ICT-2011-288048

EURECA

Enabling information re-Use by linking clinical Research and CAre

IP
Contract Nr: 288048

Deliverable: D5.1 Requirements analysis and knowledge discovery scenarios

Due date of deliverable: (31-08-2012)
Actual submission date: (15-10-2012)
Resubmission date: (07-08-2013)

Start date of Project: 01.02.2012 Duration: 42 months

Responsible WP: UOXF

Revision: <outline, **draft,** proposed, accepted>

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Service)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
Francesca Buffa	UOXF	francesca.buffa@imm.ox.ac.uk
Stefan Rueping	IAIS	stefan.rueping@iais.fraunhofer.de
Andre Dekker	MAASTRO	andre.dekker@maastro.nl

0.2 Documents history

Document version #	Date	Change
V0.1	20.05.2012	Starting version, template
V0.2	10.06.2012	Definition of ToC
V0.3	15.08.2012	First complete draft
V0.4	15.08.2012	Integrated version (send to WP members)
V0.5	31.08.2012	Updated version (send PCP)
V0.6	31.08.2012	Updated version (send to project internal reviewers)
Sign off	31.08.2012	Signed off version (for approval to PMT members)
V1.0	31.08.2012	Approved Version to be submitted to EU
V1.1	27.06.2012	Review to be submitted to EU (deliverable was approved pending editorial review)
V2.0	07.08.2013	Resubmitted version

0.3 Document data

Keywords	Requirements analysis, KD scenarios
Editor Address data	Name: Francesca Buffa Partner: UOXF Address: Weatherall Institute of Molecular Medicine, University of Oxford Phone: 0044 (0)1865 222440 E-mail: francesca.buffa@imm.ox.ac.uk
Delivery date	

0.4 Distribution list

Date	Issue	E-mailer

Table of Contents

1 Introduction

1.1 Purpose and structure of the deliverable

EURECA's goal is to enable a seamless, secure and consistent link between clinical research and clinical care systems. Such a link would enable for example the healthcare professionals to extract, in the context of each patient's case, the relevant data out of the overwhelmingly large amounts of heterogeneous patient data and treatment information. On the other end, it would help a more effective and efficient execution of clinical research, early detection of patient safety issues, faster transfer of new research findings and guidelines to the clinical setting (from bench-to-bedside).

EURECA plans to achieve this by implementing a set of loosely-coupled software services and tools which will be deployed in the context of pilot demonstrators ("the scenarios"). These scenarios will constitute the basis for tool development; however, a modular approach will be used to ensure re-usability and scalability of the solutions.

A core element of EURECA is to develop such solutions whilst fulfilling the data protection and security needs and the legal, ethical and regulatory requirements related to linking research and EHR data and allowing the extraction of data based on relevance and its contextualization to the patient case.

The above key aims and components of EURECA will be reflected in the solutions suggested for the data-mining and the purpose of this deliverable is to describe the data-mining requirements to support the EURECA scenarios. The deliverable has a specific focus on knowledge discovery scenarios, and these scenarios will be expanded and developed further here.

General key-points of the EURECA project are re-usability, compatibility and standardization of all components. In this context, we will try to build upon several previously defined concepts and methodologies whenever this is possible and point out areas for further methodological work.

The structure of this deliverable is as follows. A high-level illustration of the deliverable reviews the findings and the scenarios described in deliverable D1.1 "User needs and specification for the EURECA environment and software services" in chapter 2. These needs set up the landscape for a more in-depth requirements analysis for data-mining, which is the subject of chapter 3. Chapter 4 focuses and expands on the knowledge discovery scenarios.

2 Data-mining in EURECA

EURECA wants to deliver measurable benefit to various communities ranging from the patients themselves, to the clinical professional, research and industry.

As stated in to the EURECA DoW the main objectives are:

1. **Support more effective and efficient execution of clinical research** by:
    a. Allowing faster eligible patient identification and enrolment in clinical trials,
    b. Providing access – in a legally compliant and secure manner – to the large amounts of patient data collected in the EHR systems to be re-used in clinical research, for new hypotheses building and testing (e.g. to benefit rare diseases), study feasibility, as well as for epidemiology studies,
    c. Enabling long term follow up of patients, beyond the end of a clinical trial,
    d. Avoid the current need for multiple data entry in the various clinical care and research systems during the execution of a study.
2. **Allow data mining of longitudinal EHR data for early detection of patient safety issues** related to therapies and drugs that would not become manifest in a clinical trial either due to limited sample size or to limited trial duration, and eliminate duplicate reporting (in care and research) of identified serious side effects,
3. **Allow for faster transfer of new research findings and guidelines** to the clinical setting (from bench-to-bedside),
4. **Enable the healthcare professionals to extract, in the context of each patient's case, the relevant data** out of the overwhelmingly large amounts of heterogeneous patient data and treatment information.

Thus, the basic requirement is an approach that focuses on specific and realistic clinical questions, or scenarios, initially, whilst adopting a more modular, standard-based and scalable approach that could be easily generalized at a later stage.

The following section aims at extracting the data-mining requirements from concrete scenarios and clinical questions. The further section will expand on and describe some of the identified areas of data mining.

2.1 Scenario based requirements

The first task of EURECA clinical partners has been to develop general and modular scenarios that cover the above aims and define their specific application. Such scenarios have been described in a previous document; the "User needs and specifications for the EURECA environment and software services" deliverable D1.1, chapter 3.

2.1.1 The EURECA scenarios

A general classification was suggested in deliverable D1.1 that divides the EURECA scenarios as classified into three categories:

- **Knowledge discovery**
  - Selection of best trials for a patient
  - Trial / protocol feasibility
  - Selection and inclusion of patients into trials
  - Detection and prediction of SAEs / SUSARs
  - Pharmaco-vigilance – Automatic reporting of SAEs and SUSARs
  - Early detection and prevention of diseases
  - Personal medical information recommender
  - Develop or update guidelines for diseases
  - Data mining of consultations
  - Analyse economic data between different procedures / approaches
  - Build, optimize, validate or update a diagnostic classifier
  - Build predictive models of late morbidity
- **Data curation**
  - Long term follow-up
  - Patient diary (connection between PHR and data management tools)
- **Basic research and clinical trials support**
  - Supporting design of new trials and hypothesis generation
  - Mining of information from EHR and CT to validate research/clinical hypotheses
  - Clinical data reuse
  - Opt-out solution for new research
  - Simulation of datasets to combine
  - Rapid learning
  - High-level genomic meta-analyses in genomic/genetic
  - Association studies

EURECA scenarios are still evolving and are being defined by clinical and technical partners together within the activities of WP1. The above scenarios reflect the activities and discussion within EURECA so far. A general schema for most of these scenarios is already available and presented in Chapter 3 of the D1.1 deliverable; others will be discussed in this document.

In the next section we will discuss the requirements, the flow, the implications and challenges of these scenarios from a data-mining perspective solely. The scenarios are briefly described here again; however a full description is outside the scope of this deliverable. We reference a previous deliverable, D1.1, where an initial description is provided on how the data and the extracted information and knowledge are managed.

In the current deliverable the focus with respect to knowledge management is on the data mining methods aspect so how data is extracted and how knowledge is delivered back to the user are not discussed. In particular, we attempt a first description of the data-mining flow and requirements based on the information deposited in D1.1, which is the current knowledge and level of detail of the scenarios. We are though aware that this knowledge is evolving and a second WP1 deliverable D1.2 will discuss all scenarios in more depth. After D1.2 is finalized, we will then be able to refine the indications for the data-mining specific needs. These will be illustrated in a second WP5 deliverable D5.2, where we will have a better understanding of the detailed flow for each scenario and we will match the needs with a state-of-art review. This deliverable and D5.2 are to be considered as complementary.

2.1.2 Dissecting the scenarios data-mining components

In the following sections we attempt to extract and discuss in detail the main data-mining requirements and challenges of the EURECA scenarios. Tables summarising this information are provided for each scenario.

This analysis is based on the scenario flows provided by WP1. As we are just in the beginning of the project, these flows are still very general, but it is already possible to extract some key relevant information for the data-mining and build some first strategies. Specifically, we will focus on the following aspects for each scenario:

- **Pre-conditions:** these are the pre-conditions that need to be verified for the data-mining to start, these are not the scenario pre-conditions. The focus here is solely in data-mining.

- **Expected input database types:** these are the expected databases and databases types based on the current description of the scenario

- **Expected input data types:** these are the expected data types based on the current description of the scenario

- **Expected output:** this is the expectation regarding the output information of the data-mining, this is not necessarily the output of the scenario but sometimes the two correspond

- **Expected output formats:** this is the expectation regarding the output format of the data-mining (e.g. file types) based on the current definition and understanding of the scenarios as depicted in D1.1

- **De-anonymized data needed:** forecast on whether de-anonymized data might be needed as this can have implication for data-mining choices

- **Data-mining components:** these are the expected data-mining components based on the scenario schematic flows generated within WP1 by the clinical and technical partners, and by the initial developments and D1.2 draft.

- **Requirements and methods:** these are highlights and suggestions based on the current understanding of the scenarios and related components

- **Potential challenges:** these are the potential challenges specifically for the data-mining components

2.1.2.1 Knowledge discovery scenarios

The goal of this scenario is to find the optimal trial that fits the needs of the patient the best. The schema and initial information flow is presented in D1.1 figure 3.4.

Selection of trials for patient enrolment

The goal of this scenario is to find the optimal trial that fits the needs of the patient the best. The schema and initial information flow is presented in D1.1 figure 3.4.

