The Healthgrid White Paper

Vincent Breton, K. Dean, T. Solomonides

The Healthgrid White Paper

Vincent Breton [a], Kevin Dean [b] and Tony Solomonides [c], Editors
on behalf of the Healthgrid White Paper collaboration (see over)

a *CNRS-IN2P3, LPC*
b *Internet Business Solutions Group, Cisco Systems*
c *CEMS, University of the West of England, Bristol BS16 1QY, UK*

Abstract

Over the last four years, a community of researchers working on Grid and High Performance Computing technologies started discussing the barriers and opportunities that grid technologies must face and exploit for the development of health-related applications. This interest lead to the first Healthgrid conference, held in Lyon, France, on January 16th-17th, 2003, with the focus of creating increased awareness about the possibilities and advantages linked to the deployment of grid technologies in health, ultimately targeting the creation of a European/international grid infrastructure for health.

The topics of this conference converged with the position of the eHealth division of the European Commission, whose mandate from the Lisbon Meeting was “To develop an intelligent environment that enables ubiquitous management of citizens’ health status, and to assist health professionals in coping with some major challenges, risk management and the integration into clinical practice of advances in health knowledge.” In this context "Health" involves not only clinical procedures but covers the whole range of information from molecular level (genetic and proteomic information) over cells and tissues, to the individual and finally the population level (social healthcare). Grid technology offers the opportunity to create a common working backbone for all different members of this large "health family" and will hopefully lead to an increased awareness and interoperability among disciplines.

The first HealthGrid conference led to the creation of the Healthgrid association, a non-profit research association legally incorporated in France but formed from the broad community of European researchers and institutions sharing expertise in health grids.

After the second Healthgrid conference, held in Clermont-Ferrand on January 29th-30th, 2004, the need for a “white paper” on the current status and prospective of health grids was raised. Over fifty experts from different areas of grid technologies, eHealth applications and the medical world were invited to contribute to the preparation of this document.

The Healthgrid White Paper Collaboration[1]			
Chapter	**Coordinators**	**Section authors**	**Reviewers**
Ch 1	Ignacio Blanquer Espert, Vicente Hernandez, Guy Lonsdale	I. Blanquer, V. Hernandez, E. Medico, N. Maglaveras, S. Benkner, G. Lonsdale	Nikolay Tverdokhlebov, Sofie Nørager
Ch 2	Kevin Dean, Sharon Lloyd	K.Dean, S.Lloyd	Siegfried Benkner, Sofie Nørager
Ch 3	Richard McClatchey, Johan Montagnat, Mike Brady	K. Hassan, R. McClatchey, S. Miguet, J. Montagnat, X. Pennec	Siegfried Benkner, Irina Strizh, Sofie Nørager
Ch 4	Johan Montagnat, Xavier Pennec	V. Breton, W. De Neve, C. De Wagter, G. Heeren, G. Lonsdale, L. Maigne, J. Montagnat, K. Nozaki, X. Pennec, M. Taillet	Sofie Nørager
Ch 5	Howard Bilofsky, Chris Jones	H. Bilofsky, R. Ziegler, M. Hofmann, V. Breton, C. Jones	Irina Strizh, Sofie Nørager
Ch 6	Martin Hofmann, Tony Solomonides	M. Hofmann, T. Solomonides, N. Maglaveras	Clive Tristram, Irina Strizh, Sofie Nørager
Ch 7	Ilídio C. Oliveira, Juan Pedro Sanchez, Victoria López	M. Cannataro, P. Veltri, G. Aloisio, S. Fiore, M. Mirto, N. Maglaveras, I. Chouvarda, V. Koutkias, A. Malousi, V. Lopez, I. Oliveira, J. P. Sanchez, F. Martin-Sanchez	Irina Strizh, Sofie Nørager
Ch 8	Georges De Moor, Brecht Claerhout	G. De Moor, B. Claerhout	Sofie Nørager
Ch 9	Jean A. M. Herveg	J. A. M. Herveg	Yves Poullet, Sofie Nørager

[1] See end for individual affiliations.

