



SHARE

USE CASE SCENARIO FOR EPIDEMIOLOGY

| | |
|------------------|---|
| Document ID: | SHARE-D5 1a_v1.2.doc |
| Date: | 19/02/07 |
| Authors: | I. Blanquer, V. Hernandez, N. Jacq, T. Solomonides, M. Olive |
| Activity: | WP5: Applications |
| Document status: | FINAL |
| Document link: | http://eu-share.org/deliverables.html |
| Confidentiality: | Public |
| Keywords: | Epidemiology, Use Case Scenarios. |

Abstract: This document presents an analysis of the status of the use of ICT in epidemiological studies from the perspective of healthgrids. A use-case is presented and analysed, discussing the users, data, current practice, limitations, technological needs and barriers to adoption of healthgrid technologies. The use-case also suggests possible new approaches to epidemiological modelling and decision support through the use of healthgrids. The document should be used to validate the roadmap deliverables with respect to the needs of a relevant user community and problem.

Document Log

| Issue | Date | Comment | Author |
|-------|----------|---|--|
| 0 | 02/1/07 | First version | I. Blanquer, V. Hernández |
| 0.8 | 25/1/07 | First internal release | I. Blanquer, V. Hernández |
| 1 | 31/1/07 | First public Release | I. Blanquer, V. Hernández |
| 1.1 | 15/2/07 | Review / changes; Second public Release | M Olive; T Solomonides, I. Blanquer, V. Hernández, N. Jacq |
| 2.0 | 19/02/07 | Final Release | Y. Legré |
| | | | |

Document Change Record

| Issue | Item | Reason for Change |
|-------|---|---|
| 0.8 | Sections are completed. | Preparing first release |
| 1 | References, Glossary, all sections in general. | References filled-in, first review |
| 1.1 | Whole document Comments and General improvement | Review, changes and comments, Typos and content enhancement |



CONTENT

| | |
|---|-----------|
| 1. INTRODUCTION..... | 4 |
| 1.1. PURPOSE..... | 4 |
| 1.2. APPLICATION AREA..... | 4 |
| 1.3. REFERENCES..... | 4 |
| 1.4. TERMINOLOGY..... | 5 |
| 2. EXECUTIVE SUMMARY..... | 7 |
| 3. THE USE CASE SCENARIOS..... | 8 |
| 4. USE CASE 1: EPIDEMIOLOGY..... | 10 |
| 4.1. MEDICAL PROBLEM..... | 10 |
| 4.2. USERS..... | 11 |
| 4.3. DATA AND DATA PROVIDERS..... | 12 |
| 4.4. CURRENT PROCEDURE..... | 14 |
| 4.5. LIMITATIONS..... | 16 |
| 4.6. AIM AND BENEFITS..... | 17 |
| 4.7. TECHNICAL REQUIREMENTS..... | 18 |
| 4.8. OTHER CONSTRAINTS..... | 20 |
| 5. CONCLUSION..... | 22 |

1. INTRODUCTION

1.1. PURPOSE

The purpose of the document is to present an analysis of the status, needs, requirements and trends in the area of epidemiology. This document presents a general use-case scenario in epidemiology that reflects a wide application community and discusses the suitability and current trends in using grids. This document has a corresponding version for Innovative Medicine identified under the code D5.1.b. This document is self-contained.

This document is intended both for users (epidemiology researchers, public health authorities or industry) and developers (technologists and legal experts) to define the major hurdles that must be overcome for the adoption of grids in epidemiology. More basic information about healthgrids, is provided in the deliverable D3.1 “HealthGrid Framework” and in the HealthGrid White Paper [14].

This document is written in preparation for a roadmap towards the adoption of grid technology in Health and Life Sciences. The merged analysis of this document along with D5.1.b and deliverables D6.1 will result on the final roadmap. D3.2, D3.3 (for the current status and bottlenecks of technological and security aspects) and D4.2, D4.3 (for the current status and bottlenecks of ethical, legal, social and economic aspects will contribute to D6.1 and to subsequent development.

All deliverables can be downloaded from the SHARE web-site at www.eu-share.org.

1.2. APPLICATION AREA

The document is intended for internal and external use. It will be used as a dissemination tool for the Share project.

1.3. REFERENCES

- [1] Carmona L, Ballina J, Gabriel R, Laffon A. The burden of musculoskeletal diseases in the general population of Spain: Results from a national survey. *Ann Rheum Dis* 2001;60:1040-5
- [2] Early detection of Breast Cancer Programme. <http://www.sp.san.gva.es/DgspWeb/ppcm/>
- [3] Results of the EUROCARE-3 study, covering survival up to 1999, published in December 2003 as a monograph in the *Annals of Oncology*, appearing as Volume 14, Supplement 5.
- [4] European Network of Microbiological Surveillance, European Decision n° 2119/98/CE, <http://europa.eu/scadplus/leg/en/cha/c11548b.htm>
- [5] Basic Surveillance System of the Valencian Public Health Authorities, http://www.sp.san.gva.es/epidemiologia/Main?ISUM_ID=Groups&ISUM_SCR=groupScr&ISUM_CIPH=Jh-7XJP8mDDwkInHP3QPjJWQR2huc5Af
- [6] OGSA-DAI middleware, available at <http://www.ogsadai.org.uk/>
- [7] Integrating Healthcare Enterprise, <http://www.ihe.net>
- [8] Overview of the OpenEHR, http://svn.openehr.org/specification/TRUNK/publishing/openEHR/introducing_openEHR.pdf
- [9] European Standardisation of Health Informatics, <http://www.centc251.org/>
- [10] Virgilio Cavicchioli Neto, Henrique Fabricio Gagliardi, Alexandre Rezende, Erick Sobreiro Gonçalves, Eduardo Gallo, Fabricio Alves Barbosa Silva, Ivan Torres Pisa, Domingos Alves. “Data Access Service in a Computational Grid Platform Applied to the Monitoring and Control of Epidemics on Georeferenced Dynamic Maps”, proceedings of the e-Science conference, 2006