Pre-conditions	1) Patient data are stored in the PHR and EHR 2) Access to:- PHR, EHR, clinical trials databases and literature
Expected input database types	- PHR and EHR: relational databases, free-text components - Clinical trials: relational databases - Existing knowledge and literature: relational databases, document-oriented databases (e.g. XML)
Expected input data types	- PHR, EHR and clinical trials databases: qualitative nominal and ordinal, quantitative discrete and continuous - Existing knowledge: qualitative nominal and ordinal, quantitative discrete and continuous, string character data
Expected output	Matched and ranked trials log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML
De-anonymized data needed	Only for the specific patient receiving care
Data-mining components	1) Mining of PHR or EHR to obtain the data for each patient 2) Mining of clinical trial databases and identification of a trial for the specific patient 3) Mining of the literature and identification of suitable trials that have been completed
Requirements and methods	1) Similarity learning to discover similar datasets and similar trials.

	2) Frequent pattern mining to discover frequently occurring patterns to summarize datasets and to aid similarity learning 3) Text mining of eventual unstructured part of EHR databases and literature All the data-mining steps will require some investigation into time-dependent data mining methods that effectively address the continuously changing landscape in EHRs, trial databases, literature, public databases. Algorithms to effectively detect changes in data, annotation and structure of databases will be an advantage. Dimensionality reduction of high-throughput and imaging data might be required if these modalities are implemented in the trials and are part of the trial eligibility criteria.
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Representation of semantic similarity in such a complex data space is challenging ➢ Time dependency ➢ Data are not anonymous: privacy issues needs to be addressed, possibly by performing the data-mining at the hospital site

Trial / Protocol feasibility

Brief description of the scenario

This scenario describes if a new clinical trial is feasible to start according to the estimation of recruitment potential. Two versions of this scenario are possible:

1. Based on EHR/PHR/HIS data
2. Based on other data sources

Or a combination of these two; for more details see D1.1 figure 3.5.

Pre-conditions	1) Patient data are stored in the PHR and EHR, or in other database 2) Access to database where patients data are stored
Expected input database types	- PHR and EHR: relational databases, free-text components
Expected input data types	- PHR and EHR: qualitative nominal and ordinal, quantitative discrete and continuous
Expected output	Feasibility (yes/no) plus aggregated statistic data, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML
De-anonymized data needed	Possible
Data-mining components	1) mining of EHR or other available database to select the cohort of patients that fits recruitment criteria best 2) discovery of similar patients based on eligibility criteria 3) discovery of similar trials or similar datasets
Requirements and Methods	1) Data anonymization, or privacy preserving data mining if based on EHR, or alternatively distributed data learning as a possible solution to the ethical, legal and practical problem in transferring data to central repository 2) Similarity learning: discovery of similar patients and similar trials 3) Frequent pattern mining: discovery of frequently occurring patterns to summarize datasets and to aid

	similarity learning 4) Text mining of unstructured databases 5) Time-dependent data mining methods that effectively address the continuously changing landscape in EHR, trial databases, literature, public databases. 6) Change detection to effectively detect changes in data, annotation and structure 7) Multi-dimensional data mining 8) Dimensionality reduction if high-throughput and imaging data needs to be considered
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ representation of semantic similarity in such a complex data space is challenging ➢ Data are not anonymous: privacy issues needs to be addressed, possibly by performing the data-mining at the hospital site

Selection and inclusion of patients into trials

Brief description

This scenario describes how patients can be selected for a specific trial. The initiator could be a pharmaceutical company, a research or a clinical institution. The specific requirements for implementation will change partially depending on the initiator. There is a relation to the scenario KD15 (Personal medical information recommender). The schema is presented in D1.1 figure 3.6.

Pre-conditions	1) Patient data are stored in the PHR and/or EHR, or in other database and accessible 2) Access to imaging and genomic or other laboratory databases where patients data are stored 3) Trial eligibility criteria available
Expected input database types	- PHR and EHR: relational

	databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Genomic and other 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - Trial eligibility criteria: from relational database to small table, free-text components
Expected input data types	- PHR and EHR: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Genomic data: qualitative nominal and ordinal, quantitative discrete and continuous - Eligibility criteria: qualitative nominal and ordinal, quantitative discrete and continuous
Expected output	Patient matched to trial/trials and related statistics, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, email to treating physician
De-anonymized data needed	Possible
Data-mining components	1) Mining EHR or other databases to select the cohort of patients that fits inclusion criteria of the trial 2) Discovery of similar patients based on eligibility criteria 3) Discovery of similar trials or similar datasets
Requirements and Methods	1) Privacy preserving data mining if based on EHR/PHR 2) Distributed data learning as a possible solution to the ethical, legal and practical problem in transferring data to central repository 3) Similarity learning: discovery of similar patients and similar trials 4) Text mining of unstructured databases 5) Frequent pattern mining: discovery of frequently occurring patterns to summarize datasets and to aid

	similarity learning 6) Time-dependent data mining methods that effectively address the continuously changing landscape in EHR, trial databases, literature, public databases. 7) Change detection to effectively detect changes in data, annotation and structure 8) Multi-dimensional data mining, in particularly for genomic and imaging databases if needed to be considered 9) Dimensionality reduction if high-throughput and imaging data needs to be considered
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ representation of semantic similarity in such a complex data space is challenging

Detection and prediction of SAEs and SUSARs

Brief description

This scenario describes how SAEs and SUSARs can be detected and predicted before a treatment is given to a patient (see D1.1, figure 3.7).

Pre-conditions	1) Patient data are stored in the PHR and/or EHR, or in other database and accessible 2) Access to imaging and genomic and other LIMS/laboratory databases where patients data are stored 3) Access to SAE databases and literature
Expected input database types	- PHR and EHR: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components

	- Genomic and other 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - SAE: relational database, document-oriented databases (e.g. XML), free-text components - Existing knowledge and literature: relational databases, document-oriented databases (e.g. XML)
Expected input data types	- PHR and EHR: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Genomic data: qualitative nominal and ordinal, quantitative discrete and continuous - SAE: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data
Expected output	Predicted risk plus summary statistic and related images/plots, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML
De-anonymized data needed	Possible
Data-mining components	1) Data mining in databases of EMA for SAEs and literature mining to find association between a specific profile and SAEs 2) Assign patient to a risk group based on the molecular analysis of the pharmacogenomics in the blood of the patient
Requirements and Methods	1) Frequent pattern mining: discovery of frequently occurring patterns to summarize datasets and to aid similarity learning 2) Similarity learning to discover similar SAEs profiles 3) Classification rules to assign patient to risk group

	4) Text mining of unstructured databases (e.g. part of EHR and literature) 5) Time-dependent data mining methods that effectively address the continuously changing landscape in SAEs, literature and other databases, with a view to extend to data stream mining in the future 6) Change detection to effectively detect changes in data, annotation and structure 7) Multi-dimensional data mining, particularly for imaging and 'omic databases 8) Dimensionality reduction if high-throughput and/or imaging data need to be considered
Possible challenges	➢ Complex structured and unstructured databases ➢ Time dependent data

Pharmaco-vigilance – Automatic reporting of SAEs and SUSARs

Brief description

This scenario describes how SAEs and SUSARs are detected in specific patients and will be reported automatically to regulatory bodies. Details and scenario schema can be found in D1.1, figure 3.8.

Pre-conditions	1) Patient data are stored in the PHR and/or EHR, or in other database and accessible 2) Access to imaging and LIMS/laboratory databases where patients data are stored 3) Physician has validated the patient data 4) Access to SAE EMA databases
Expected input database types	- PHR and EHR: relational

	databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory data: relational databases, document-oriented databases (e.g. XML), flat file databases - SAE: relational database, document-oriented databases (e.g. XML), free-text components
Expected input data types	- PHR and EHR: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - SAE: qualitative nominal and ordinal, quantitative discrete and continuous
Expected output	Data, tables and plots for report compliant to GCP criteria, full log of all operations and version of software used for GCP
Expected output formats	Excel, pdf, power point, HTML
De-anonymized data needed	Possible
Data-mining components	1) At regular time points query databases to find matching events, e.g. HIS/EHR/PHR, for SAEs and SUSARs and generate automatic report.
Requirements and Methods	1) Classification rules to assign patient to risk group 2) Similarity learning to discover similar SAEs profiles 3) Text mining of unstructured databases (e.g. part of EHR) 4) Multi-dimensional data mining, particularly for imaging and 'omic databases 5) Dimensionality reduction if high-throughput and/or imaging data need to be considered
Possible challenges	➢ Complex structured and unstructured databases

Early detection of cancer / individual risk / prevention

Brief description

According to the patient's personal life style data (social networks), his genetic data and clinical data (EHR, PHR, HIS, etc.) the personal risks for diseases can be listed. This might help to detect cancer earlier by starting a screening program for the patient or advice the patient to change his/her lifestyle to prevent cancer, if such a program exists. The scenario is outlined in D1.1, figure 3.9.

Pre-conditions	1) Patient data are stored in the PHR and/or EHR, or in other database and accessible 2) Access to imaging and LIMS/laboratory databases where patients data are store 3) Access to literature and knowledge based databases 4) Access to social network
Expected input database types	- PHR and EHR: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components - Social network: relational, semi-structured, structure-free databases
Expected input data types	- PHR and EHR: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - Social network: qualitative nominal and ordinal, quantitative discrete and continuous, string data - Literature: qualitative nominal and

	ordinal, quantitative discrete and continuous, string data
Expected output	Individual cancer risk and summary statistic data/plots, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, emails to relevant user (e.g. physician)
De-anonymized data needed	Yes
Data-mining components	1) Data Mining of literature to find risks for diseases (Cancer) 2) Data mining of EHR/social networks to describe the lifestyle of a patient 3) Prediction of the individual cancer risk
Requirements and Methods	1) Privacy preserving data mining 2) Text mining of unstructured databases 3) Frequent pattern mining to identify frequent patterns in different datasets 4) Rule discovery to identify and predict risk for disease 5) Classification rules to assign patient to risk group 6) Time-dependent data mining methods that effectively address the continuously changing landscape in literature and other databases, with a view to extend to data stream mining Change detection to effectively detect changes in data, annotation and structure 7) Multi-dimensional data mining, particularly if imaging and genomic data are considered 8) Dimensionality reduction, particularly if imaging and genomic data are considered
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Representation of semantic similarity in such a complex data

	space is challenging ➢ Large amount of time-dependent data, time-dependent data-mining with a view to methods used in data stream mining ➢ Data are not anonymous: privacy issues needs to be addressed. If complex scenarios involving tool such as social networks are executed privacy preserving data-mining methods need to be developed and implemented.