Contents

1. From Grid to Healthgrid: Prospects and Requirements

1.1. RATIONALE

Evidence-based medicine requires medical decision making to be based on sound knowledge of the patient combined with peer-reviewed scientific evidence, rather than informed guesswork and personal skill. It is also widely accepted that there is a pressing need to move away from manual management of patient information to digital records. Countries in the EU are investing heavily to establish electronic patient record systems. Technically the problem is one of standardization and ensuring that systems are developed that interface through common 'languages' to enable the sharing of information. Technology to secure the information can also be complex and expensive to deploy. Moreover, access to many different sources of medical data, usually geographically distributed, and the availability of computer-based tools that can extract the knowledge from that data are key requirements for providing a standard healthcare provision of high quality.

Research projects in the integration of bio-medical knowledge, advances in imaging, development of new computational tools and the use of these technologies in diagnosis and treatment suggest that grid-based systems can make a significant contribution to this goal. The benefits of improved access are raised to a new level, not merely enhanced by integration over a grid.

Grid technology has been identified as one of the key technologies to enable the 'European Research Area'. A major challenge is to take this technology out of the laboratory to the citizen, thus reaching far beyond eScience alone to eBusiness, eGovernment and eHealth. The benefits of grid technologies in areas involving long simulation processes covering a large set of experiments have been clearly proven. For example, High Energy Physics (HEP) is one of the main application fields of grid technologies [1, 2, 3]. Although grid technologies have clear potential for many applications (those demanding computing or storage power, dealing with geographically distributed information or requiring ubiquitous access), the take up of grid is slow. Reasons for this are the lack of adequate infrastructure, lack of users' confidence and, most frequently, the shortage of applications.

A Healthgrid should be an environment where data of medical interest can be stored, processed and made easily available to the different healthcare participants: researchers, physicians, healthcare centres and administrations, and in the long term perspective citizens. If such an infrastructure were to offer all necessary guarantees in terms of security, respect for ethics and observance of regulations, it would allow the association of post-genomic information and medical data, opening up the possibility for individualized healthcare.

This white paper presents a survey of the healthgrid technologies, describing the current status of grid and eHealth and analysing mid-term developments and possibilities. There are numerous driving forces that are fostering the deployment and exploitation of the secure, pervasive, ubiquitous and transparent access to information and computing

resources that grid technologies can provide. Many technical problems arising in eHealth (standardization of data, federation of databases, content-based knowledge extraction, and management of personal data) can be solved with the use of grid technologies. However, there are many barriers to overcome. The paper considers the procedures from other grid disciplines such as High Energy Physics or numerical simulation and discusses the differences with respect to healthcare, with the intention of outlining a path forward towards the successful deployment of grid technologies for eHealth and ultimately the creation of a Healthgrid.

1.2. INTRODUCTION TO HEALTHGRID

1.2.1. The European Health Sector

eHealth deals with the use of Information and Communication Technologies (ICT) to develop an intelligent environment that enables ubiquitous management of citizens' health status, assists health professionals in coping with some major challenges or integrates the advances in health knowledge into clinical practice.

Many eHealth applications have been developed for dealing with information management and procedural challenges of current healthcare. eHealth is not only a good strategy for improving healthcare quality, but also a good business. The eHealth or Health Telematics sector is becoming the third industrial pillar of healthcare after the pharmaceutical and the medical imaging device industries. It is estimated that health expenditure on ICT systems and services would rise from 1% to 5% by 2010 [10], there are more than 1,500 health care sites on the Internet today and eHealth retailers predict revenues ranging from $22B to $348B (US) by the year 2004. Health care is the second most frequently searched topic on the Internet [11].