- [11] VOMS, Virtual Organisation Membership Service, <http://edg-wp2.web.cern.ch/edg-wp2/security/voms/>
- [12] PERMIS, PrivilEge and Role Management Infrastructure Standards Validation, <http://www.permis.org/en/index.html>
- [13] Understanding Shibboleth, <https://spaces.internet2.edu/display/SHIB/UnderstandingShibboleth>
- [14] V. Breton, K. Dean and T. Solomonides, editors on behalf of the HealthGrid White Paper collaboration, "The HealthGrid White Paper", Proceedings of HealthGrid conference, IOS Press, Vol 112, 2005
- [15] *Nosocomial Infections as a Suitable Case for the Grid* Tony Solomonides **Proceedings of the IADIS International Conference: e-Society 2003**. Lisbon, Portugal. June 2003. A Palma dos Reis and P Isaias, Editors pp 1037-1038. ISBN 972-98947-0-1 IADIS Press, Lisbon
- [16] *Conceptual Modelling of an Epidemiological Information System* Tony Solomonides, Mohammed Odeh, Richard McClatchey, Jean-Marie Le Goff, João Carriço and Jonas Almeida **Proceedings of the IADIS International Conference: e-Society 2003**. Lisbon, Portugal. June 2003. A Palma dos Reis and P Isaias, Editors pp 794-799. ISBN 972-98947-0-1 IADIS Press, Lisbon
- [17] See the 'Dossier of Concerns' by 23 academics at <http://homepages.cs.ncl.ac.uk/brian.randell/Concerns.pdf>

1.4. TERMINOLOGY

Glossary

| Term | Definition |
|---------------------|---|
| Data provenance | Reliability of a piece of data according to the origin and their processing. |
| Data quality | Integrity, completeness and accuracy of a distributed set of data. |
| Effectiveness | Validation of the suitability of a therapy to solve a medical problem based on a sample population under ordinary clinical practice conditions |
| Efficacy | Validation of the suitability of a therapy to solve a medical problem based on a sample population under optimal conditions, such as a clinical trial |
| Efficiency | Analysis of the suitability of a therapy in the whole population and under the real conditions of medical practice |
| EHR | Electronic Health Record |
| Epidemiology | Scientific study of factors affecting the health and illness of populations |
| Federation | Integration of different sources of information |
| HealthGrid | Adoption of Grid Technologies in a Health IT environment. |
| ICD-9 | International Coding of Diseases, Version 9. |
| ICD-O | International Coding of Diseases, oncology. |
| LOINC | Logical Observation Identifiers Names and Codes |
| Medical Informatics | Usage of computer methods for storing and processing medical information related to patients' health. |
| Morbidity | Incidence of illness; measure of damage to health. |
| Mortality | Rate of death resulting from a particular condition. |
| OGSA-DAI | Open Grid Services Architecture – Data Access Integration |



| | |
|-----------------|--|
| Ontologies | A conceptual framework for the integration of information from diverse sources. A set of attributes and functions associated with a concept. |
| Prevalent cases | Cases that have been identified and diagnosed in the past. |
| SNOMED | Systematized Nomenclature of Medicine. A terminology for defining medical terms. |
| Terminologies | Subset of an ontology that defines a vocabulary. |
| UMLS | Unified Medical Language System, a set of Knowledge Sources for easing medical informatics developments. |
| Use-case | Example complete scenario that describes a particular problem, actors, data and procedures to tackle that problem. |

2. EXECUTIVE SUMMARY

This deliverable presents a use case concerning epidemiology and analyses the current status and trends that could be related to the introduction of healthgrid technologies. Epidemiology is defined as the scientific study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine. It is considered a cornerstone methodology of public health research, and is highly regarded in evidence-based medicine for identifying risk factors for disease and determining optimal treatment approaches to clinical practice. Although there are many different activities in the frame of epidemiology, this deliverable tries to concentrate in a use case that could be general enough to reflect current practice and to capture near-future trends in the light of possible adoption of healthgrid technologies. This use case can be summarised as the development of extensive (population-level) retrospective studies of the morbidity and mortality of treatments, population features and additional clinical factors. The results of these studies can be used to develop prospective analysis that could guide medical practice.

In this use case, the main users (epidemiologists, public health authorities, pharmaceutical companies), gather the data from different sources (primary care, hospital, demographic information, prescription, mortality records, social assistance, environment), integrate them and execute advanced statistical methods (such as Bayesian inference, Markov chains, decision trees, etc.).

