled*: All security measures of the restricted access tier apply, and additionally a review by a committee (e.g. data access committee) is needed before the data is released to the requestor. We also refer to this tier as tier 3.

In this case, the request will be passed to the responsible data controllers. Data access committees will decide if access can be granted on the basis of the available informed consent and existing ethics approval; additional approval may be necessary. Data use agreements regulate the data transfer. Among other items they should confirm that no attempts will be undertaken to re-identify data. Before data release, data will typically be anonymized.

Figure 3 gives an overview of how patient data are enriched with additional data from external databases and used to generate patient profiles that are compared with reference data. Only the profile information is used for treatment decisions and prognosis. The patient's own data contain the personally identifiable data that must be protected. Access and release will only be possible if a proposal has been approved by a data access committee and all accompanying steps (check of informed consent and ethics approval, completed agreements, anonymity when released) have been performed.

Figure 3: Data flows and research infrastructures involved data production



Access to the third level will require an application to a data access committee. The access request must specify which cases/patients are to be selected by completing certain data fields and values.

7.3 Security measures

In order to allow data access in a secure and privacy preserving manner, the security measures identified in deliverable D5.3 will be needed. Section 8.2 in the report to D5.3 provides more details and examples for implementation.

The following key elements of the security architecture described in D5.3 are instantiated for the process described in this report:

- The three different access tiers (open, restricted and committee controlled) are needed.
- Federated authentication using Shibboleth is preferred. The possibility to use a local authentication of the consumer is the fall-back option if the requestor has no account at a trusted identity provider.

- The process of committee approval can be supported using the Resource Entitlement Management System (REMS [18]). In this context SAML, will be used to convey authorization information.
- All communications will be protected using standard SSL/TLS connections.
- Selected sensitive data-at-rest (e.g. a patient's demographics) will be stored encrypted.
- Anonymization will be the standard measure to secure data release in the two restricted access tiers.
- Pseudonymity will be used for research data used internally. If supported by positive Data Access Committee (DAC) approval, it is an option for drill-down in the DAC-controlled restricted access tier.
- Accountability, auditing, and provenance are of high importance. All relevant actions are logged by the bridge endpoints which serve as data providers. These relevant actions are account creation, login events, queries against restricted data, application for individual-level data, and access to and sharing of individual-level data. Provenance traces should be kept where Open Provenance Model (OPM)[8] or PROV-based[9] provenance data modelling can be helpful. Relevant resources and lists of implementations for this task are presented in the report to deliverable 5.3.
- As a non-technical counterpart, the legal framework and the forms provided by BioMedBridges deliverable 5.2 [19] will be used. This includes data use agreements templates, and access committee guidelines.

We refer to the report to deliverable 5.3, where the underlying security architecture has been presented in detail. In D5.3, articles by Mello et al [20], Malin et al [6], and Curcin et al [21], representing relevant contributions to data sharing and architectures are summarized. Focussing on clinical trial data, Mello et al. have analysed policies of data sharing, its benefits, risks, and legal implications, leading to an identification of core principles. Malin et al. have discussed technical and policy approaches and given relevant recommendations. Curcin et al. have covered the important aspect of provenance and the need for provenance-aware system implementations.

[8] http://openprovenance.org/
[9] http://www.w3.org/2011/prov/wiki/Main_Page

7.4 Instantiation of the pseudonymization measure in the context of Work Package 8

For the collection of research data (see above), data separation and pseudonymity are needed to secure follow-up data. Basically, the data will be split into two repositories with two independent (sub)applications: one, the “patient list application”, will contain the patients’ demographic data – name, social security number, address – and the other, “personalised medicine data application”, will contain the sensitive PM data with pseudonymous identifiers only. TUM-MED will provide a solution for the patient list application, which can be extended to double pseudonymity (also called “double coding”). It will be integrated with the PM database.

Under this concept, if the patient list application is compromised, an attacker has access to the demographic data only and cannot associate the rest of the data to the identifying data of the patient list without having to compromise the personalised medicine data application too. An attacker has to compromise two independent systems to link the demographic data of a patient to the corresponding personalised medicine data.

Both applications should have independent user management, i.e. each application is responsible for its own authentication and authorization process.

The workflow for data entry by the clinical personnel could look like this:

A user logs in to the patient list application and creates a new patient data entry by entering the demographic data and saving them in the patient-list application. After this step, the user clicks on a link to perform the context switch to the personalised medicine data application. During the context switch, the user gets logged into the personalised medicine data application automatically and the data of the patient selected is displayed. Now the user can enter additional data into the personalised medicine data application.