Service-based applications in eHealth are an important issue in general business. Application Service Provision (ASP) hosting, for example, makes it possible for service providers to specialize in installing and maintaining applications and services for their subscribing customers. ASP shifts the burden of hardware infrastructure to the providers and frees customers who only need an Internet browser to access the software. The general advantages of ASP, such as staff and resource specialization, broad marketability or scalable investment are complemented by the situation within the health sector: the healthcare market is fragmented, as many people use proprietary systems; many processes are tedious and could be better streamlined; and healthcare organizations have comparatively old legacy computer systems and less ICT staff than other sectors. However, there are some disadvantages. The ownership of mission-critical client functions is much more important in the case of healthcare. Moreover, health records persist over long time-frames and thus require long-term storage, subject to strict legal requirements on data protection and security. High service-level provision is also critical in healthcare, while medical information can require high bandwidth connections to meet minimum delay requirements. Nevertheless, electronic processing of medical data has opened many possibilities for improving medical tasks such as diagnosis, surgical planning or therapy, both in daily clinical practice and clinical research.

Linking databases with medical information is necessary, but it is only part of the solution. Further processing of the information to extract knowledge is a must, since the sheer volume of information makes it impossible to search directly. Data mining can provide the means to analyse relevant information and perform population-oriented health studies

Electronic processing of medical data is at different stages of evolution in different places and even in different departments. Hospital Information Systems are widely used for in-patient management. Laboratory and image diagnosis departments have an important degree of electronic management of patient data. The adoption of these technologies in Primary Care is less advanced. However, the main challenge is the so-called Electronic Patient Record (EPR). EPR promises coherent access to and management of the complete patient information of an individual or a population. There is a great deal of effort already invested to achieve this aim, including on standardization.

Security is the most important issue. Personal data (any piece of information in which its owner can be identified, either directly or in combination with information that is available or can otherwise located) is confidential, so access to the information must be performed only by authorized and authenticated persons, and data must be encrypted to guarantee its confidentiality and integrity. Moreover, electronic archiving of personal data is strictly regulated by European and national laws. Pervasive access and fault tolerance are other important aspects, since medical practice requires round-the-clock availability.

Medical information is voluminous and dispersed. Large resources are needed to store patient records comprising images, bio-signals, plain text, videos, photographs or other forms of digital data. Moreover, healthcare provision structure is distributed and information is not consolidated among hospitals, primary care and casualty departments. Linking federated databases requires computing effort and complex structures. Medical information is far from ‘standard’. It is often stored in mutually incompatible formats and standards are neither complete nor universally accepted. Even the use of a standard protocol may not imply that independently derived data representing a specific ‘the same’ piece of information will be identical. Tuning and quality of equipment and expertise of the staff all affect the final results.

In the medium term, it is reasonable to expect that most of the services in healthcare will use computer-based resources to store, process and share patient information. Technologies are converging to a mature status and high-bandwidth communication networks are being deployed among healthcare centres throughout Europe, although, of course, there are still differentials between member states. A new key enabling technology is the grid.

1.2.2. Introduction to Grid

Grid computing aims at the provision of a global ICT infrastructure that will enable a coordinated, flexible, and secure sharing of diverse resources, including computers, applications, data, storage, networks, and scientific instruments across dynamic and geographically dispersed organizations and communities (known collectively as Virtual Organizations or VOs). Grid technologies promise to change the way organizations tackle complex problems by offering unprecedented opportunities for resource sharing and

collaboration. Just as the World Wide Web transformed the way we exchange information, the grid concept takes parallel and distributed computing a major step forward towards what is sometimes called 'utility computing', providing a unified, resilient, and transparent infrastructure, available on demand, in order to solve increasingly complex problems.

Grids may be classified into computational grids, data/information/knowledge grids, and collaborative grids. The goal of a computational grid is to create a virtual supercomputer, which dynamically aggregates the power of a large number of individual computers in order to provide a platform for advanced high-performance and/or high-throughput applications that could not be tackled by a single system. Data grids, on the other hand, focus on the sharing of vast quantities of data. Information and knowledge grids extend the capabilities of data grids by providing support for data categorization, information discovery, ontologies, and knowledge sharing and reuse. Collaborative grids establish a virtual environment, which enables geographically dispersed individuals or groups of people to cooperate, as they pursue common goals. Collaborative grid technologies also enable the realization of virtual laboratories or the remote control and management of equipment, sensors, and instruments.