Personal medical information recommender

Brief description

This scenario describes how people can obtain objective information about trials, treatments etc. about their specific disease. It defines the condition of a patient and does data mining in all available data sources (see D1.1, figure 3.10):

Pre-conditions	1) Patient data are stored in the PHR and/or EHR, or in other database and accessible 2) Access to imaging and LIMS/laboratory databases where patients data are stored 3) Access to literature and knowledge based databases 4) Access to clinical trial databases
Expected input database types	- PHR/EHR and clinical trials: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components

Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data
Expected output	Summary of medical data and related medical knowledge, some ranked criteria to provide information, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, email to user (e.g. physician or patient)
De-anonymized data needed	Yes
Data-mining components	1) Data mining of literature and trial databases to identify and extract information about the disease and possible trials 2) Summary analysis of information and reporting 3) Analysis and summarization of patient data
Requirements and Methods	1) Privacy preserving data mining 2) Text mining of unstructured databases 3) Frequent pattern mining to identify frequent patterns in different datasets 4) Association rule mining to identify common patient profiles 5) Classification rules to assign patient to risk group 6) Time-dependent data mining methods that effectively address the continuously changing landscape in SAEs, literature and other databases, with a view to extend to data stream mining 7) Change detection to effectively detect changes in data, annotation and structure 8) Multi-dimensional data mining, particularly if imaging and genomic data are considered 9) Dimensionality reduction if high-

	throughput or imaging data are considered
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Representation of semantic similarity in such a complex data space is challenging ➢ Time dependency for both clinical, knowledge-based and literature databases

Develop or update of guidelines from clinical trial data and literature mining

Brief description

This scenario describes how guidelines can be developed and regularly updated from data mining of clinical trials and literature; details and schema re presented in D1.1, see figure 3.11.

Pre-conditions	1) Patient data are stored in the EHR, or in other database and accessible 2) Access to imaging and LIMS/laboratory databases where patients data are stored 3) Access to literature and knowledge based databases 4) Access to clinical trial databases 5) Access to guideline 6) Guideline of interest selected
Expected input database types	- PHR/EHR and clinical trials: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-

	oriented databases (e.g. XML), flat file databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components
Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data
Expected output	Structure list of guidelines updated items, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, direct update of guideline databases
De-anonymized data needed	No
Data-mining components	1) Data mining in CT/HIS, Literature and trial databases 2) Limit the data-mining to data published after the date of the most recent guideline 3) Generate automatic listing of the updated items
Requirements and Methods	1) Time-dependent data mining methods that effectively address the continuously changing landscape in literature and other databases, with a view to extend to data stream mining 2) Time-restricted data mining, selection of data after a given date filter 3) Text mining of free text databases and literature 4) Association rule mining to find common themes and updates for the selected guideline 5) Frequent pattern mining to understand data structure and aid association rule mining 6) Change detection to effectively detect changes in data, annotation

	and structure 7) Multi-dimensional data mining, particularly in imaging and genomic data are considered 8) Dimensionality reduction if high-throughput or imaging data are considered
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Representation of semantic similarity in such a complex data space is challenging ➢ Time dependency both in terms of changing data structure and content, and in terms of the need of time filters

Data mining of consultations

Brief description

In prospective clinical trials many consultations are performed. A part of the questions of such consultations are repeatedly asked. This scenario generates an automatic answer to questions asked during consultations. Details and schema are in D1.1, Figure 3.12.

Pre-conditions	1) Patient data are stored in the EHR, or in other database and accessible 2) Access to literature and knowledge based databases 3) Access to clinical trial databases 4) Access to consultation
Expected input database types	- PHR/EHR and clinical trials: relational databases, free-text components - Consultation: free-text, semi-structured databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components

Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data - Consultation: free-text, string data
Expected output	Structure and maybe ranked list of answers, Summary statistics regarding data collection and mining, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, possibly direct update of FAQ database
De-anonymized data needed	Yes
Data-mining components	1) Data-mining of structured documentation from consultation 2) Data-mining of free text from consultation or data-mining of structured information extracted from the free text 3) Selection and analysis of answers to the same consultation question 4) Literature and other databases mining to validate results
Requirements and Methods	1) Free text mining, particularly for consultations 2) Association rule mining to find relevant data and profiles matching the consultation 3) Frequent pattern mining to understand data structure and aid association rule mining, and to reveal frequently recurrent consultation questions 4) Time-dependent data mining methods that effectively address the continuously changing landscape in literature and other databases 5) Privacy preserving data mining
Possible challenges	➢ Representation of semantic similarity in such a complex data space is challenging ➢ Text data from consultations can be non-anonymous; this needs to be addressed either by anonymization tools used before the mining or by using privacy

	preserving data-mining techniques

Analyse economic data between different procedures (for funding reasons) compared to outcome and quality of life / data of hospital stays, expected side effects, etc.

Brief description

By joining data from EHR, clinical trials, literature and open databases economic aspects of different procedures (diagnostic and/or therapeutic) can be analysed in respect to outcome and quality of life in an individual patient. The relevant steps for this scenario are described in D1.1, figure 3.13.

Pre-conditions	1) Patient data are stored in the EHR, or in other database and accessible 2) Access to imaging and LIMS/laboratory databases where patients data are stored 3) Access to literature and knowledge based databases 7) Access to economic data and related ranking criteria
Expected input database types	- PHR/EHR and economic data: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components
Expected input data types	- PHR/EHR and economic data: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data

Expected output	Ranked list of procedures according to economic and clinical outcome criteria, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML
De-anonymized data needed	No
Data-mining components	1) Data-mining of literature 2) and data-mining of open source databases: identify best diagnostic and treatment procedures for patients 3) Data mining of economic databases to produce ranking of procedures
Requirements and Methods	1) Text mining of literature and free-text component of databases Association rule mining to discover similar procedures and criteria 2) Frequent pattern mining to aid association rule mining 3) Time-dependent data-mining with a view to methods used in data stream mining, to address rapidly changing data type and structure 4) Classification and ranking analysis to produce ranking of procedures
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Representation of semantic similarity in such a complex data space is challenging ➢ Large amount of time-dependent data, time-dependent data-mining with a view to methods used in data stream mining ➢ Data are not anonymous: privacy issues needs to be addressed. If complex scenarios involving tool such as social networks are executed privacy preserving data-mining methods need to be

	developed and implemented.

Build, optimize, validate or update a diagnostic classifier

Brief description

By analyzing data from literature together with patient data produced using high-throughput assays, imaging, pathology, information from clinical trials and EHR databases, build a diagnostic classifier or update existing ones. The relevant steps for this scenario are described in D1.1 (Diagnostic sarcoma classifier, BR4).

Pre-conditions	1) Patient data are stored in the EHR, or in other database and accessible 2) Access to imaging and genomic and other LIMS/laboratory databases where patients data are stored 3) Access to literature and knowledge based databases 4) Access to clinical trial databases
Expected input database types	- PHR/EHR and clinical trials: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components
Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data - Consultation: free-text, string data

Expected output	Patient diagnostic group with all related statistics, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML
De-anonymized data needed	Yes
Data-mining components	1) Data-mining of literature and open source databases to find data and profiles that match the patient profile 2) Analysis of data from high-throughput assays, imaging, pathology, EHR/PHR and clinical trials to build a diagnostic classifier 3) Classify new patients 4) Validate classification using data from different clinical centres
Requirements and Methods	1) Privacy preserving data-minig or anonymous data 2) Text mining for EHR free-text components and literature data 3) Association rule mining to find profiles and data that match the patient data and to discover similar patient profiles 4) Frequent pattern mining to aid association rule mining 5) Time-dependent data-mining to account both for changes in database structures, information content and changes in techniques/type of data 6) Distributed learning: data from different centres can be merged without having to transfer data to central repository 7) Classification rules to classify new patients once the classifier is developed
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Data are not anonymous: privacy issues needs to be addressed. Distributed learning could provide a solution.

Build and validate predictive models of late toxicity

Brief description

By analyzing data from literature together with patient data including imaging, pathology, information from clinical trials and EHR databases, build a mathematical model of late toxicity or update existing ones. The relevant steps for this scenario are described in D1.1, they are included in the personal medical information recommender scenario.

Pre-conditions	1) Patient data are stored in the EHR, or in other database and accessible 2) Access to imaging and genomic and other LIMS/laboratory databases where patients data are stored 3) Access to literature and knowledge based databases, and information on existing models via either literature or repositories
Expected input database types	- PHR/EHR and clinical trials: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components, formulas
Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data - Models: quantitative data, ordinal data

Expected output	Predicted risk and associated statistics/plots, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, email to user (e.g. physician)
De-anonymized data needed	Yes, but only for patient in care
Data-mining components	1) Data-mining of relevant literature 2) Data-mining of relevant clinical trials 3) Analysis of data from imaging, pathology, EHR/PHR and clinical trials together with long-term follow-up to build models of late toxicity 4) Mining of clinical trials and other databases to find datasets for validation of models 5) Predict late toxicity risk using pre-existing models and classify patients into toxicity groups
Requirements and Methods	1) Association rule mining to discover existing models and toxicity profiles 2) Frequent pattern mining to aid association rule mining Time-dependent data-mining accounting both for information and data structural changes 3) Distributed learning: data from different centres can be merged without having to transfer data to central repository in the learning/refinement phase of the model building if this is included in the scenario 4) Classification rules for new patients to be assigned to risk group
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Data are not anonymous: privacy issues needs to be addressed. Distributed learning could provide an alternative solution to

	anonymization and storage of data in a central repository.