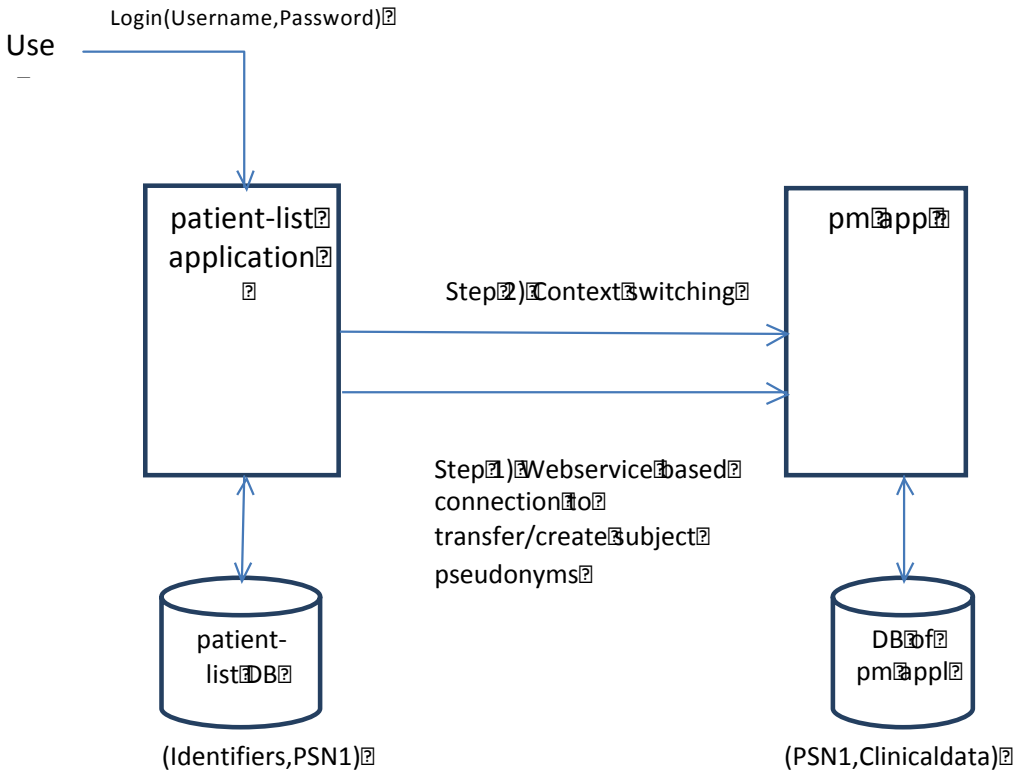
The interaction between the patient-list and the personalised medicine data application could look as in Figure 4. For each new patient created in the patient-list application, this application assigns a new pseudonym to the patient. This pseudonym has to be transferred to the personalised medicine data application. This transfer/creation is done using a webservice call originating from the patient-list application to the personalised medicine data application (Step 1).

During step 2, the personalised medicine data application is called, the user is logged in, and the data belonging to the selected patient are displayed. To realize this, username and password of the user and the pseudonym of the patient are transferred to the personalised medicine data application. This step happens only client-side (most likely in the browser of the user, and the context switch will most likely be implemented using Javascript).

Requirements on the personalised medicine data application side are:

- Authentication and authorization mechanisms
- Webservice API to create/update(/delete) new patients (represented by a pseudonym)
- Possibility to perform the context switch, i.e. a “scriptable” GUI

Figure 4 Secure Data Management for Research Data in FIMM personalized medicine use case: Usage scenario diagram for possible interaction of the patient-list application with the personalized medicine data application: Identifiers in “clinical data” have been replaced by pseudonyms



8 Processes for secure sharing of and access to personalized medicine data

Building on the work carried out in the secure access work package (WP5), we show a process by which a provider of personalised medicine data can share and the consumer of the data can gain access to the data in a secure manner. We focus on the most restrictive data bridge (No. 6 as described in Section 0), where a consumer tries to access the enriched, person-related data of WP8.

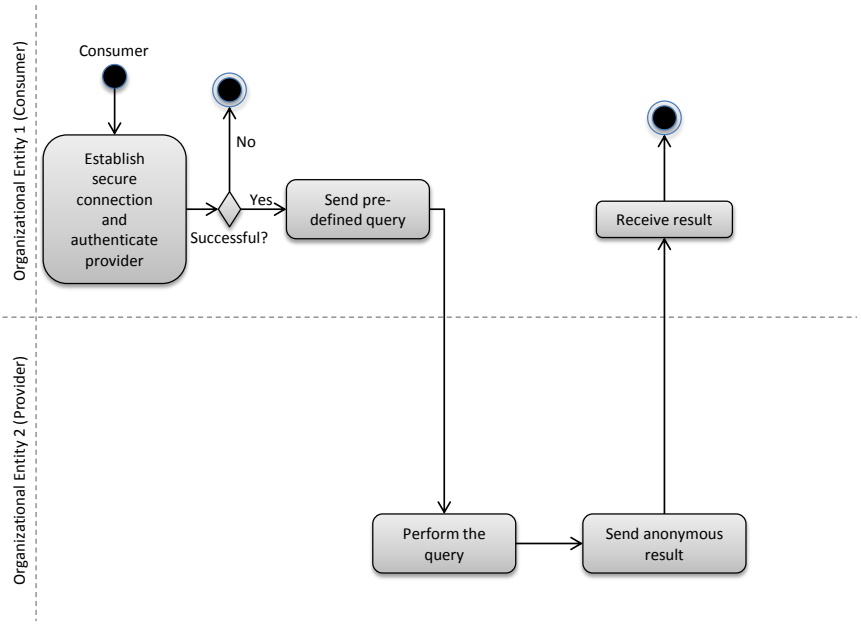
In the following the process steps to access personalised medicine data are described.