From the original experiments investigating possibilities offered by broadband networks, grid technologies have entered into a phase where production capabilities are available, e.g. NASA's Information Power Grid, CERN's DataGrid, or NSF's TeraGrid, to name a few. However, the vision of large scale resource sharing has not yet become a reality in many areas. This can be attributed mainly to the lack of commonly accepted standards, as well as to the diversity and fragmentation of available grid middleware, tools and services. The Global Grid Forum (GGF), with participants from industry, research, and academia is the main body driving global standardization efforts for grid services, protocols, and interfaces.

According to a recent survey of twenty European grid projects, the most widely used middleware is the Globus toolkit followed by Unicore. Over the last two years, however, the Globus toolkit, which has been originally designed for the needs of High Performance Computing (HPC) resource sharing in the academic community, has undergone a significant shift towards the adoption of a service-oriented paradigm, and the increasing support for and utilization of commercial Web Services technologies. The Open Grid Services Architecture (OGSA) was a first effort in bringing grid technologies and Web Services together. The recent decision of GGF to base the implementation of OGSA on the forthcoming Web Services Resource Framework (WSRF), currently standardized by OASIS, is a further significant step in this direction and will allow the utilization of standard Web Services technologies, which enjoy large scale industry support, for grid computing.

Future developments of grid technologies will be characterized by a full adoption of the service-oriented paradigm and Web Services technologies, a complete virtualization of resources and services, and the increased utilization of semantic information and ontologies (cf. Semantic Grid). Significant efforts will have to be undertaken in order to provide appropriate high-level tools and environments that hide the complexity and reduce the costs of grid application development. The availability and adoption of advanced security standards, support for Quality of Service and the establishment of associated grid business models and processes, will be pre-requisites for large scale adoption of grid technologies.

1.2.3. HealthGrids

Healthgrids are grid infrastructures comprising applications, services or middleware components that deal with the specific problems arising in the processing of biomedical data. Resources in healthgrids are databases, computing power, medical expertise and even medical devices. Healthgrids are thus closely related to eHealth.

Although the ultimate goal for eHealth in Europe would be the creation of a single healthgrid, i.e. a grid comprising all eHealth resources, naturally including security and authorization features to handle subsidiarity of independent nodes of the healthgrid, the development path will mostly likely include a set of specific healthgrids with perhaps rudimentary inter-grid interaction/interoperational capabilities.

The future [13] evolution of grid technologies addresses most precisely problems that are very appropriate for healthcare. Healthgrid applications are oriented to both the individualized healthcare and the epidemiology analysis. Individualized healthcare is improved by the efficient and secure combination of immediate availability of personal clinical information and widespread availability of advanced services for diagnostic and therapy. Epidemiology healthgrids combine the information from a wide population to extract the knowledge that can lead to the discovery of new correlations between symptoms, diseases, genetic features or any other clinical data.

The following issues are identified as key features of healthgrids:

- Healthgrids are more closely related to data, but hospitals are reluctant to let information flow outside hospital bounds. For a large-scale deployment of healthgrids, and thus for opening an attractive business, it is important to leverage security up to a trustworthy level of confidence that could release a generalized access to data from the outside (see also below). Although data storage remains the responsibility of the hospital, many business opportunities can arise from data sharing and processing applications. Federation of databases requires computing effort and complex structures.

- Management of distributed databases and data mining capabilities are important tools for many biomedical applications in fields such epidemiology, drug design or even diagnosis. Expert system services running on the grid must be able to interrogate large distributed databases to extract such knowledge as may lead to the early detection of new sources of diseases, risk populations, evolution of diseases or suitable proteins to fight against specific diseases.