2.1.2.2 Data curation

Long-term follow-up

Brief description

This scenario deals with the curation of data in long-term follow-up including also survival follow-up, primary and secondary outcome measures follow-up, safety reporting for adverse reactions are study treatment completion. The relevant steps for this scenario are provided in D1.1, figure 3.14.

Pre-conditions	1) Patient data are stored in the EHR, or in other database and accessible 2) Access to imaging and genomic and other LIMS/laboratory databases where patients data are stored 4) Access to national registries 5) Access and permission to update clinical trials databases 6) Patient has been selected from clinical trial database
Expected input database types	- PHR/EHR and clinical trials: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases
Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous
Expected output	Follow-up items retrieved, log of operations and version of software used
Expected output formats	direct update of clinical trial database, brief HTML report

De-anonymized data needed	Yes, but only for the patient in care and enrolled in the clinical trial
Data-mining components	1) Data mining of HIS/EHR/PHR to find new data about the selected patient 2) Data mining of National registries to find data about the selected patient
Requirements and Methods	1) Association rule mining if identifiers are not identical between databases and match between different ID systems is required 2) Text mining for part of EHR and maybe national registry databases 3) Time-dependent data mining 4) Privacy preserving data mining
Possible challenges	➢ complex structured and non-structured data: both patient and trial data can include EHR/PHR data, imaging data, pathology data, genomic and genetic data ➢ Data are not anonymous: privacy issues needs to be addressed. Distributed learning could provide an alternative solution to anonymization and storage of data in a central repository.

Patient Diary

Brief description

This scenario deals with the possibilities of a patient diary. Such a diary can be used in clinical trials; the details are provided in D1.1, figure 3.15.

No specific data-mining component specified

2.1.2.3 Basic research and clinical trials support

Supporting design of new trials and hypothesis generation

Brief description

Clinical trials are often generated to test a research question; analysing all available data from previous trials, guidelines, literature and others, can support the generation of research hypotheses such as to identify potential biomarkers. Some of the key steps involved in this scenario are identified in D1.1, figure 3.16.

Pre-conditions	1) Access to clinical trial databases 2) Access to imaging and genomic and other LIMS/laboratory databases related to the clinical trials 3) Access to literature and other public databases
Expected input database types	- clinical trials: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases - Literature and knowledge-based databases: relational database, document-oriented databases (e.g. XML), free-text components, formulas
Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous - Literature: qualitative nominal and ordinal, quantitative discrete and continuous, string data
Expected output	Hypothesis generation with supporting documentation, data and summary statistics, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, other reporting system, email to trial chair
De-anonymized data needed	No
Data-mining components	1) Data mining of clinical trial

	databases 2) Data-mining of literature 3) Data-mining of other databases
Requirements and Methods	1) Similarity learning to find similar clinical trials and/or datasets 2) Association rule mining to find common profiles and data features 3) Frequent pattern leaning to aid association rule mining 4) Rule discovery to find new data features/structures/links 5) Text mining of literature 6) Time-dependent data-mining to address change in data and databases structure 7) Multi-dimensional data mining as imaging and genomic data are key in this scenario 8) Dimensionality reduction as data from high-throughput assays and imaging are likely to be considered
Possible challenges	➢ complex structured and non-structured data: data can include EHR/PHR-like data, imaging data, pathology data, genomic and genetic data ➢ Large amount of time-dependent data, time-dependent data-mining with a view to methods used in data stream mining

Clinical data reuse

Brief description

Re-use the clinical data into the trial eCRF systems to avoid double data entry.

No specific data-mining component specified

Opt-out solution for further research

Brief description

Provide a platform where patients can select which research they do not like to do with their data or biomaterial. Such a scenario is based on the fact that every patient agrees to share his data to any research project and that he can disagree to specific research projects at any time by using the above described website. The relevant steps for this scenario are described in D1.1, figure 3.17.

Pre-conditions	1) Access to anonymized EHR/PHR databases information 4) Access to imaging and genomic and other LIMS/laboratory databases related to the clinical trials 5) Research question available for patient selection
Expected input database types	- EHR/PHR: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data: relational databases, document-oriented databases (e.g. XML), flat file databases
Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous
Expected output	Patient matched to research questions with relevant statistics and supporting documentation, data and summary statistics, log of operations and version of software used
Expected output formats	Excel, pdf, power point, HTML, other reporting system, email to patient
De-anonymized data needed	Yes, but only after the patient consented to study and would like to enrol, so after data-mining
Data-mining components	Data-mining on anonymized data to select patients to answer specific research question
Requirements and Methods	Methods will depend on specific realization of this scenario and research questions. Possible methods involve:

	1) Similarity learning to find patients matching the research profile 2) Text mining of EHR 3) Time-dependent data-mining to address change in data and databases structure 4) Multi-dimensional data mining if imaging and genomic data are part of this scenario 5) Dimensionality reduction as data from high-throughput assays and imaging might be considered
Possible challenges	➢ complex structured and non-structured data: data can include EHR/PHR-like data, imaging data, pathology data, genomic and genetic data ➢ Large amount of time-dependent data, time-dependent data-mining with a view to methods used in data stream mining ➢ Maintenance and design of a website where the patient has access to disagree with the research project

Similarity of datasets to combine

Brief description

Detection and identification of similar datasets to use in meta-analyses or combined analyse. This scenario is further described in D1.1 but it is very general at the moment. It can be seen as a sub-task or use case of other scenarios as the technique described it is required by some of the other scenarios.

Pre-conditions	Access to relevant anonymized EHR/PHR or other relevant databases information
Expected input database types	- EHR/PHR: relational databases, free-text components - Imaging: relational databases, document-oriented databases, free-text components - Laboratory and 'omic data:

	relational databases, document-oriented databases (e.g. XML), flat file databases
Expected input data types	- PHR/EHR and clinical trials: qualitative nominal and ordinal, quantitative discrete and continuous, string data - Imaging data: DICOM - Laboratory data: qualitative nominal and ordinal, quantitative discrete and continuous
Expected output	Similar datasets selected and ranked based on similarity metrics, related statistics provided, log of all operations and software versions
Expected output formats	Excel, pdf, power point, HTML, other reporting system
De-anonymized data needed	No
Data-mining components	Data mining of CT databases, EHR, literature and other databases to find similar datasets that can be combined
Requirements and Methods	Methods will depend on specific realization of this scenarios. Possible methods involve: 1) Similarity learning to find similar datasets 2) Frequent pattern leaning to aid similarity learning 3) Text mining of EHR 4) Time-dependent data-mining to address change in data and databases structure 5) Multi-dimensional data mining if imaging and genomic data are part of this scenario 6) Dimensionality reduction as data from high-throughput assays and imaging might be considered
Possible challenges	➢ complex structured and non-structured data: data can include EHR/PHR-like data, imaging data, pathology data, genomic and genetic data ➢ Large amount of time-dependent data, time-dependent data-mining

	with a view to methods used in data stream mining

2.1.3 Summary

The above analysis of all scenarios identified some key areas that need to be considered for EURECA data-mining. These are still general as the specific implementation or methods will need to be tailored to the specific scenarios, also in accordance with the EURECA environment and with the choices for semantic interoperability. Table 2.1 summarize the results so far for all scenarios described either in the previous paragraph of this section, or in the last section of this document (EURECA knowledge discovery) or in WP2 deliverable 2.1.

		Data mining requirements											
	Scenario	**Similarity learning**	**Frequent pattern mining**	**Association rule discovery**	**Classification**	**Text mining**	**Time-dependent data mining**	**Change detection**	**Multi-dimensional data-mining**	**Dimensionality reduction**	**Data stream mining**	**Privacy preserving data mining**	**Distributed data learning**
Knowledge discovery	Selection of best trials for a patient	x	x			x	X	x	x	x			
	Trial / protocol feasibility	x	x			x	X	x	x	x		x	x
	Selection and inclusion of patients into trials	x	x			x	x	x	x	x		x	x
	Detection and prediction of SAEs / SUSARs		x	x	x	x	x	x	x	x	x		
	Pharmaco-vigilance				x	x							
	Early detection and prevention		x	x	x	x	x	x	x		x	x	

	of diseases												
	Personal medical information recommender			x		x	x		x	x	x		
	Develop or update guidelines for diseases		x	x		x	x		x	x	x		
	Data mining of consultations		x	x		x	x					x	
	Analyse economic data between different procedures / approaches		x	x		x	x				x		
	Build, optimize, validate or update a diagnostic classifier		x	x	x		x		x	x		x	x
	Build predictive models of late morbidity		x	x	x	x	x						x
Data curation	Long term follow-up					x	x					x	
	Patient diary (connection between PHR and data management tools)												
Basic research and clinical trials support	Supporting design of new trials and hypothesis generation	x	x	x		x	x		x	x			
	Mining of information from EHR and CT to validate research/clinical hypotheses	x	x	x		x	x		x	x			
	Clinical data reuse												
	Opt-out solution for new research					x							

	Similarity of datasets to combine	x	x			x	x		x	x			
	Rapid learning			x		x	x		x	x	x		
	High-level genomic meta-analyses in genomic/genetic		x	x		x	x		x	x			
	Association studies		x	x		x	x		x	x			

2.2 User needs based requirements

At the start of the project a survey was performed that was aimed at gathering opinions about the projects, the possible applications, and about the methods and the datasets which could be available.