The two main parties involved in the processes are the data consumer (often also called user) and the data provider. The identity provider is a third party that can play a role during the authentication process. For further definitions and concepts refer to the report to 5.3.

8.1 Access to and sharing of open data

The PM use case will provide access to some open, anonymous, coarse grained aggregates and metadata to any consumer. Even in this case it is recommended that the communication takes place over a secure channel. To this end, the data consumer establishes a SSL/TLS connection with the provider, authenticating the provider by his certificate. The consumer can execute pre-defined queries for the open data. We refer to D5.3, where we cite Malin et al. [6] who suggest prior approval by a DAC, here too. These pre-defined queries will produce anonymous data. Concerning anonymity, we refer to D5.3, Section 8.2.5. This process is depicted in Figure 5.

Figure 5: Access to open data



8.2 Registration process

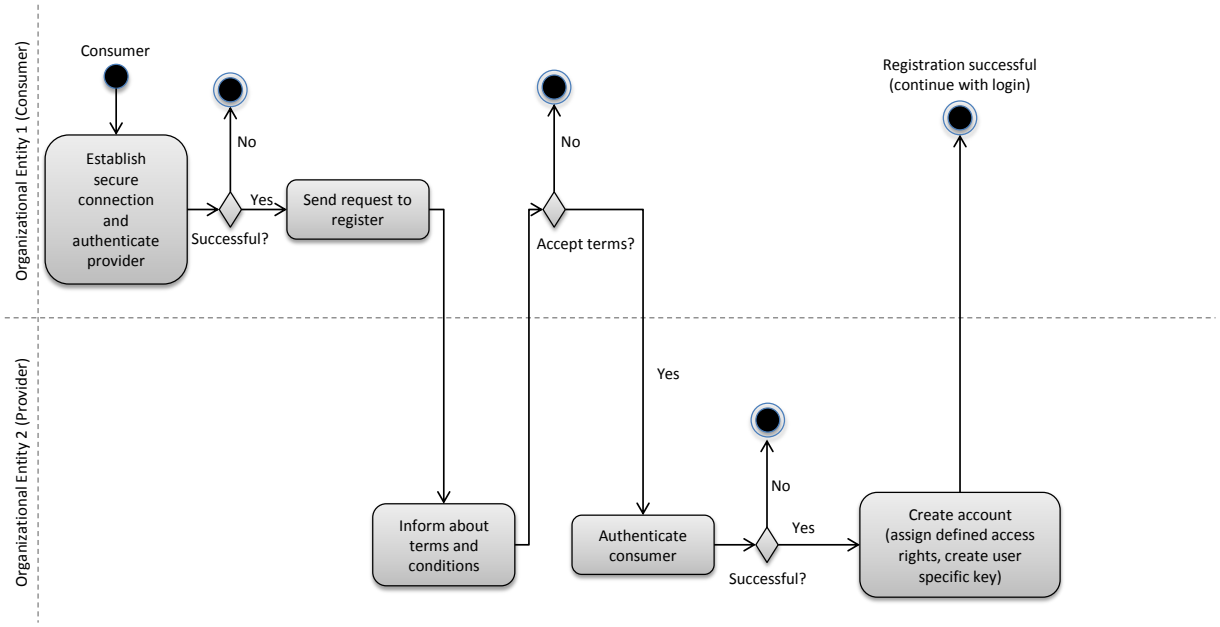
For tiers 2 and 3, an account has to be registred In order to enable the system of the data provider to recognize the consumer. To create an account, the consumer first establishes a secure SSL/TLS connection with the data provider. The provider authenticates themselves using a SSL/TLS certificate from a trusted certification authority (CA). As a next step, the consumer sends the registration request to the system and thereby establishes a session with the data provider. Each subsequent communication can therefore be attributed to the requesting consumer. The session information is conveyed by a session token. This session token is only valid for a defined amount of time.

As a next step, the consumer is presented with the terms and conditions. These include rules for accessing aggregate, metadata and individual-level data. Access to these tiers is described in the following sections.

The account can either be created as a local account or, preferably, the consumer/user can be authenticated by the identity provider of her/his institution. We recommend the use of the federated Shibboleth system [6] for authentication as most home institutions of the researchers (consumers) are providing this kind of authentication service.

If the local account creation or federated authentication is successful the system creates the consumer account. The account information created also includes a consumer specific secret key which will later be used to derive consumer specific pseudonyms. The registration process has to be performed only once and is shown in Figure 6.