There are many technical limitations due to the data, as the large size, the distributed structure and the variable, often poor, quality of coding or the difficulties in each extraction. Many of these difficulties can be solved by the use of grid data integration techniques, although there are difficulties associated with provenance that are structural and cannot be solved without the modernisation and agreement on standards interoperability of data sources. Other technical limitations concern data processing, with opportunities now to adapt, migrate and exploit data mining techniques which have proved successful in industrial areas. The adoption of such techniques requires a large increase of the computational resources, which can also be faced from the point of view of grid computing technologies.

On the other side, there are other issues that affect the processing of epidemiological data, such as the legal, social and ethical aspects that European and national regulations impose concerning the patient consent, user authentication and authorization, permission revocation, logging and auditing. The use of grid technologies can help with some of these problems. Finally, there are other aspects related to medical informatics and more precisely with data coding and integration that should converge with grid technologies providing extensive interesting results.

So, from the analysis of the problems in epidemiology, grid constitutes an enabling technology that can solve many technological issues and foster scientific development in the area, provided that the requirements for security, broadly conceived, data integration and reliability of epidemiological studies are adequately addressed.

3. THE USE CASE SCENARIOS

A use-case scenario represents a significant example in the relevant application area. Such an example must be generic enough to be representative but specific enough to enable a clear and accurate analysis. The use-case must be related to a real application case, or at least to a reasonable foreseeable application. It need not necessarily involve computers currently but must be sensible for the use of grids. A use-case mainly describes how a set of actors solve a problem using some data and following a specified procedure requiring some technical means, which produce some benefits and have some limitations. The “what, how, who and where” involved must be clearly defined. The purpose of defining and analysing the use case scenario is to cross-validate the requirements and needs identified in the roadmaps from workpackages 3 and 4.

The use-case scenarios must consider all the actors, data, processes, limitations and benefits, and present them in a clear way. Thus, for each case the following data are selected:

- **Medical Problem.** This should describe a medical problem with wide scope, indicating the main implications in health management and outlining its importance.
- **Users.** The use-case could involve different users depending on the interest for the results. Medical users, industry users, researchers, policy makers and the public could share the same use-case but would consider different targets or input sources. The maturity in the use of ICT, the size of the population and their relevance should be outlined to assess the importance of the use-case.
- **Aim and Benefits.** The benefits will be clearly identified for all the users involved, as well as the side-effects. Desired aim and potential benefits that may not be achieved should also be outlined.
- **Data and Data Providers.** The data sources that are necessary for the use-case should be identified. Important factors in this part are the availability of the data, the restriction in its usage, privacy issues, representativeness and coverage, quality of the data and ICT maturity, issues generally identified under the label ‘provenance’.
- **Current Procedure.** The means used to solve the medical problem must be described. This means will surely involve computer-based processing, but possibly not at large or for only part of the problem. Actors involved in the process and identified in the “users” section must be mentioned.
- **Limitations.** The current or foreseeable procedure will be bound by some restrictions and limitations that reduce its impact and relevance. These limitations can be technical (lack of performance or resources) or scientific (lack of knowledge) due to the procedure itself.
- **Technical Requirements.** The technical requirements needed for the current procedure should be identified. Technical requirements to achieve the expected aim should be also identified if possible.
- **Other Constraints.** Along with the technical constraints, there are normally be legal, ethical, regulatory and socio-economic constraints that must be faced to achieve the aim.



The use-case scenarios selected in the SHARE project are epidemiology and innovative medicine. The case of epidemiology is described in this deliverable.

4. USE CASE 1: EPIDEMIOLOGY

This section describes, according to the definition of a use-case provided before, the use case scenario 1, related to epidemiology. Epidemiology is “the scientific study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine. It is considered a cornerstone methodology of public health research, and is highly regarded in evidence-based medicine for identifying risk factors for disease and determining optimal treatment approaches to clinical practice”¹.

The interest of epidemiology is not the individuals but the population, and the main source is the evidence. So the more representative and the larger the population is, the more accurate the results are. This normally implies wide geographic coverage and large data sources.

An important tool for the epidemiologist is the development of statistical, mathematical, philosophical, biological, and psychosocial theory dealing with the problem of the health of the population. Defining the diseases, drawing disease causal chain / chains, and formulation of health strategy are important aspects of epidemiology.

Modern epidemiologists use disease informatics as a tool. Moreover, epidemiologic studies are generally categorized as descriptive, analytic (aiming to examine associations, commonly hypothesized causal relationships), and experimental (a term often equated with clinical or community trials of treatments and other interventions). From the point of view of this use-case, we add the *predictive* use of epidemiological methods in decision support systems for use by public health authorities. To this end, simulation and modelling may also be undertaken for the purpose of modifying future behaviour as for describing it. [14], [15]

4.1. MEDICAL PROBLEM

Many are the problems that could be tackled by grids in Epidemiology. However, most of them follow a similar structure: Analysis of a large volume of distributed data through complex simulation, statistical procedures or knowledge discovery methods. Thus, in this deliverable we focus in a very significant case that involves most actors of the healthcare ydeliver chain.