2.2.1 Results from user need questionnaires

The description of the survey and an in-depth analysis of results can be found in a previous document D1.1 “User needs and specifications for the EURECA environment and software services”.

In brief five main points were highlighted there:

1. The described clinical scenarios are important as they cover the needs of the participants of the survey.

2. Tools need to be developed build out of use cases as open source tools and stored in the VPH toolkit.

3. Legal issues need to be solved to share data if the data producer wants to share and not only use EURECA tools to work on his own data.

4. Some of the answers highlighted that data should not leave the centre, thus distributed data analysis/mining should be considered

5. Need for standardization

2.2.2 Implication for data-mining requirements

The implications of the above points for data-mining are:

1. The need to base the specific requirements on the scenarios

2. The need to develop open source tools

3. The need to investigate and test privacy preserving data mining

4. The need to investigate distributed learning

5. Modular approach needed

3 Data-mining approaches and methods

The previous sections analyzed the scenarios and user need questionnaire to specify the main components and requirements for the data-mining.

Below we will define and describe the main data mining areas mentioned in the previous section. Their application indication in the context of EURECA is illustrated in section 2.1.2 and summarized in table 2.1.3.

The one provided here is an extremely brief overview and it is provided here exclusively with the purpose of aiding the interpretation of the previous requirement analysis and related discussion provided in section 2.1.2.

A state-of-art review of existing data-mining methods, with in dept analysis of existing application and current limitations, is out of the scope of this deliverable and but will be performed in deliverable D5.2. Please see deliverable D5.2 with respect for a state-of-art review.

3.1.1 Data mining techniques in EURECA

Data pre-processing: dimensionality reduction

New clinical trials include techniques such as high-throughput assays and imaging techniques which produce a very large amount of data points/variables. Thus, data pre-processing has become a very important step in data analysis.

Many of the variables (e.g. gene) in these datasets are correlated and not independent. Methods such as dimensionality reduction help to reduce the number of such variables to some smaller set of independent variables. This helps gaining statistical power in analyses where usually the number of variables is much higher than the number of cases.

Dimensionality reduction approaches can be applied before the analysis or, for example in classification problems, whilst building the classifier.

Similarity Learning

Similarity Learning consists of classification on pairwise similarities. In contrast to traditional machine learning, similarity learning does not assume that objects are well represented in a Euclidean feature space. This is useful for problems in bioinformatics, information retrieval and many other areas with diverse object representations.

In EURECA for example we want to find similar clinical trials. Since clinical trials more frequently now include pathology, genomic and imaging data, the representation of semantic similarity will need to be defined in extremely complex data space.

Association Rule Discovery

Association rule mining considers the problem of discovering association rules between items in a large databases; it has been applied extensively for example to databases of sales transactions but less so to the clinical or medical sciences.

Algorithms have been proposed and tested for categorical data and less so for numerical data (1-5).

Classification

Classification is a mining technique based on machine learning; it is used to classify each item in a set of data into one of predefined sets of classes or groups.

The data classification process involves a learning phase and classification phase. In the learning phase a set of training data are analyzed by classification algorithm; then in the classification phase data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Methods applied to classification problems vary from linear models, to decision trees, Bayesian Classification, Neural Networks, and Support Vector Machines (SVM). Depending on the specific realization of the EURECA scenarios different classification methods will be used.

Clustering

Clustering is a data mining technique that defines groups of objects that have similar characteristic. Contrarily to classification where objects are assigned into predefined classes, clustering both defines the classes and assigns objects to them.

By using clustering techniques we can identify particular regions in object space and can discover overall distribution pattern and the correlations among data attributes. Classification approach can also be used for identifying groups or classes of objects but it becomes computationally costly.

Types of clustering methods include partitioning methods, hierarchical agglomerative methods, density based methods, grid-based methods, model-based methods.

Prediction

Prediction is a data mining technique that attempt to discover the relationship between independent variables and relationship between dependent and independent variables. Regression techniques can be applied to prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict.

Frequent pattern mining

Frequent pattern mining is a core data mining technique that focuses on identifying and extracting frequently occurring patterns from different types of datasets, including unstructured ones. By doing this frequent pattern mining can summarize effectively very complex datasets.

Frequent pattern mining is also used as a tool in combination with other data-mining techniques such as association rule mining and classification.

Text mining

Text mining is a branch of data mining that refers to learning by using automatic extraction of information from free text. Information from different text documents and/or resources is extracted and then linked to generate new rules or hypotheses. These are typically organized and explored with other data-mining methods.

In text mining the data patterns are extracted from natural language text rather than from structured databases; such automated processing of natural language is challenging and methods to perform such a task are still limited. Typically, it requires dividing text mining in specific relatively small tasks that can be performed automatically.

An example of application to genomics is the study of co-occurrences of words in publications to infer related function of genes or proteins, or generate hypotheses that can then be tested in further studies.

Time-dependent and data stream mining

In many modern scientific and medical research domains, new knowledge and data are stored and recorded in large data streams of transactional data that rapidly and continuously grow over time. Not all scenarios described above present all the data stream characteristics but several of them present the challenge of being data that change over time and a large amount of data to be processed.

These types of data require a different data mining approach with respect to the ones used for classical static databases; both as the data change in time but also because the dimension of the data does not allow the classical re-sampling and training approaches often used in the machine learning and data mining community. For example, several classification algorithms require a recursive processing of the data.

Methods for analysis of such data have been described and a recent collection has been published, edited by Aggarwal, which describes many of the advances in the area (6).

3.1.2 Modularity and standardization

One key aspect central to EURECA is a modular approach where tools can be developed for specific scenarios but then be easily re-used and extended to different or larger scale scenarios.

In this context, the aim is to build upon several previously defined concepts and methodologies whenever this is possible and point out areas for further methodological work.

One area where we would like to re-utilize previous concepts is the high-level generation and handling of workflows.

For example, we will take on the concept of data-mining process patterns previously described in the p-medicine project (deliverable 11.1) and we will bring this concept to the EURECA application domain.

Briefly, the concept of data mining process patterns extend the classical data mining workflows to a more general process description which also include a representation of the manual work that needs to be done in order to successfully apply a workflow to a specific problem.

3.1.3 Distributed data learning

For some of the EURECA scenarios, and also as one of the requirements in the survey, the need was highlighted for analyzing the data locally at the hospital site and then merging results or further mine external public databases.

The field of distributed data mining studies and addresses the problem of analyzing data residing at different locations, or nodes, without necessarily having to collect them at a central site, for a recent review see (7). In this case, the challenge is to solve learning problems with minimal exchange of information between nodes; thus the algorithms need to use and summarize the limited amount of exchanged information very efficiently.

Several methods have been applied to distribute data mining including machine learning methods and Bayesian Network Learning (6, 8) In the last section of this document we will discuss specific scenarios where a distributed data approach mining is chosen and we will discuss specific methods. Furthermore, we will focus on the main different solutions available and their applicability to EURECA in a following state-of-art review on data-mining methods.

3.1.4 Privacy preserving data-mining

One of EURECA key goals is to deliver an environment that fulfils the data protection and security needs and the legal, ethical and regulatory requirements related to linking research and EHR data. In addition, several conflicting interests of different stakeholders must be taken into account to ensure the practicability of data mining solutions. The main problem are conflicting interests on which information should be protected, and which information should be freely available, in particular when considering information that should be made public outside of the EURECA

contractual framework, e.g. in the form of scientific publications or open models for decision support.

The most relevant of the aforementioned stakeholders and interests are

- Patients require their personal information to be kept private. On the other hand, many patients are willing to freely share at least part of their information on public websites such as Facebook or Patientslikeme. This can become problematic when considering background knowledge attacks on released information. In addition, patients may also profit from their information being released, as new treatments can be found. Hence, it is not clear where to best draw the line between privacy and data publication.
- Trial chairmen (and the general public) are ultimately interested in finding new scientific knowledge that may help their patients, and publishing it in scientific journals. Rules of good scientific practice may require them to make available a large part of the data that is the basis of their findings.
- Data protection officers are responsible for following all applicable legal and regulatory requirements.
- Hospital and pharmaceutical companies, being economic entities, also have to consider patient data as an economic asset, and may not be willing to freely share data with other entities, as long as questions of intellectual property rights and financial compensation have not been cleared.

Vast research on methods for privacy-preserving data mining exists. The main directions of privacy-preserving data mining can be described as follows:

Privacy-preserving data publishing deals with the question of releasing data in such a way that all sensitive information is removed. The released data can then be processed with standard data mining methods. Approaches include randomization, k-anonymity (9), l-diversity (10) and t-closeness (11).
In EURECA, however, this approach does not seem promising. The problem is that once relevant information is removed, it cannot be recovered. Hence, it may be that clinically important information cannot be found, which clearly violates the interests of trial chairmen and the general public. The key point is here that under the applicable privacy law it is still possible to go back to patients and ask them for additional consent for their data to be processed and made public once it is known that an important discovery is made.