Figure 6: Registration process



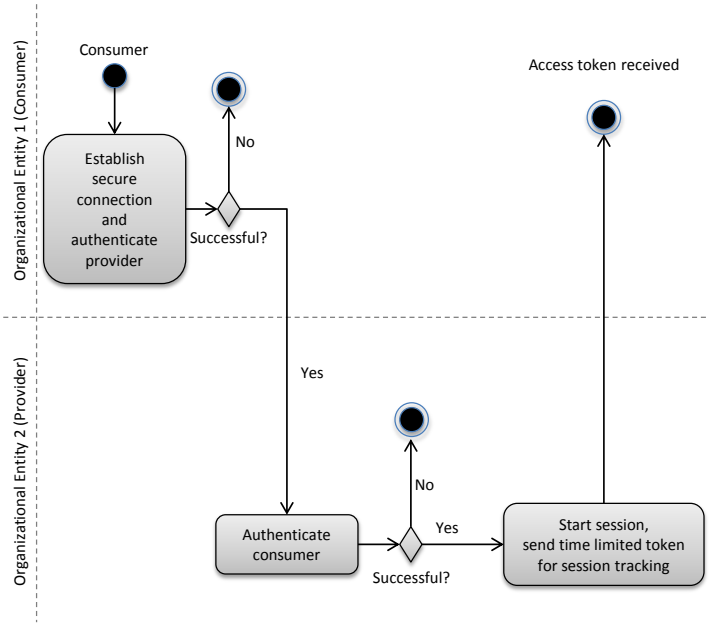
8.3 Authentication process

The authentication or log in process of the consumer is the prerequisite for access to and sharing of restricted data. As a first step, a secure channel is created between the consumer and the provider. During the establishment of the secure channel the provider authenticates themselves to the consumer using their TLS/SSL certificate. Following this, the authentication of the consumer can be performed locally, i.e. the data provider authenticates the consumer or another entity, the identity provider authenticates the consumer. We recommend the use of the federated Shibboleth system for authentication, as most home institutions of the researches (consumers) are providing this kind of authentication service, for details see D5.3.

In both cases, after a successful authentication, a session is established between the data consumer and the data provider. This session is time limited. The process is shown in Figure 7. In the following it is assumed that

this session has been established via a session token which will be transferred to the data provider for each request.

Figure 7: Authentication process



8.4 Access to restricted data

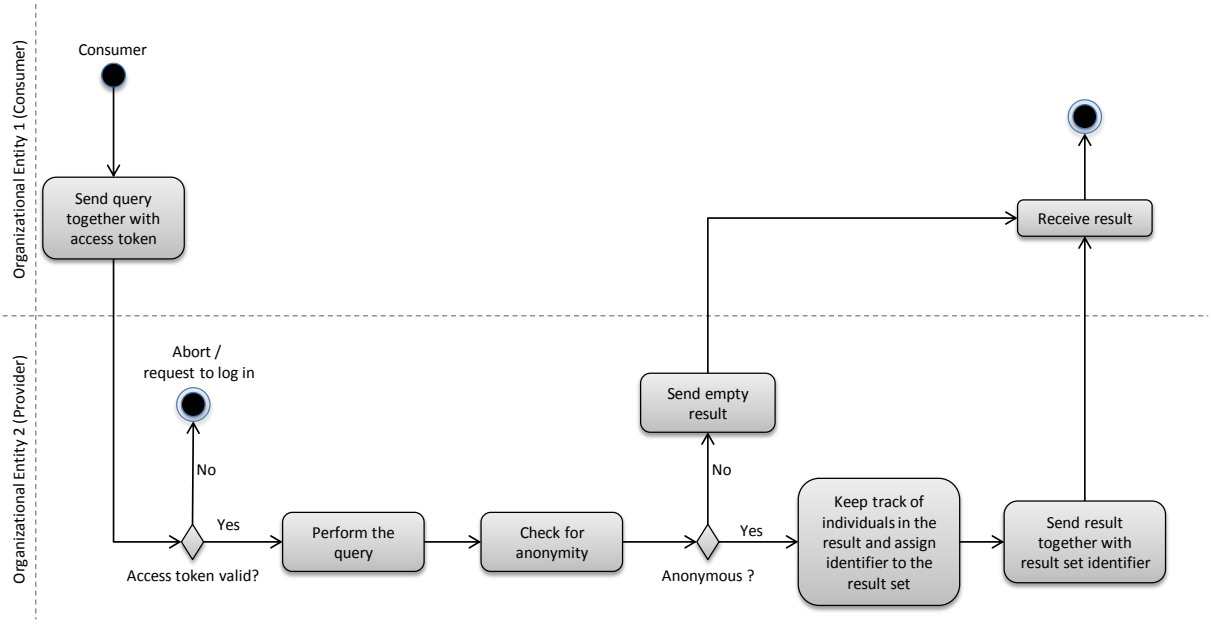
In this process step, the data consumer/requestor aims at finding data matching his/her research questions in order to feed them into his/her database. To this end, she/he formulates a query which returns the number of matching individuals for a set of specific criteria. To prepare an application to individual-level data, an (anonymous) identifier for the query result can be returned. This step will grant access to anonymous data for which oversight is desired and/or access to data underlying intellectual property (IP) protection requirements.

First, the authentication process described in the previous section has to be completed. A session is established and the communication channel is secured.