Epidemiology must deal with multiple (and normally distributed) data sources, performing complex statistical and data-mining analysis on them. Without the analysis of these databases, the epidemiologist must rely on the results of the efficacy (validation of the suitability of a therapy to solve a medical problem based on a sample population under optimal conditions) or effectiveness studies (similar validation but under clinical practice conditions) provided by the euticalpharmacy industry, although these usually do not fit the real profile of the population. Despite the use of randomised samples, failure to match sufficiently closely the (age, gender, etc) structure of the population may affect the significance of the results. Efficiency studies (analysis of the suitability of a therapy in the whole population and under the real conditions of medical practice) are difficult due to the large number of data that must be recorded and processed.

¹ Wikipedia: <http://en.wikipedia.org/wiki/Epidemiology>

Thus, the analysis of the efficiency of drugs in the treatment of different diseases is generally not deeply analysed. The evolution of populations and specific features of the factors causing the diseases make some treatments inefficient or even unsafe. For example, it has been suggested that anti-rheumatoids have a large impact due to adverse reactions [1] and even it is one of the most important causes of death in aging population. However, it is not possible to substitute its use in all cases. There are alternatives which are less aggressive, but which cannot be used on a large scale for several reasons. Thus, knowing the exact impact of the drugs on the population would enable selection of selecting groups at risk and provision of different treatments. This approach can be applied to many other examples, such as the early detection of resistance or the efficiency analysis of vaccines. Moreover treatments are defined at global level without considering the diversity of diseases and patients.

Other study areas relevant to the epidemiologists are the Oncological Information Systems, the Microbiological Surveillance Networks, the Biological and Environmental Risk Management, or Bio-banks. All these areas share similar concepts with differences in their implementation. Oncological Information Systems deal with the identification and follow-on of cancer cases in the population and are very exhaustive. Normally, a high coverage has been reached in several specific topographies and morphologies, even at European level through the development of European Cancer Survival Studies (EUROCARE, European International Society of Pediatric Oncology), in which the experts have developed standardised protocols. Microbiological Surveillance Networks are more concerned with detecting abnormal trends on infectious diseases and early detection of outbreaks (Valencia RVM [5], French REMI). Biological and Environmental Risk studies normally aim at detecting and following-on healthcare and industrial practices that could jeopardise the health of the staff or the users. Finally, Bio-banks, storing both patient data and biological samples are a very good research source, both for public and private research. Indeed, they constitute an interesting business and constitute a way to regulate the distributed – and even illegal in some cases – databanks available in healthcare centres.

So, summarising the different use cases stated before, the medical problem can be described as the development of extensive (population-level) retrospective studies of the morbidity and mortality of treatments, population features and additional clinical factors. The results of these studies can be used to develop prospective analysis that could guide medical practice, develop novel models and provide decision support for intervention. Healthgrid technology would make it possible to conduct ‘real-time’ concurrent studies and to recognise epidemic trends in time for intervention. [14]

4.2. USERS

The main users or actors that are involved in the process are, on one hand, those who are interested in the population, such as epidemiology researchers and public health authorities, and on the other, those who are aiming more at developing treatment, such as the drug developers and researchers.

Epidemiology authorities are both the producers and the consumers of the results obtained from the study of the population. The results can guide policies and treat recommendations to improve the general public health and reduce morbidity and even mortality. Results are obtained from research performed by the epidemiology researchers; they are the actors who

define the experiment, the data and the processing. Their results are also of great interest to pharmaceutical companies, who only have the results of clinical trials which involve selected patients and under very controlled conditions.

Public health officials and policy makers, including governments, are deeply interested in controlling epidemics, whether nosocomial (such as MRSA in the UK) or pneumococcal infections of children in nursery settings (see the EURIS project). Among the potential outcomes they are interested in are simple spread models which combine understanding of prescription levels and resistance, transport networks, etc, to provide some decision support.

Finally, general practitioners are also indirect users of the results. General practitioners should use the public health guidelines that epidemiologists define for the general treatment of the population. Based on their experience, they can alternatively select other treatments if the individual profile of the patient differs from the general population. They can also warn the risk population about diseases that are being currently diagnosed.

4.3. DATA AND DATA PROVIDERS

Epidemiology studies require as input a large amount of data. Data must be relevant to the medical problem being analysed and population selected must be representative of the general population. The extent of the information could be:

- Representative reduced sample of the population. Effectiveness, safety and efficacy studies are based on reduced samples. This is so due to the difficulties of recording and accessing the data and the processing means available. All hospitals perform clinical trials.
- The greatest coverage possible. Real epidemiology must work with the largest volume of high quality data possible (ABUCASIS, Generation Scotland, EURO CARE).

Normally, the problems that the epidemiologist has to face regarding the data collection are:

- Data are distributed. Data are recorded at the point-of-care and only in few cases are transferred and stored centralised. Due to the lack of transparent and automatic means, epidemiology studies collect manually the data, only existing defined protocols in the best cases.
- Data are heterogeneous. Data are stored following different code schemas and data structures. Different code schemas are not necessary directly interoperable and require semantic integration. The different data structures should be integrated in a common, minimal subset.
- Data are primarily collected for healthcare and not for epidemiology making this kind of research more difficult. Data are obtained during care delivery by multiple operators and under different conditions, and their suitability for epidemiological research is not a primary concern. Depending on the skills, emergency level and willingness of the data operators, accuracy and quality of codification may vary. This may be considered a major motivation for a systematic approach to data provenance.
- Data sources are not homogeneous. Data are stored in different back-ends, following incompatible data formats.