- Security in grid infrastructures is currently adequate for research platforms, but it must be improved in the future to ensure privacy of data in real healthgrids. Encrypted transmission and storage is not sufficient, integrity of data and automatic pseudo-nymization or anonymization services must be provided to guarantee that data is complete and reliable and privacy leakages can not appear due to unattended use of the resources. Biomedical information must be carefully managed to avoid privacy leakages. Failure on privacy in biomedical personal information causes irreparable damage, since there is no way to retrieve the situation. Secure transmission must be complemented with secure storage with strictly controlled authenticated and authorized access. Automatic pseudo/anonymization is necessary for a production stage.

- Robustness and fault tolerance of grids fits very well to the needs for ‘always on’ medical applications. Grid technologies can ease the access to replicated resources and information, just requiring the user to have a permanent Internet connection.
- Research communities in biocomputing or biomodelling and simulation have a strong need for resources that can be provided through the grid. Compliance with medical information standards is necessary for accessing large databases. There are many consolidated and emerging standards that must be taken into account. Complex and multimedia information such as images, signals, videos, etc. is clearly a target for grid and is more sensitive to data formats.
- Finally, flexibility is needed for the control of VOs at a large level. The management of resources should be more precise and dynamic, depending on many criteria such as urgency, users’ authorization or other administrative policies.

Today most grid applications for health follow the classic high-throughput approach. Numerical simulation of organs obtained from patients’ data [14, 15] is used to aid understanding and to improve the design of medical devices. Patient-customized approaches can be found at research level in areas such as radiotherapy, cranio-facial surgery or neurosurgery.

Other areas of application deal with large-scale information processing, such as medical imaging. Breast cancer imaging has been the focus of several successful grid projects [16, 17] and eHealth projects suitable to migrate to grid [18]. These efforts have concentrated on federating and sharing the data and the implementation of semi-automatic processing tools that could improve the sensitivity and specificity of breast cancer screening programs. Much effort has been invested to reduce the information needed to be exchanged and to protect privacy of the information.

The concept of a patient-centric grid for health has also been explored [19]. The main aim of this approach is to make the information available to the whole health community (patient, relatives, physicians, nursery), considering access rights and language limitations.

Bioinformatics is the area where grid technologies are more straightforwardly introduced. The main challenge faced by bioinformatics is the development and maintenance of an infrastructure for the storage, access, transfer and simulation of biomedical information and processes. Current efforts on biocomputation [20] are coherent with the aims of grid technologies. Work on the integration of clinical and genetic distributed information, and the development of standard vocabularies, will ease the sharing of data and resources.

1.3. DEFICITS, OPPORTUNITIES AND REQUIREMENTS FOR INDUSTRY

Grid technology is still a ‘moving target’. The rapid evolution of platforms and versions leads to major difficulties in the development of applications to a production stage. Industry has to define and exploit business models on the grid, but it needs more stability and standardization on grid infrastructures before it can develop viable business models. Indeed, current grid middleware lacks several components that would be necessary for business exploitation:

- Grid middleware lacks reliable and complete accounting services that can clearly identify providers, consumers and resource usage in a scenario in which a wide range of heterogeneous resources, owned by different entities, are shared. The whole economics of the grid is still to be worked out.
- Current efforts at robustness and fault tolerance have improved middleware reliability, crucial for exploitation in healthcare applications, but it is still not at a production level.
- Security and privacy models for the grid are not adequate for applications that can be certified by end users and health authorities.
- Reliable benchmarking must be performed to certify that all components perform with the quality of service and robustness that healthcare applications require. Middleware certification is even more important in healthcare applications, taking account of possible impact on patient health, and on legal and ethical considerations.
- Grid exploitation may encounter a serious problem in the use of software licences. Current software licenses usually prevent its use in grid environments in which the computers and the users are not clearly defined. New licence models will appear with the development and new business models. Until then, successful applications should better focus on the exploitation of own or public licence software.
- Before developing business-relevant applications, there is a clear need of a production infrastructure in which applications can be run. Many services can be implemented and tested and deployed for validation. Validation of healthcare applications can then be undertaken on such a platform, although final exploitation can be deployed on separate resources.