The consumer can now send a query for aggregate data together with the access token to the data provider. The session token is validated and, if successful, the aggregate query is executed. Pre-defined queries plus query-set-size control will be used to ensure anonymity (see D5.3, Section 8.2.5 and

[22]). The identifiers of the entities contributing to the result can be stored temporarily with an identifier assigned to this set. The process is depicted in Figure 8.

Figure 8: Access to restricted data



8.5 Applying for access to individual-level data

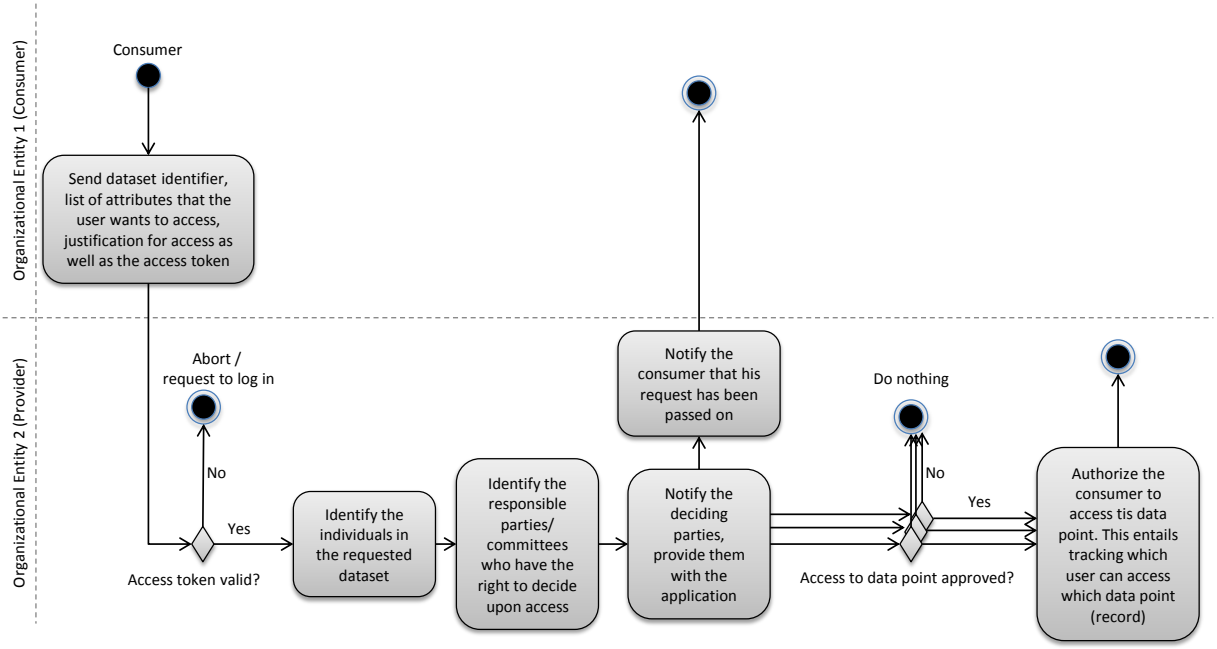
If, for example based on the insights from the aggregates, the user decides to apply for access to individual-level data, interaction with the third access tier starts. The request needs to be checked and approved by the appropriate deciding parties, i.e. the DAC. Each request can require the approval by several deciding parties. DACs may request the scientific reasons in an application for access. They will also make sure that access is covered by informed consent and necessary ethics approval.

After the authentication process has been completed, a session is established and the communication channel is secured.

The consumer now sends the access token together with the application for access including the cohort/data set identifier, a justification for access and the attributes s/he wants access to. The provider validates the provided access token. If the token is valid, the provider identifies the responsible data controller(s) as well as the responsible DAC by using the dataset identifier provided. The requestor is notified once the deciding committee receives the

application. The decision of each deciding party is documented electronically, and access rules are added to the applicant's account. The decision also includes the requirements concerning how to sanitize (typically: anonymize) the data before release. In case access has been granted, the consumer will be notified the next time she/he logs on. This process is shown in Figure 9. As a suitable implementation to support the authorization workflow, REMS [18] has been suggested in D5.3. REMS is an open source tool managing access rights to research resources that assists both researchers requesting data access and DACs granting access.

Figure 9: Applying for access to individual-level data



8.6 Access to individual-level data

If access to individual-level data is granted, the typical mechanisms take place: the consumer has to be authenticated, the session established and the communication channel secured.

To download data, the consumer can submit the query together with his/her session token. After validation of the session token, the query is executed. All records to which the consumer does not have access are filtered out of the result.

The basic security measure for release of individual-level data is anonymisation. This means that internal identifiers are recoded in a way that they are consistent, but no linkage to these will be possible after the data is released to the consumer. The filtering step includes further means to anonymize the individual-level data (as described in D5.3, Section 8.2.5). To protect the data during transfer they are encrypted before they are sent to the consumer. The respective key is delivered to the consumer out of band (i.e. over another communication channel other than the main channel, e.g. by phone). This process is depicted in Figure 10. On the side of the data consumer, all measures described by and agreed upon in the data use agreements have to be performed.