- Data volumes are large. The amount of data depends on the size of the population and the input sources, although in all cases is updated daily on the order of millions of records (increasing daily by orders of Gigabytes). A fixed snapshot of the whole population is of the order of Terabytes.

The different pieces of information that are used belong to different kinds of sources

- Demographic information. On one hand, identification keys and other sensible data are needed at the integration step, although they are disaggregated later. On the other, geographic information about residence is very important.
- Primary care. Information about general practitioner and specialist visits, including diagnosis, date and symptoms.
- Prescription. Information about the drugs prescribed, the dose and the period. A change of treatment or a prolongation of it would be key data for the efficiency analysis.
- Hospital Information. Information about hospital episodes and treatments, nosocomial infections, in the same sense as in the previous cases, but considering the differences on the source.
- Mortality records. Information about deceases is not trivial to obtain, since there are normally separated and protected by strictly regulations. Normally, the mortality records can only verify the correctness of the information for a complete data query.
- Social assistance. Social information about economic conditions is important for treatment monitoring.
- Environment information. Information about weather (an important factor in the spread of many pathogens, rheumatoid diseases, etc.), pollution or exposure to risk factors.

As health management is distributed, data are collected from different sources. This distribution relates to both the data sources for the information from a specific individual, but also to consolidate the whole population. The integration of data relates to several levels: person, time and place. Currently, most countries and regions in the developed countries are making efforts to consolidate information from patients and populations, although very few are aiming at consolidating the whole population.

The owners of the data are generally public health authorities. Private centres are obliged to report about the mandatory notification diseases (which vary from country to country), although this could involve a minimum part of the information that could be difficult to integrate.

One significant advantage of a healthgrid approach is that it enables integration in a virtual database without forcing the data to be brought together to one physical database. In the UK, where the so called 'spine' is being implemented as an essentially monolithic database of patient records, 51% of GPs told a recent survey that they would not be willing to transfer patient records onto the spine without explicit consent from each patient. [17]

4.4. CURRENT PROCEDURE

Currently, epidemiology is mainly limited to efficacy, safety and effectiveness studies. These studies are performed through clinical trials on controlled populations with low interaction of treatments and by pharmaceutical companies mainly. The data are directly and extensively recorded from the point-of-care by the companies who pay the studies. Reports are published and made available to clinics and public authorities who authorise the use of a drug.

In very specific cases, whole population is considered, but normally focused only on a specific disease. For example, Oncological Information Systems [2, 3] or Microbiological Surveillance Networks [4].

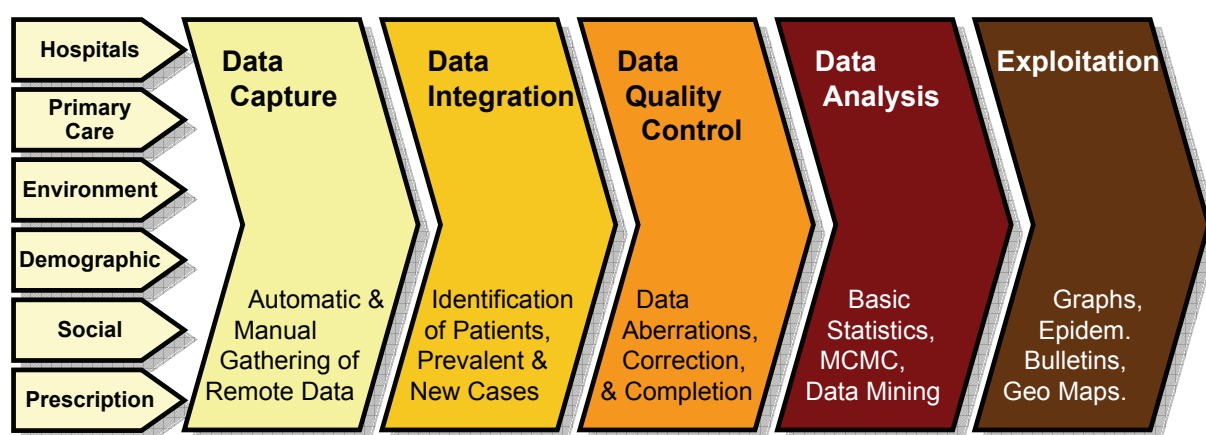


Figure 1: Steps of the Epidemiological Analysis Procedure, from Data Gathering to Exploitation of the Results.

In most cases, current procedure (described in figure 1) involves the following steps

- Data capture. Data are normally stored, distributed and must be retrieved for processing. Different approaches are followed, ranging from creating central warehouses with aggregated information or virtualising a federated repository. Normally, first approach is preferred. Data capture could not normally be performed in a fully automatic way. Repositories are heterogeneous and data are normally incomplete, requiring expert intervention. Even setting aside compatibility problems, data capture is not trivial. Data must be electronically recorded and access to download data must be provided. Since many systems use proprietary solutions, even this extent cannot be guaranteed. Moreover, the process should be as automatic as possible, which could require interfering with codes. Currently this process is performed:
 - Manually on the systems that have the possibility of raw or customised dumps of data periodically by the system operators. Data are packed and manually sent through secure mechanisms.
 - Manually through remote administration tools by authorised users in the epidemiologic services.