There are at least three scenarios in which healthgrid technologies can be successful from a business point of view:

- Consolidation of resources: Integral solutions for applications, data and resources at centre and region are needed. (Current distributed database technologies do not yet offer the level of interoperability or the capability of providing other resources, apart from data, to make this a reality.)
- Efficiency leveraging: Ideal applications from the business point of view are those requiring large peak resources followed of inactivity periods.
- Reduction of production costs in applications where the return on investment is low but the social impact can be high. Joint public-private consortia may succeed in healthcare goals, such as rare disease drug discovery, that do not offer economic profit but may benefit significant populations. Providing resources for *in-silico* experimentation may stimulate the discovery of affordable, effective drugs for neglected diseases.

There is a long way to go before exploitation, and industry should assist and guide research on healthgrids in order to profit from reliable and interesting results.

1.3.1. The pharmaceutical Industry

The convergence of biotechnology and ICT are providing novel drug development methods, as a consequence of which pharmaceutical industry requires enormous amounts of computing capacity to model, discover and test interactions of drugs with receptors, and thus to decide which should be synthesized and tested.

Drugs that come to market are the results of several years of research. There is a need to accelerate the development process and reduce time to market for new drugs. One way this can be done is by increasing the number of calculations processed for docking analysis. Computation with virtual compounds produces a large volume of information which is hard to analyse both in terms of time and cost. These results must be stored for further analysis, creating the need for mechanisms to share securely and privately the information among federated databases.

In fact, there is an overload of information, but there is a lack of interoperability between different applications and data sources. Current tools cannot handle the results in an effective way, nor do they extract enough knowledge. This means that there is a lot of wasted information and unused results. Collaboration between scientists and researchers from industry is crucial for success in the pharmaceutical industry.

The next step in drug development is to integrate phenotype with genotype information and environmental factors, leading to 'personalized' drugs, leads to the need for on-demand analysis, requiring more resources and tools.

1.3.2. Medical Information Technologies Industry

Most important challenge in medical IT is the need to reach the maximum degree of interoperability, seamless access and processing of distributed electronic medical information. This challenge, based on the electronic patient record, requires the interaction of industry, research and standardization bodies.

These aims are not achievable solely through the integration of distributed databases. First, not all the information is comparable or compatible, not only in terms of format, but also due to differences in procedures, devices, human intervention or other factors. Federation of data must be achieved at a semantic level for interoperability to become a reality. Secondly, much medical information is not currently processed electronically. Vital signs, perception tests and laboratory analyses are usually captured and stored, even in digital form, but not available for further processing through lack of connectivity or incompatibility. Interfaces for equipment and storage formats are currently being developed and standardized, but take-up is slow.

The integrated electronic patient record will require a significant increase in resources for storage and processing, so that clinical institutions will certainly have to consider sharing computing services. Interoperability among devices will be a strict necessity. New services may then be made available on this infrastructure, including clinical aid applications, such as computer aided diagnosis, image processing, vital sign feature extraction, clinical output evaluation or even simulation.

1.4. DEFICITS, OPPORTUNITIES AND REQUIREMENTS FOR HEALTHCARE AND MEDICAL RESEARCH

The situation in 'routine' healthcare is very different from that of medical research. The main target for healthcare-oriented grids is to access large amounts of data securely and efficiently, with occasional need for high processing power. Medical research however deals with a wider set of issues. Computing resources, knowledge extraction from very large databases and means for solving grand-challenge problems are important concerns in different applications.