A specific option to support browsing is the release of pseudonymous data. It has to be noted that pseudonymity goes beyond just replacing direct identifiers, so additional security measures as described in D5.3, Section 8.2.5 are needed. The DAC application has to describe the scientific process and justify the needs for accessing and using the data while data use agreements have to explicitly cover the specific process. Figure 11 illustrates the process: pseudonymity with consumer-specific pseudonyms will allow the requestor to later refine the query. Pseudonymous identifiers included in the result will be encoded for the consumer, using the consumer-specific secret key created during account generation.

Figure 10: Release of anonymized individual-level data

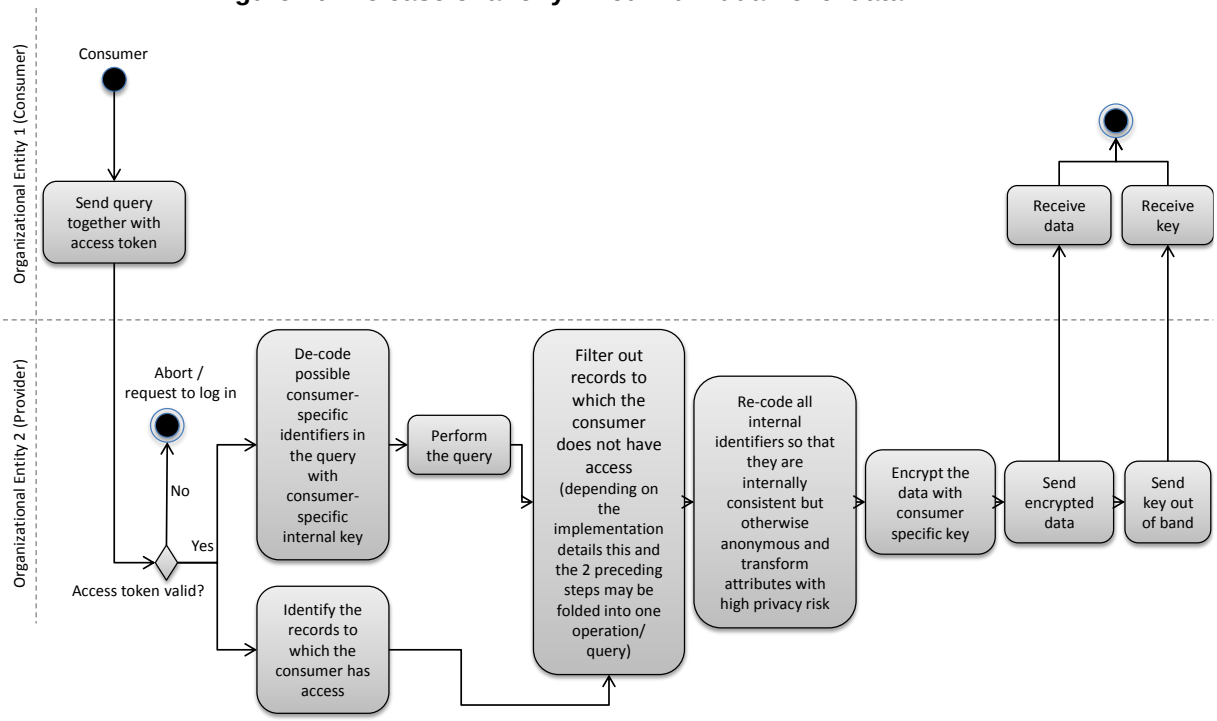
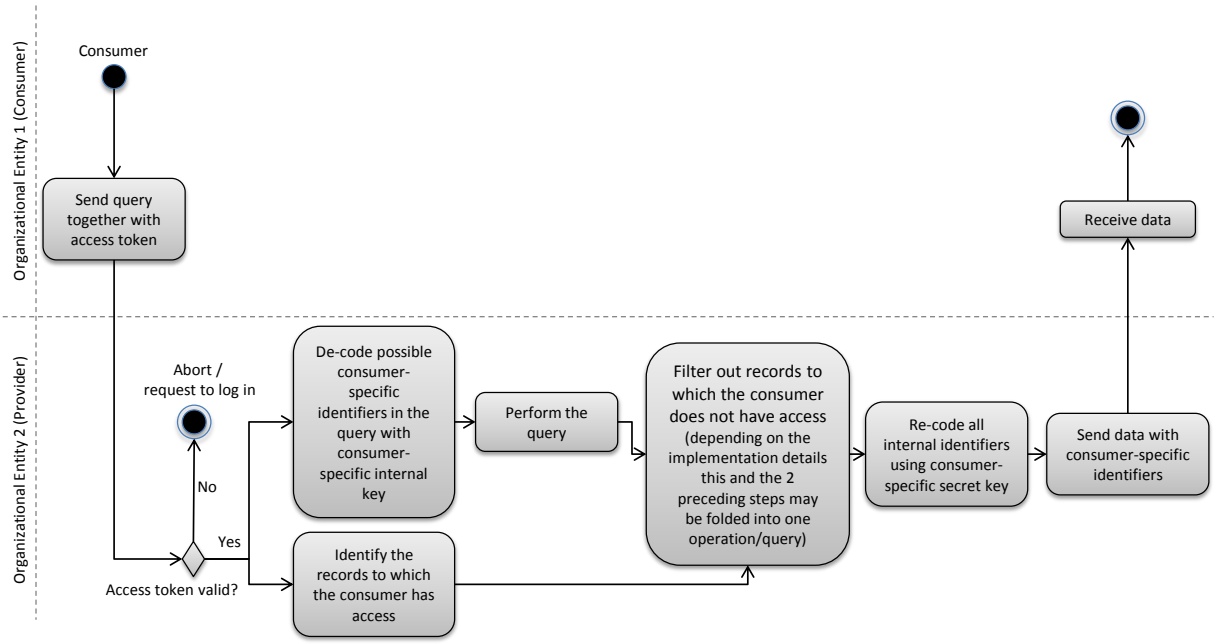