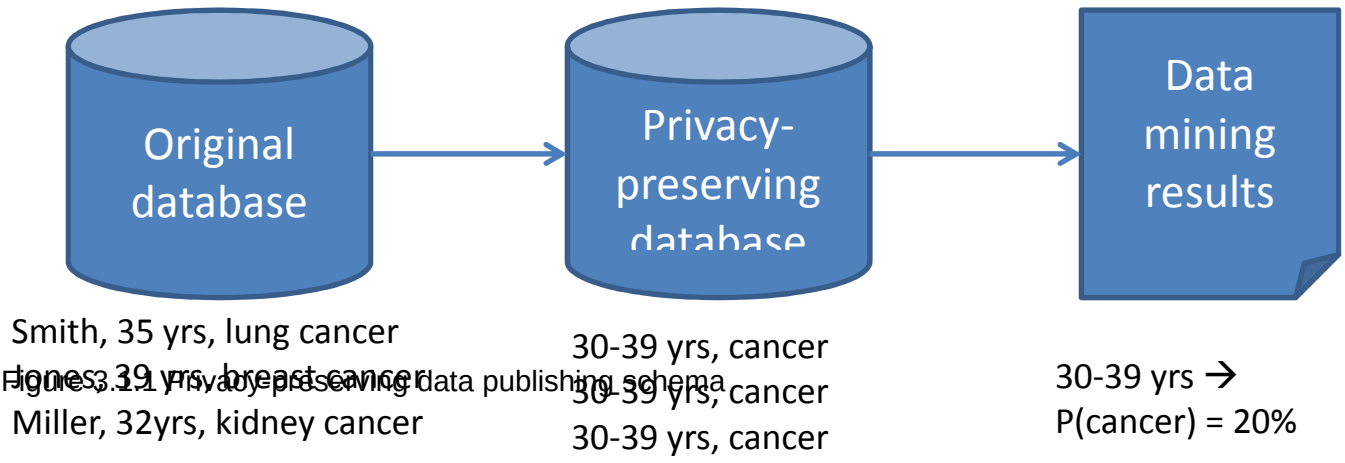


Figure 3.1.1 Privacy-preserving data publishing schema

Secure distributed computation removes the need for a central database in which data mining takes place. Instead, models are computed in a distributed way, such that it is guaranteed that the content of the individual, distributed databases are kept private except for what can be learned from the global results. Several approaches, based on techniques like the solution of the Millionaire's problem or the secure sum algorithm exist.

In EURECA, this approach will be interesting because it matches the EURECA distributed architecture (seeing hospitals as the local databases). In particular, it will help to cover the interests of hospitals and pharmaceutical companies, which most likely will object to the transfer of their data to a central database. The implementation of this approach consists of putting a computational endpoint at every hospital, which execute the local data mining queries and gives information back according to the privacy-preserving protocol to a central instance which coordinated the computation and assembles the final result.

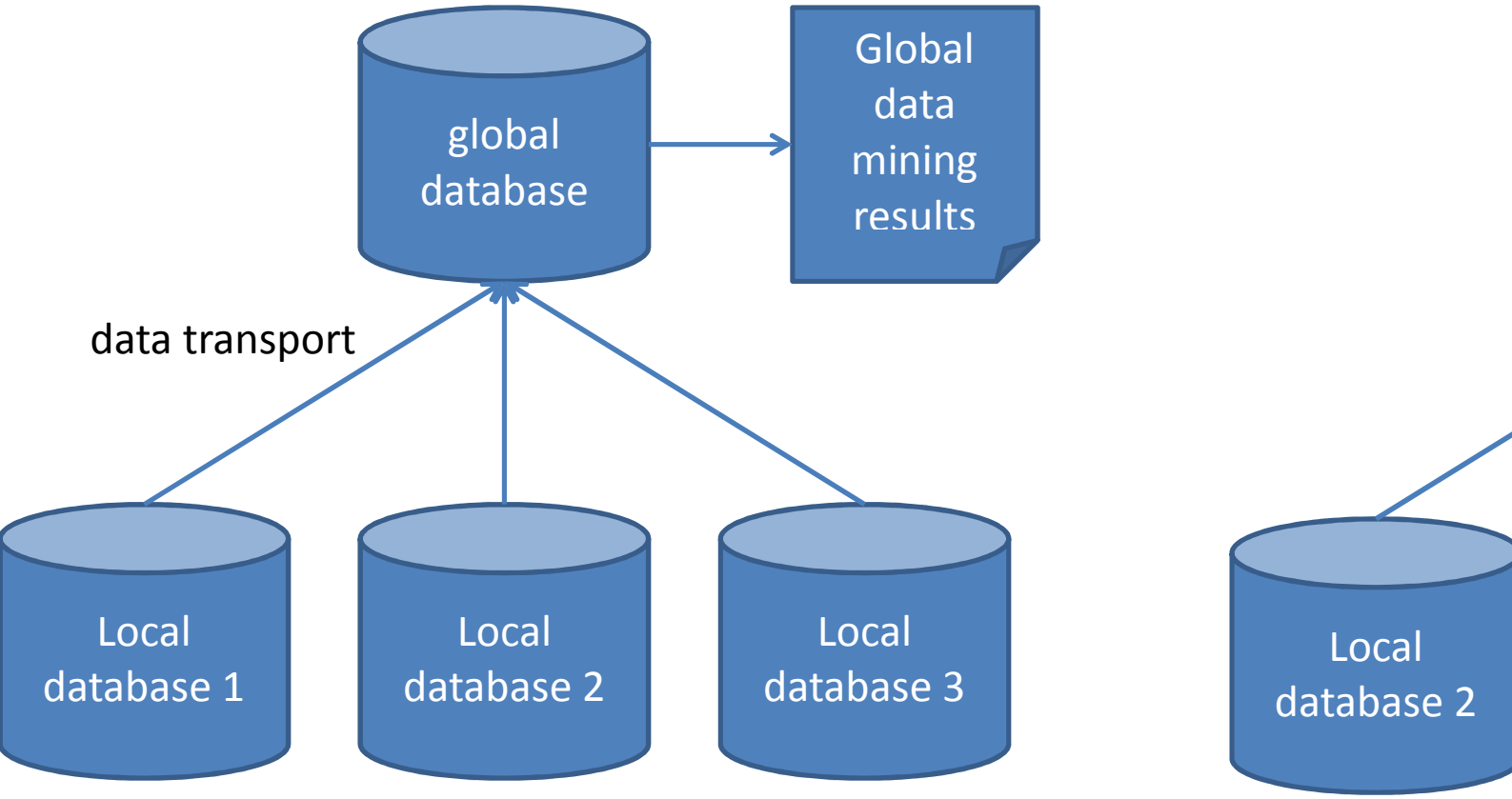


Figure 3.1.2 Secure distributed computation schema

Privacy-preserving model publishing deals with the question of quantifying the information that is contained in a released piece of knowledge such as a statistic, or a data mining model such as a complete decision tree (12-14).

An example of this scenario could be a neighbour, who knows that the patient regularly goes to the university hospital of the city to get treated for cancer. The neighbour might then look up running clinical studies of the university hospital on the internet, look for the latest publications of these studies, and try to see if he can use the published results to infer knowledge about the patient from his public data.

For EURECA, this is a central problem, as it addresses the problem of exporting valuable knowledge outside of the contractual framework. The problem is that while collaborating parties (such as the project consortium) can be bound by contracts to observe all legal and ethical constraints, the ultimate goal of medical research is to generate new medical knowledge and make it available to the general public. Hence, it must be possible to weigh the scientific importance of a new discovery against the privacy implications, so that an adequate decision can be made.

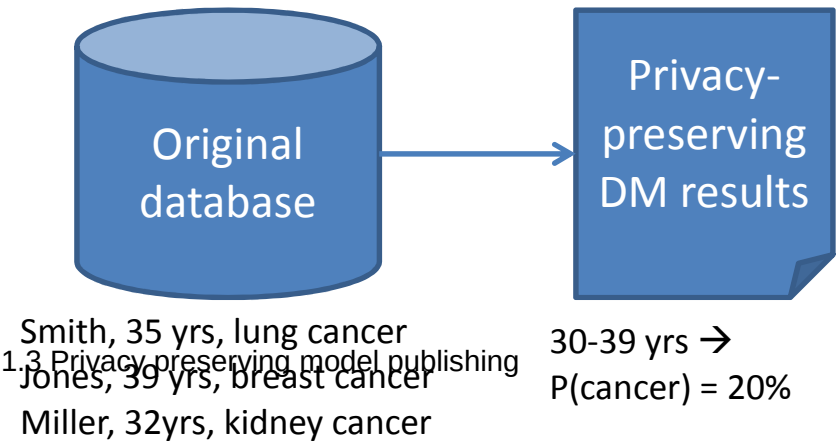


Figure 3.1.3 Privacy preserving model publishing

To give a closer idea of privacy preserving data-mining, a recent example that has been successfully applied to this area is micro-clustering, for a review see e.g. (6). Micro-clustering is a technique based on condensed representation of the data which show k-anonymity; these representations act as pseudo-points that can be used in the data-mining as a surrogate of the actual points. These and other methods will be reviewed in a following deliverable focusing on a state-of-the art review of current data-mining methods which could find an application in EURECA.

In the context of EURECA the privacy-preserving requirements are different depending on the specific application. In D1.1, chapter 4, the legal and privacy requirements for the EURECA scenarios have been discussed. In this respect, the following categorization was adopted:

- **Research domain**

  In the research domain, in most cases the data can be anonymized before the data-mining occurs, thus privacy preserving data-mining techniques are not required and there will be tools in EURECA to guarantee anonymization is compliant with European regulations.

  The following scenarios were classified as research domains:

- Develop or update guidelines for diseases
- Data mining of consultations
- Analyse economic data between different procedures / approaches
- Opt-out solution for new research

- **Care domain**

  In the care domain personal data are always needed; thus data-mining algorithms applied to this area will need to address the privacy problem either by applying specific privacy-preserving data-mining technique or by working in combination with other EURECA tools and algorithms to create privacy-preserving workflows.

  The following scenarios were classified care domain:

  - Detection and prediction of SAEs / SUSARs
  - Early detection and prevention of diseases
  - Personal medical information recommender
  - Long term follow-up
  - Patient diary (connection between PHR and data management tools)
  - Clinical data reuse

- **Trial support and execution**

  In the trial support and execution setting the situation can be more complex, and some scenarios do need access to personal data, others do not.