Biocomputing medical applications are one family of "killer applications" for biomedical grid research. The maturity of genetics and biomedical technologies brings them closer to medicine, and grand-challenge computing problems of biocomputing are currently being migrated to grid [20]. Biomedical modelling and simulation is another important arena for grid applications. Biomedical models are highly coupled, involve complex physics and require intensive numeric computing. Coupling the models is essential to achieve a realistic simulation that could give useful feedback to medical science and medical instrument technology. The long-term aim of the "virtual human being" can only be technologically feasible with very large computing resources. National e-science infrastructures may not be sufficient for such a large goal.

Healthcare grids' key issue is to be provided with the proper services for querying, storing and retrieving multimedia medical data from a data grid. Privacy Enhancing Techniques must be considered to allow medical data access from outside the borders of the medical database holders. Coordination with EPR initiatives is fundamental to avoid replication of effort and to ensure the applicability of results. Connection to medical information systems such as Hospital Information Systems, Picture Archiving Computer Systems, Radiology, Laboratory and Primary Care Information Systems will be very important for access to data, while the development of libraries of services will ease the process of building up medically-relevant applications.

Last but not least, the grid is an important opportunity for the spreading of knowledge in developing countries. Sharing medical data, procedures, services and expertise with research centres in those countries where these tools are not widely available may be a first step towards improving healthcare delivery and, at the same time, medical expertise.

1.4.1. Medical information processing

The ultimate goal of biomedical and health informatics is to support the continuity of individualized health care from prevention to rehabilitation. However, integration of informatics and technology tools in clinical practice has progressed far slower than expected, and the communication gap between clinicians and informaticians is still significant.

The difficulties in widely implementing research results have been discussed extensively in recent years. Some factors arise both in research and in implementation, and are related to intrinsic difficulties in medical informatics, such as the complexity of information and organization, human factors, and diversity of cultures, especially in relation to financial and business aspects. For example, where specific algorithms have been developed and applied efficiently to a very narrow range of specific cases, extended

validation would be necessary before use in healthcare. A broad biomedical and health informatics platform, enabling interconnection and integration of resources, while supporting evidence-based medicine and validation of research results, would thus contribute to the acceptance of technological developments in the medical world.

A key point in medical informatics is the management of medical information, and the efficient and quality certification of information and knowledge flow between all the players involved in the health delivery process. Previously obtained knowledge has to be captured and organized in a structured form in order to be retrieved in the right context and in an organized manner, thus contributing both to educational and to research purposes, while simultaneously supporting new healthcare diagnoses and the generation of new medical knowledge.

The basic strategies and scope of medical informatics has also been reconsidered in the context of its relationship with bioinformatics. A potential for collaboration between the two disciplines could involve topics such as the understanding of molecular causes of disease, the efficient disease management of chronically ill patients and the integration of clinical and genetic data. An interesting perspective is the combination of pervasive computing, facilitating the transmission and collection of biological data on a real-time basis outside a clinical setting, with the biomarkers and other indicators, resulting in a new phase for home care systems.

Concluding, there is an emerging need for exchange, synthesis and ethically-sound application of knowledge - within a complex system of interactions among researchers and users, in an interdisciplinary environment - to accelerate the capture of the benefits of research through more effective services and products, a strengthened health care system and ultimately better health. These requirements support the applicability of grid technologies, which provide the functional and architectural framework to facilitate such synergies while addressing the underlying ethical and privacy issues.

1.4.2. Biomedical modelling

Research in the physics of human biomedical processes has made much progress recently. The consolidation of accurate and complete simulation tools for many engineering processes has contributed to the development of biomedical models of the structural dynamics, fluid dynamics, chemical processes, and electric potential propagation which describe with high degree of accuracy the physics of many organs and tissues.

All these models are generally applied to restricted small areas or do not reach the desired accuracy due to the large memory requirements that fine meshing for numerical analysis requires. Moreover, the complexity of human biomedical models lies on the high degree of coupling among the chemical, structural, magnetic and electric processes. This complexity requires further improvement of biomedical models and use of unprecedented computing and memory resources.