Figure 11: Browsing of pseudonymous individual-level data



9 Delivery and schedule

The delivery is delayed: ☐ Yes ☑No

Appendix A: Survey 2

Questionnaire

The following questionnaire serves for getting a concrete basis to build the security architecture and framework for BioMedBridges. Therefore we want to extend the usage scenarios to Use Cases. The aim is to get on the one hand more technical information and on the other hand information about the parts of the Use Case, that are not covered by the usage scenario. Therefore we provide questions concerning data stores, processes, external entities and data flows in your Use Case. You can find a prefilled version in the bottom with the information gathered from the usage scenarios. Please look through them and correct them if necessary. We are highly interested in the technical details; try to be as concrete as possible. If your Use Case deals with further data stores, processes, external entities, data flows just add them in the same way. This can be done every time when something has changed.

General

What are the differences between your usage scenario and your Use Case? What is not covered by the usage scenario? In the following we want to get a clearer impression of your Use Case. For each data store, data flow, process, external entities (meaning interactions crossing the border of your system(s) to users, web services etc.) existing in your Use Case please answer the questions separately.

Data store

What kind of data store do you use (e.g. database)? Please give a concrete description (what kind of database or database management system, e.g. MySQL?).
What kind of data do you store in that data store?
Does your data contain personal data? Is the data pseudonymous/anonymous?
What format does this data have?
What security measures already exist (e.g. authentication system via email, authorization system via role-based access control, data validation, encryption, audit trail, k-anonymity etc.)?

Process

What program/calculation is executed?
What is the input of the process? Which data format?
What is the output of the process? Which data format?
Is there a process which is executed together with this process (name the abbreviation, e.g. P2)?
What security measures already exist (e.g. authentication system via email, authorization system via role-based access control, data validation, encryption, audit trail, k-anonymity etc.)?

External entities

Who is the external entity (user, web service, server etc.) outside your system, you develop for the Use Case?
Do you need an authentication/authorization mechanism for the external entity getting access, to check who the external entity is and which rights it has? Why is it needed?
Do you already have an authentication/authorization mechanism? If so, which?

Does the external entity itself have an authentication/authorization system your process/data flow/data store has to use? If so, which?

Data flow

Which data flow do you describe, between what process/data store/entity?
Is there a data flow that exists in parallel?
What data is transferred?
How is the data transferred?
Is the transferred data confidential?
Is there data in the data source that is not allowed to be transferred, e.g. besides anonymous data also personal data?
What security measures already exist (e.g. authentication system via email, authorization system via role-based access control, data validation, encryption, audit trail, k-anonymity etc.)?

10 Background information

This deliverable relates to WP8 Use case: Personalized Medicine. Background information on this WP as originally indicated in the description of work (DoW) is included below.

WP8 Title: Use case: Personalized Medicine
Lead: 16: UH
Participants: EMBL, KI, UDUS, TUM-MED, UH

WP8 will integrate complex data sets to understand disease pathogenesis and improve biomarker and treatment selection

Work package number	WP 8	**Start date or starting event:**			month 1			
Work package title	Use case: Personalized Medicine							
Activity Type	RTD							
Participant number	1: EMBL	3: KI	5: UDUS	7: TUM-MED	16: UH			
Person-months per participant	16	8	5	8	32			

Objectives

1) Definition of a process for secure sharing of and access to personalized medicine (PM) data.
2) Definition of existing PM data types and mappings between them.

3) Pilot the use of PM data to support the clinical decision making process.

Description of work and role of participants

Use case: Personalized Medicine - integrating complex data sets to understand disease pathogenesis and improve biomarker and treatment selection

Task 1. Develop a process for secure sharing of and access to PM data
Building on the work carried out in Secure access work package (WP5) we will develop a process by which a producer of the data can share and the user of the data can gain access to the PM data in a secure, legal yet easiest possible manner. FIMM will have the role of a prototype PM data provider as well as a user. TUM, as the leader of WP5 will provide expertise in privacy protection as well as secure sharing and access matters.

Task 2. Define types of PM data and mapping between them
Measurements made with different technologies may not be (and usually are not) directly comparable even though the underlying thing measured (e.g. certain mRNA level) may be exactly the same. This creates a situation where a user of the data may inadvertently be "comparing apples with oranges". To avoid that we will catalogue data types (as well as pertinent standards) relevant to PM and provide mapping between them if applicable. FIMM will provide PM domain expertise. KI, as the leader of the Standards work package (WP3) will provide know-how of existing standards.