  The following scenarios were classified as trial support and execution:

  - Selection of best trials for a patient
  - Trial / protocol feasibility
  - Selection and inclusion of patients into trials
  - Pharmaco-vigilance – Automatic reporting of SAEs and SUSARs
  - Supporting design of new trials and hypothesis generation
  - Similarity of datasets to combine
  - Rapid learning

This “legal” categorization is maintained throughout this document; whilst a more detailed discussion on specific data-mining requirements and related available algorithms will be provided in a following document.

The main goal of research of privacy-preserving data mining in EURECA will be to find an approach with strict guarantees on data privacy that addresses the practical needs in the scenarios that are described in this document. A main criterion will be to find clear, understandable descriptions of privacy problems (e.g. which information of which patient might be at risk, or which part of a set of information to be published is problematic) such that a productive discussion between all relevant stakeholders is possible, and privacy problems can be easily detected and solved.

4 EURECA knowledge discovery: focus on initial data mining components

In this section we focus on some scenarios with a large data mining component; these scenarios are described here and will be analyzed further. The requirement analysis above indicated several needs which further specify the needs for the EURECA environment. As not every tool should be built from scratch it is important to dissect the scenarios into use cases of highest granularity. This approach will help to build tools in a modular way and reuse granular tools in different scenarios. This will be described in detail in D1.2 (Definition of relevant user scenarios based on input from the users) and further documents. Please refer to future deliverable D1.2 for further evolution of scenarios and related flow.

4.1 Secure distributed data mining

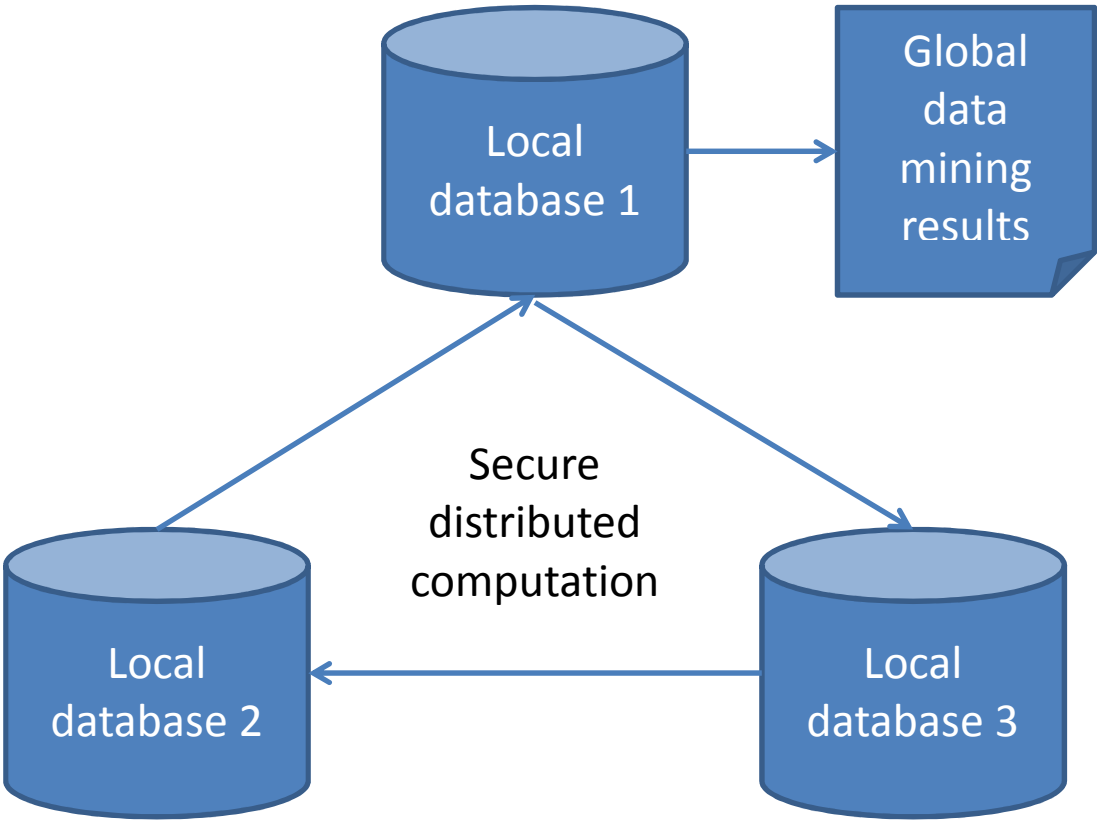
Scenario

As an example of privacy-preserving data mining, consider a group of hospitals, all with a local database. We assume that all local databases are accessible under a common data schema, and that they all contain information about patients with a certain condition, e.g. patients for which a certain treatment has led to the development of a severe adverse event.

In a realistic setting, hospitals might be willing to cooperate to find an explanation for the occurrence of the severe adverse event; however, as the local databases contain much more information about the patients and their treatment, they might not be willing to export their data to a central database.

Contextualization and brief state of art

In this case, secure distributed data mining can be applied to securely compute all rules in the union of all local databases that are significant to describe the occurrence of the severe adverse event in question, while keeping all other information hidden from the participating parties. In particular, secure distributed rule discovery (15) is of interest to EURECA, on the one hand because decision rules are valuable for the use in Clinical Decision Support systems, on the other hand because understandable rules make it easier for the clinician or bioinformatician to check whether the discovered patterns are interesting, and hence to act upon the discovered knowledge. Note that only the global, securely computed result is to be accessed by the investigating scientist, the underlying database and the individual patient data in them are hidden. Hence, the investigating scientist has much less information about the underlying data than in the standard case, which means that he/she might need to access additional sources of information – scientific publications, general medical knowledge – to interpret and evaluate the data mining results, and might also contact the local database owners for additional information. In this case, an understandable representation of the results that can be communicated to other parties might be necessary.



The basic idea of the approach is that using the well-known secure sum algorithm, sums of different values can be computed securely over the local databases, which allows to securely executing relevant queries, and build a rule mining algorithm on top of it. It is then guaranteed that the output of the algorithm is identical to the output of the standard algorithm applied on a single global database.

The final choice of the algorithms to be implemented in a secure distributed fashion, however, will depend on the performance of several data mining approaches in a non-distributed prior study.

Application indication

Privacy-preserving data mining is most relevant in those EURECA scenarios where very detailed, privacy-sensitive data from possibly many hospitals needs to be combined for knowledge discovery. Furthermore, distributed data mining can be a solution when complex data mining results need to be exported to third parties (like pharmaceutical companies or regulatory bodies) or the general public (like scientific publications or open source decision support modules). In particular, these are the following knowledge discovery scenarios:

- Detection and prediction of SAEs and SUSARs
- Pharmaco-vigilance
- Guidelines development
- Data mining on consultation data
- Diagnostic sarcoma classifier
- Analyse economic data between different procedures

4.2 Rapid learning

Scenario

In rapid learning research we learn outcome prediction models from clinical data using machine learning (see Figure 4.2.1 below). Typically outcomes are tumour related such as survival, tumour progression, local control, pathological complete response, distant metastasis but also toxicity related outcomes such as quality of life, cosmetic result, shoulder movement, heart toxicity etc. Typically a model consists of prognostic and predictive features extracted from the clinical data, and predicts an outcome for a certain treatment choice. By comparing predicted outcomes for different treatments, an optimal treatment can be chosen. The exact input, treatment and outcome features are very disease dependent. Examples of models for lung, head and neck and lung cancer can be found at www.predictcancer.org.

Contextualization and brief state of art

Several examples of rapid learning and related methodologies have been discussed in the context of Oncology care and research (16). The basic requirement for machine learning is that a machine (and thus not a human) can learn from the data. This means that a machine should not only be able to read the data (syntactic interoperability) but also that the machine should be able to understand the data (semantic interoperability). A further requirement is that there should be lots of data to learn from as the accuracy of the model is correlated to the amount of data the machine has seen (a machine learns differently than a human). This means that data from many hospitals needs to be available for learning. In conclusion, rapid learning needs large, semantic and syntactic interoperable datasets from many different hospitals.

It is not a requirement for rapid learning to have the data in one place. Distributed learning approaches are available in which the learning applications work locally on the data and publish the knowledge they have extracted from that dataset. By combining local knowledge a global model can be learned that has used all the data to learn from, without data leaving the hospital.

To fulfil the requirements stated above, a number of tools are proposed

- An ETL (Extract Transform Load) tool that connects to clinical data sources, performs the semantic and syntactic transformation into a common data model and stores the data according to this data model.
- An API that allows access to the stored data including the possibility to query and retrieve the data.
- A standardized application environment including machine learning libraries in which a specific rapid learning application can be deployed.
- A tool that allows access to the outside world to publish the knowledge produces by the rapid learning application.