- Automatically through remote administration tools and macros. Used in practice when the technical means exist but the access to the source code of the recording programs is not possible.
- Automatically on the systems that have been modified to do periodic dumps and transfers of this data. Desirable but very rare.
- Patient identification. Data from different sources but concerning the same individual must be identified. Different data sources have different, normally incompatible, identification means. Most widely used means are specific population databases, since passport numbers, social affiliation numbers or even identity card numbers are not univocal in cases such newborn. Thus, identity federation schemas must be implemented to merge the pieces of information to build-up the patient information file. Keeping patient personal data implies high privacy restrictions, but it is necessary to keep track of the evolution of patients. Personal data storage can be skipped if all data, retrospective and current, can be retrieved for each study. In those cases, personal data can be disaggregated once the patient identification is performed.
- Identification of prevalent cases and new patients. In most epidemiological studies, it is very important to locate the temporal scope of the cases. Identification of first visit and diagnosis and follow-on episodes will give information about the effectiveness of the treatments.
- Quality control. Different pieces of information could be incomplete, poorly coded or even incorrect. Cross-checking should be used prior to the analysis to avoid introducing errors. Deficient coding could make pieces of information useless. Normally, data are cross-validated, trying to detect:
 - Aberrations. Information pieces that can be directly discarded (e.g. impossible anatomical location, sex-incompatible diseases, erroneous dates, etc.). In these cases the valid data that remains is easily assessed and can be performed automatically.
 - Incoherent data. Data in different fields can be incoherent but possible separately. This process is extremely complex since the correct field cannot be easily assessed. Normally they are manually reviewed by experts or directly discarded.
 - Incomplete data. Data in some entries could be very general or not carefully coded, and information in other fields could increase its value. This could be done automatically but verification is normally required by an expert.

Individual health management must be much more accurate, since a single erroneous datum can lead to an important negative effect on an individual's health. On the contrary, in epidemiology there is not a need for a total certainty of the accuracy of the whole data. The results have a confidence interval in which they are valid, and health management decisions of individual patients are not taken directly on those results. Epidemiology studies are more interested on exhaustively screening the population, assuming that a reasonable small number of mistaken pieces of information are considered

- Analysis. Currently, analysis is simply performed by aggregation of data and the analysis of the tendency of new cases. This aggregation is performed by geographical area, sex, period (week/month) or age segment. For example, the weekly epidemiological survey of the Valencian Authorities, records the new cases on Tuberculosis, Flu, Legionellosis, Chickenpox, Hepatitis, Rubella, Parotiditis for women and men, considering the geographical distribution of the population and their age. This information is compared with respect to the annual tendency records.
- Exploitation. The results of the analysis are normally filtered to be used by medical experts. Normally information is consolidated as simple graphs or tables or used for the surveillance methods, which automatically raise warnings when specific conditions are met. Again, a good example is the on-line weekly epidemiology bulletins of the Valencian Regional Authorities.

4.5. LIMITATIONS

Safety studies are necessary for the approval of a treatment or a drug. However, proving that a treatment does not cause undesired negative effects is not enough. Effectiveness studies can probe that the drug or treatment has the desired effect on the disease, but limiting to a sample of the population could be misleading, since:

- Populations are carefully selected, so they could reflect neither the structure nor the changes in the features of the patients. Indeed, the sample population is normally monitored more often or using special procedures which are not considered in normal practice (at least in all the centres).
- Interactions with other treatments are not extensively tested, so that many adverse effects remain hidden until real use on a population. The target population usually does not implicate a large number of drugs or other a short variety of them is tested.
- Cost-effectiveness of the treatment is not evaluated. The suitability of different treatments could vary depending on the social or biological profiles of the individuals.

These limitations are a consequence of several problems, which cannot be faced in the context of efficiency studies involving the whole population. The sources of these problems are at different levels and must be considered separately. The most important problems are:

- Data are incomplete or inaccurate. If this limitation is due to the fact that data are not electronically available or not recorded, there is clearly no additional technological solution that can solve this issue from the epidemiologist point of view. However, in many cases it is a problem of interoperability or need for manual intervention, which can be solved using upper middleware layers. Finally, on top of all this, there is a need for defining common policies.
- Integration of data from the whole population is not performed. Even if it is available, the integration might not be possible due to limitations on the coding and structures. Federation schemes (such as OGSA-DAI [6]) can solve the problems due to structuring, although the semantic restrictions will remain. Interoperability among schemas is being targeted in medical informatics activities, such as IHE [7], OpenEHR [8] and the CEN/TC 251 [9] standards. Additionally, healthgrid projects such as IBHIS and Health-e-Child are developing mechanisms for the integration of both

syntactically and semantically heterogeneous data. Coding quality might be automatically improved with the integration of data and the development of specific services. In this point there is an important need for the involvement of medical experts and medical informatics community.