Task 3. Develop a PM informatics pipeline
As a proof of concept that the tasks above facilitate the interoperability of different PM data types we will develop a prototype PM informatics pipeline to support the decision making process in PM. This prototype pipeline will utilise the data type specifications and standards established in Task 2 and be subject to constraints of access procedures established in Task 1. FIMM will be a prototype PM data producer and user. EMBL-EBI as the leader of Technical Integration work package (WP4) will provide expertise on general framework and architecture of the implementation.

Deliverables

No.	Name	Due month
D8.1	Process specification for secure sharing of and access to PM data	30
D8.2	Definition of PM data types (report)	30

D8.3	Demonstration of interoperability between different types of PM data	48

11 References

[1] BioMedBridges, Deliverable 5.1 - Report on regulations, privacy and security requirements, http://www.biomedbridges.eu/deliverables/51-0, June, 2013.

[2] M. Howard und S. Lipner, The security development lifecycle: SDL, a process for developing demonstrably more secure software, Microsoft Press, 2006.

[3] M. Deng, K. Wuyts, R. Scandariato, B. Preneel und W. Joosen, A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements, Requirements Engineering, Volume 16, Issue 1, pp 3-32. http://dx.doi.org/10.1007/s00766-010-0115-7, March 2011.

[4] J. Donald S. Le Vie, Understanding Data Flow Diagrams, last access April 2014.

[5] NIST Special Publication 800-30, Guide for Conducting Risk Assessments, September 2012.

[6] B. Malin, D. Karp und R. H. Scheuermann, „Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research.," *Journal of investigative medicine : the official publication of the American Federation for Clinical Research,* Bd. 58, Nr. 1, pp. 11-8, 2010.

[7] C. Heeney, N. Hawkins, J. de Vries, P. Boddington und J. Kaye, „Assessing the Privacy Risks of Data Sharing in Genomics.," *Public health genomics,* 2010.

[8] B. Malin, G. Loukides, K. Benitez und E. W. Clayton, „Identifiability in biobanks: models, measures, and mitigation strategies.," *Human genetics,* pp. 1--10--10, 2011.

[9] K. El Emam, E. Jonker, L. Arbuckle und B. Malin, „A systematic review of re-identification attacks on health data.," *PloS one,* Bd. 6, Nr. 12, p. e28071, 2011.

[10] F. K. Dankar, K. El Emam, A. Neisa und T. Roffey, „Estimating the re-identification risk of clinical data sets.," *BMC medical informatics and decision making,* Bd. 12, Nr. 1, p. 66, 2012.

[11] K. Benitez und B. Malin, „Evaluating re-identification risks with respect to the HIPAA privacy rule.," *Journal of the American Medical Informatics Association : JAMIA,* Bd. 17, Nr. 2, pp. 169-77, 2009.

[12] G. Church, C. Heeney, N. Hawkins, J. de Vries, P. Boddington, J. Kaye, M. Bobrow und B. Weir, „Public access to genome-wide data: five views on balancing research with privacy and protection.," *PLoS genetics,* Bd. 5, Nr. 10, p. e1000665, 2009.

[13] K. E. Emam, Guide to the de-identification of personal health information, 1st Hrsg., Auerbach Publications, 2013.

[14] F. K. Dankar und K. El Emam, „A method for evaluating marketer re-identification risk," in *Proceedings of the 1st International Workshop on Data Semantics - DataSem '10*, New York, New York, USA, 2010.

[15] T. Truta, F. Fotouhi und D. Barth-Jones, „Disclosure risk measures for microdata," *15th International Conference on Scientific and Statistical Database Management, 2003.,* Nr. 3, pp. 15-22, 2003.

[16] C. J. Skinner und M. J. Elliot, „A measure of disclosure risk for microdata," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* Bd. 64, Nr. 4, pp. 855-867, 2002.

[17] D. Lambert, „Measures of Disclosure Risk and Harm," *Journal of Official Statistics,* Bd. 9, 1993.

[18] M. Linden, T. Nyrönen und I. Lappalainen, Federated authorisation: Resource

Entitlement Management System, TERENA Networking Conference (TNC), 2013.

[19] BioMedBridges, Deliverable 5.2 - Tool for assessment of regulatory and ethical requirements; including supportive documents, http://www.biomedbridges.eu/deliverables/52-0, December 2013.

[20] M. M. Mello, J. K. Francer, M. Wilenzick, P. Teden, B. E. Bierer und M. Barnes, „Preparing for Responsible Sharing of Clinical Trial Data," *N Engl J Med,* 24th October 2013, 369(17):1651-8.

[21] V. Curcin, S. Miles, R. Danger, Y. Chen, R. Bache und A. Taweel, „Implementing interoperable provenance in biomedical research," *Preprint submitted to Future Generation Computer Systems,* 9 December 2013.

[22] N. Adam und J. Worthmann, „Security-control methods for statistical databases: a comparative study".*ACM Computing Surveys (CSUR), 21(4).*