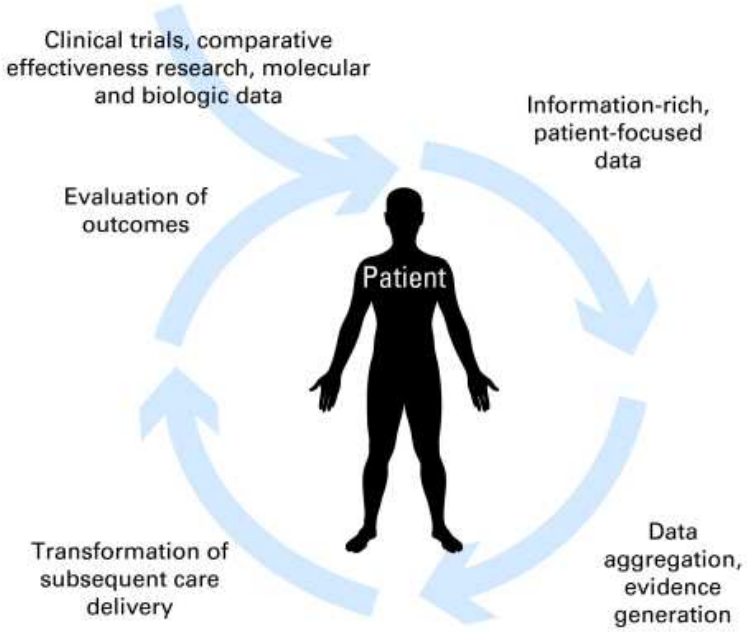


Figure 4.2.1. A representation of Rapid Learning from (16)

Application indication

Rapid learning approach will constitute a scenario on its own (see list in the above section; “Basic research and clinical trials support” scenarios) but will also be relevant for the following “knowledge discovery” EURECA scenarios:

- Develop or update guidelines for diseases
- Build, optimize, validate or update a diagnostic classifier (e.g. scenario BR4, Sarcoma diagnostic classifier)
- Build predictive models of late morbidity

And the “Basic research and clinical trials support” EURECA scenarios:

- Supporting design of new trials and hypothesis generation (see also section below)
- Mining of information from EHR and CT to validate research/clinical hypotheses
- Clinical data reuse

4.3 Supporting design of new trials and hypothesis generation, and high-level genomic meta-analyses

Scenario

A clinical trial often starts from the formulation of a clinical research hypothesis. Such a hypothesis generation process is usually a consequence of analysing available data from previous trials, guidelines, literature and other sources. It can also help to find information related to biomarkers that are relevant for the disease and can be tested during the trial execution (for a review in the area see (17)).

In this context a very much related scenario is the high-level genomic meta-analysis scenario (Table 2.1). High-throughput genomic technologies have brought a revolution in molecular biology, and have enabled scientists to study biological systems at a different scale. However, this technology has not yet delivered the expected benefit to the clinical setting and patient management by producing validated effective biomarkers.

It is now recognized that integrating information of multiple relevant genomic studies in meta-analyses could help this translation process; however this is often challenging due to heterogeneity of data annotation and processing and the integration has been limited to a small amount of genomic and clinical data.

Several tools have been recently suggested that address in part the heterogeneity due to different platforms and different data processing algorithms/workflows, see e.g. (18, 19); however the complexity and potential of a large scale integration of biological and clinical information has not yet been addressed. This does not only require integration of data but also integration of abilities and knowledge domains which often resides with different individuals.

For this reason we introduce here the concept of high-level (distributed) genomic meta-analyses. These can be seen as an extension of existing meta-analysis and knowledge-based analysis approaches, where information from clinical databases, biological data and the drug and disease knowledge-base are fully integrated in a collaborative analysis that combines data but also human intervention from domain experts.

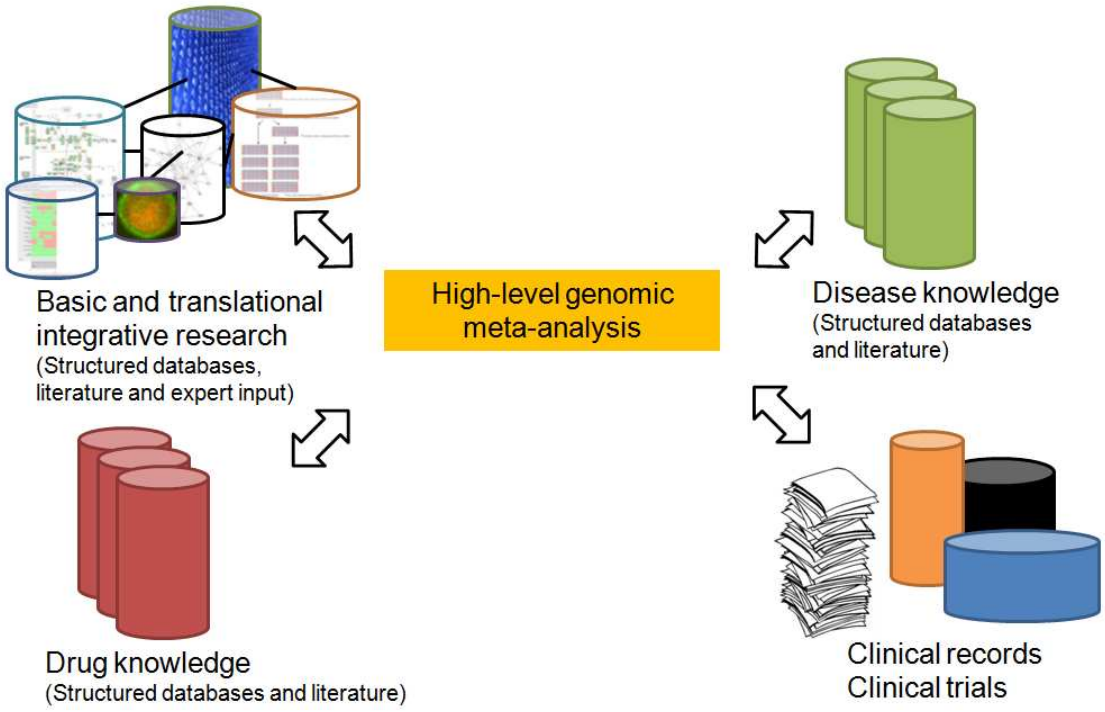


Figure 4.3.1 High-level genomic meta-analysis example schema

Contextualization and brief state of art

Methods such as the above mentioned rapid learning could aid this high-level genomic approach; in fact most of the data mining requirements and tools needed are similar (see previous section, suggested tools list). Furthermore, due to the fast accumulation of scientific, clinical knowledge, and an interactive expertise component, this scenario would benefit from some of the methods used in data stream mining, where mining is performed over continuously and rapidly changing streams of data (see Section 3.1.1, Data stream mining).

One of the basic differences between the present approach and the Rapid Learning approach as described above is that a machine (and thus not a human) can learn from the data but also from the human (expert) intervention, where the latter can be seen as personal communication or actions performed on a machine. This means that a machine should still be able to read the data (syntactic interoperability) and to understand the data (semantic interoperability); but also to rapidly incorporate and integrate knowledge and interpretation from multiple users provided both as structured information and/or for example as free-text.

The point of view of the present approach and related scenario is the use of clinical and biological information, together with human expertise, for both hypothesis generation for future clinical trials and patient management. Distributed data mining is often necessary in such a context both for practical limitation to transferring large amount of data and for privacy preserving reasons (see Section 4.1).

Application indication

This approach will constitute a scenario on its own but it will also be relevant for the following “knowledge discovery” EURECA scenarios:

- Develop or update guidelines for diseases
- Build, optimize, validate or update a diagnostic classifier (e.g. scenario BR4, Sarcoma diagnostic classifier)

And the “Basic research and clinical trials support” EURECA scenarios:

- Supporting design of new trials and hypothesis generation (see also section below)
- Mining of information from EHR and CT to validate research/clinical hypotheses
- Clinical data reuse

5 SUMMARY

The EURECA environment and the tools to be developed are based on the needs gathered by the questionnaire and scenarios built in WP1. The present requirement analysis was based on the scenarios and questionnaire results described in deliverable D1.1.

Specifically, we have dissected each one of the scenarios and extracted the components relevant for the data mining and highlighted the challenges; this will guarantee that the tools implemented or developed for data mining and the environment will support the user needs.

We have also briefly described the techniques that are available in the area of application; however, a further document will review these methodologies in depth, and describe the needs for validation and development.

6 REFERENCES

1. Agarwal R, Aggarwal C, Prasad V. A tree projection algorithm for generation of frequent itemsets. High Performance Data Mining Workshop; 1999; Puerto Rico; 1999.
2. Aggarwal C, Wof J, Yu P. A new method for similarity indexing for market data. ACM SIGMOD Conference; 1999; 1999.
3. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD International Conference on Management of Data; 1993; 1993.
4. Agrawal R, Srikant R. Fast algorithms for mining association rules. 20th Int Conference on Very Large Data Bases, VLDB94; 1994; 1994.
5. del Jesus MJ, Games JA, González P, Puerta JM. On the discovery of association rules by means of evolutionary algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2011;1.
6. Aggarwal C. Data Streams, Models and Algorithms: Springer; 2007.
7. Bekkerman R, Bilenko M, Langford J. Scaling Up Machine Learning: Parallel and Distributed Approaches: Cambridge University Press; 2011.
8. Heckerman D. A tutorial on learning with Bayesian Networks: Microsoft Research; 1995.
9. Samarati P, Sweeney L. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. Technical Report; 1998.
10. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramanian M. ℓ-diversity: Privacy beyond kappa-anonymity. 22nd International Conference on Data Engineering; 2006; 2006.
11. Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy beyond k-anonymity and l-diversity. 23rd International Conference on Data Engineering; 2007; 2007.
12. Dwork C. Differential Privacy; 2006.
13. Dwork C, Smith A. Differential Privacy for Statistics: What we Know and What we Want to Learn. . Journal of Privacy and Confidentiality 2009;1:135.
14. Friedman A, Wolff R, Schuster A. Providing k-Anonymity in Data Mining. VLDB Journal 2008;17:789.
15. Grosskreutz H, Lemmen B, Rüping S. Secure Top-k Subgroup Discovery. ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning; 2010; Barcelona, Spain; 2010.
16. Abernethy AP, Etheredge LM, Ganz PA, et al. Rapid-learning system for cancer care. J Clin Oncol 2010;28:4268-74.
17. Cesario A, Marcus F. Cancer Systems Biology, Bioinformatics and Medicine; 2011.
18. Castro MA, Wang X, Fletcher MN, Meyer KB, Markowetz F. RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. Genome Biol 2012;13:R29.
19. Wang X, Kang DD, Shen K, et al. An R package Suite for Microarray Meta-analysis in Quality Control, Differentially Expressed Gene Analysis and Pathway Enrichment Detection. Bioinformatics 2012.