- Statistical methods needed to analyse the information are costly. Basic statistics (aggregation, averaging, specific indexes, etc.) require a reasonable computational cost compared to the needs for data storage. However, more advanced data mining techniques, Bayesian inference, or Markov chains require much more computational resources and are not applied currently. Most tools used in this area are a mixture of commercial solutions and open-source tools. Well-known statistical analysis products have their compatible open-source counterparts which can be integrated easily on the Grid. Normally the computing model required is a combination of high-throughput and parallel computing.
- Finally, the legal and ethical management of data define many regulations that must be complied with. Patient consent, auditing, privacy management, users authorisation, logging and revocation management are issues that prevent from the aggregation of databases coming from different sources and database managers. However, grid services could facilitate the on-demand integration of data sources, allowing each source to retain ethical control over its data, and improved mechanisms for protecting patient privacy and ensuring appropriate consent are also being explored.

4.6. AIM AND BENEFITS

The main aim of the epidemiological case study selected is to have the ability to analyse the effects on the real population of the drugs and treatments focusing on factors such as:

- Hidden or not well-quantified interactions with other drugs, via correlation of prescription and morbidity. Many treatments present undesired effects, especially in the case of old or chronic patients with the prescription of multiple drugs. These interactions are not addressed even on effectiveness studies and can only be seen on a large cohort and during a long observation period. A well-known case is the use of anti-rheumatoid drugs, which are an important cause of vascular diseases ending with fatal results. Anti-rheumatoid drugs cannot be substituted in all the cases but doses and prescription can be restricted in the cases that cannot be avoided.
- The co-occurrence of other diseases alters the effectiveness of drugs and treatments. It is well-known in very specific cases or diseases (such as AIDS and tuberculosis), making some treatments inefficient. Moreover, the influence of migrations in European countries is modifying the profile of the average patient, changing the original profile (e.g. vaccination) of the citizens. Effectiveness studies should be performed, not only at a population level, but splitting the population in clusters that could be identified
- Effectiveness on different population groups, via correlation of demographic and clinical profiles and treatments. It is clear that the integration of genetic profiles, the concurrency of other diseases and demographic information could give more clues

about the suitability of treatments for different groups of people. This can only be obtained aiming at general level.

- Verify effectiveness published by providers with respect to the real population. Vaccines and antibiotics are important targets for this issue. Currently, health management authorities must rely on the results published by the pharmaceutical producers, which have results on controlled populations and normally do not have large impact results. Normally, only severe cases are detected. Assessing “in production” the effect of vaccines on population will feedback the health policies, and even the industry.

4.7. TECHNICAL REQUIREMENTS

Technical requirements are related to those needs that can be solved by the adoption of consolidated technology, and normally are only bounded by economic and social problems. Along with the technical requirements, there are, on one side, legal, ethical and economic requirements that could prevent feasible technical requirements to be adopted, and new scientific needs that require an advance in the knowledge.

The technical requirements are focused on several main areas:

- Data restrictions: Source data, mainly primary care and hospital data must be extensively recorded and made available in an integrated fashion, even if it is recorded and stored distributed. Data formats and coding must be compatible and followed by the users. The problems faced are:
 - Data Capture. Data must be automatically and periodically extracted from the distributed repositories. This must be done through demons, ‘cron-like’ processes or remote administration and directly transferred. The use of manual procedures is neither scalable nor robust.
 - Integration of the data. Data must be syntactically and semantically compatible. This requires more than a basic matching of fields, since different coding schema could be incompatible and reflect different levels of detail. A semantic integration will require consulting different sources of data to make the rightmost matching. The use of middleware solutions such as OGSA-DAI could provide the processing means and data compatibility schemas as demonstrated in the literature [10]. Integration of more semantically complex data may require specialised services.
 - Data Coding. On the other side, the quality of the results relies on the quality of the data, mainly constraint by the data coding. There exist standard coding schemas for coding and transferring medical data, such as ICD-0, ICD-9, ICD-10, HL7, OpenEHR, EN13606, etc., ontologies, such as UMLS, SNOMED, LOINC, or even natural language (such as MENELAS). and de-facto standards. Finally, although there could be a de-facto agreed application or format for a concrete area, the users could just write plain text. In these last cases, there is a need of text-processing tools that could analyse the plain text and try to deduct the suitable coding. Data provenance techniques, which associates a degree of “accuracy” to the data according to the source, the

consistency with other sources of data and the processing performed, are of great interest for automating the integration and quality enhancement processes.

- Scientific restrictions: New statistical tools and models should be developed to deal with the large populations and correlation studies. Many data-mining tools used in economics, resource provider industry or other industrial areas can be adapted. With the introduction of genomics, genetics and proteomics, new methods are being applied.
- Computational restrictions: These new methods that are being developed require large computational resources, which are needed both due to the methods themselves and to deal with the size of the large populations covered.
 - Basic statistics (aggregation, averaging, specific indexes, regression, etc.) require a reasonable computational cost compared to the needs for data storage.
 - Model-based statistics, such as Monte Carlo Markov chain (MCMC) methods and Decision Trees define models of the diseases which are trained and filled-in with the data recorded in healthcare delivery. Those models define probabilities and transitions between states that can be used for prospective analysis. Although they have moderate computing requirements, these requirements increase with the complexity of the models and the data sizes.
 - Data mining techniques, such as Bayesian inference and clustering methods are well used in many other knowledge discovery processes. The computational requirements of those methods are far higher than basic statistics.
- Reliability. Healthcare delivery is performed around the clock, and thus requires very reliable and robust systems. Although epidemiology studies are less strict on terms of performance and quality of service, this does not mean that reliability is not important. Data gathering and integration is a critical task that must be performed seamlessly to avoid data replication and gaps. Large computing processing tasks could be more flexible to resubmission or queuing waiting times, in the same degree as any other research activity. Normally, epidemiological studies are performed periodically (weekly, monthly, yearly) and most of these processes are being automated, as the size of data increases, so a minimum quality of service must be guaranteed to ensure the publication of results on time.
- Network Restrictions: Different data providers and resource centres should be connected through high-speed links. The bandwidth needed will depend on the accessing means and on the data sources. There would be cases (primary care for example), in which a continuous transferring of short records will be necessary, being the latency very important. Other cases, (such as microbiological sources, demographic information, etc.) are accessed periodically, and can probably be programmed to be performed during night or weekends.
- Security Restrictions: Information is confidential, so security measures must be implemented to deal with the legal regulations. Normally, data stored in data

warehouses and federated views of distributed databases are pseudo-anonymised. However, during the identification of patients and the annotation of data, personal identity must be preserved. This requires on one side, implementing high-end authentication and authorisation mechanisms, and on the other side implementing privacy means for data transfer and data storage

- Authentication of users. The use of X509 certificates is wide-spread in the scope of medical informatics. This is being combined in production environments with smart cards containing private keys, and protected with a password. This schema would be even more common with the progressive introduction of electronic identity cards (in experimentation in several European countries).
- Authorisation of users. The authorisation of users is normally performed through Access Control Lists. This would be enough for healthcare data access, involving patients assigned to the practitioners. The difficulties that appear in such environments are due to the large number of users (order of tens of thousands) and data items (order of hundreds of millions) are inefficient for complex authorisation mechanisms such VOMS [11] or PERMIS [12]. More flexible environments in which the authorisation could be delegated and trusted would be more practical, even risking the principles of data ownership of Grid environments. Very distributed mechanisms, such as Shibboleth [13] should be explored.
- Encryption of data. Due to legal regulations, it might be necessary to encrypt data when stored temporally or permanently in external storages. Medical data are bounded by the maximum level of privacy (e.g. level 3 of LOPD in Spain) and external storages are not clearly taken into account. Since data protection regulations could evolve to more restrictive models, it will be important to take into account more advanced protection means. On-the fly encryption and automatic key sharing over different administrative domains seem to be reasonable techniques.

4.8. OTHER CONSTRAINTS

Legal regulations restrict the storage and use of electronic data. Patient consent for population-level information could be unmanageable and encryption techniques for large population data inefficient and even fragile. Skipping the restriction of the laws through anonymisation and data dissociation could be more realistic.

Along with the privacy restriction issues that have been stated in the previous section, other issues that must be taken into account are:

- Traceability of accesses. Accesses to data must be traceable during the integration process and at least meanwhile the private information has not been removed. This is compulsory due to the European and national regulations, which require identifying the potential leakages of privacy and the responsible persons.
- Management of patient consent. The patient consent must clearly outline the usage of the medical data for epidemiological studies.



- Traceability of pseudo-anonymised data. Data could be removed in the future if the donors or closely related people want to revoke the permission. It will be necessary to keep track of the modifications.

5. CONCLUSION

In conclusion, this document has defined the epidemiological case study to be considered for the analysis with respect to the Grid technical requirements identified. The main conclusions of this analysis will validate the trends defined on deliverables D3.2, D3.3 (for the current status and bottlenecks of technological and security aspects) and D4.2, D4.3 (for the current status and bottlenecks of ethical, legal, social and economical aspects).

The technical, ethical, legal, social and economical aims that are outlined on these deliverables should be sufficient to cover the needs and limitations of the application areas defined in this deliverable and deliverable D5.1.b (Use Case Scenario for Innovative Medicine), and should consider:

- **Data management.** Most requirements deal with distributed access to heterogeneous sources of data, coded using different formats and with different degrees of quality. Data indexation, semantic integration, reliable transference, identity identification and other problems must be faced to end up with a mostly-automatic distributed system for gathering and integrating the available data, verifying the correctness and assisting epidemiologist on constructing the source databases.
- **Privacy.** Confidential information must be managed in a secure way and only for the integration purposes, being disaggregated of the clinical consolidated data to enable processing. However, although not possible for most users, patient traceability must be implemented to able data managers to remove data whose permission could be revoked.
- **Security.** Identification and authorisation of users must be scalable and strong enough to enable accessing to the parts of the data a user is authorised to. Single sign-on, authorisation delegation and distributed processing are key issues.
- **Processing.** New ways of processing data are requiring large computational resources for knowledge discovery, simulation tools and other data mining techniques. The large computing demand is due to the larger amount of data and the new methods. Computing means are needed in a periodic and punctual fashion, with reasonable quality of service.

From the consolidation of deliverables D3.2, D3.3, D4.2, D4.3, D5.1a and D5.1b, the final roadmap will be prepared.