Thus, the evolution towards the “virtual human” model is a major long-term aim of biomedical computing. Tackling such a problem requires the close cooperation of many groups, sharing computing resources, models and data. Accurate medical models are not freely available, and usually represent the most valuable capital of a research centre. Means for cooperating without compromising Intellectual Property Rights (IPR) are necessary.

1.4.3. Genomics

Genome-wide sequencing projects have been completed for many organisms, including *Home Sapiens* [4] and *Mus Musculus* [5]. This reversed the conventional approach to biomedical discovery, in which understanding a certain biological function required identification (and sequencing) of one or more genes involved in that function: the current situation is that thousands of genes have been sequenced but still wait for any functional information to be assigned to them.

The fact that genes of unknown function represent over 70% of all genes, suggests that current comprehension of most biological and pathological processes is far from complete. As a consequence, new technological platforms that take advantage of the genome sequence information to explore gene function in a systematic way are evolving at an incredibly fast pace. Application of microarray technology [6] to more translational research fields, such as cancer research, has revealed its enormous potential as a diagnostic support tool in clinical management. Recent work has shown that it is possible to exploit gene expression profiling of tumour samples to define sets of genes (signatures) whose expression correlates, positively or negatively, with specific clinical features, such as metastasis-free survival in breast cancer [8], and response to therapy [7]. Other types of massive datasets currently generated in genomics projects include: protein expression levels, measured by proteomic screening; protein-protein interaction datasets in various organisms; protein structure data; genomic sequencing of additional organisms, and comparative genomics; sequence polymorphisms in human populations, mutational analysis in human cancer and in hereditary diseases; loss-of function analysis in various organisms by small interfering RNA (siRNA)-based approaches [9].

As a consequence of these genomic research activities, biomedical databases are continually and exponentially increasing in number and size, together with bioinformatic tools that extract information from them. Major research laboratories (e.g. NCBI in the USA and EBI in Europe) collect and regularly update information. These data can be analysed using a web interface to a number of well-known applications (mainly data mining programs), that are CPU intensive and require large amounts of I/O.

Often a data analysis process requires the pipelining of results through different applications. The retrieval of results from a web-based application is an awkward and error prone task involving ‘screen scraping’, electronically capturing the content of the screen. This is further complicated by the changes to web interfaces. Even though the computing resources dedicated to any single researcher are limited, concurrent access to the web applications leads to the congestion of the major resource centres. Hence, biologists prefer to download the database files and to process them locally.

This has two major consequences: every single researcher has to track the database update process to keep his/her copy of the data up-to-date; the massive download of huge amounts of data worsen the performances of the web site and of the applications of the download centre.

Another relevant aspect is the lack of a standardization of the published databases: cross-referencing of data is made difficult (if not impossible) by redundancies and incoherencies, there is neither standard query language, nor central management of data,

and finally, different processing applications require the same data in different formats. Data quality control and, accordingly, confidence in the results obtained is poor.

A grid infrastructure is expected to overcome many of the drawbacks of the existing web-based approaches to genomic data handling and mining, by offering new services such as the transparent access to computing resources for CPU-intensive processes which is important due to the high computing demand of the biomedical problems. Another important task is the creation and management of shared, coherent relational databases to resolve incoherencies and inconsistencies in the actual databases and to provide the infrastructure to gather data coming from genomic experiments, providing the means to manage replicated copies of the data files and their coordinated updating.

Finally, database security (all aspects concerning data confidentiality), data transfer channel encryption and, last but not least; user authentication and authorization must be considered as a main requirement.

1.5. REFERENCES

1. "LHC Computing Grid Project", http://lcg.web.cern.ch/LCG/default.htm.
2. "The CrossGrid Project", Technical Annex http://www.lip.pt/computing/projects/crossGrid.
3. "Project Presentation" The DATAGrid project, http://www.eu-dataGrid.org.
4. "Human Genome Resources", http://www.ncbi.nlm.nih.gov/genome/human.
5. "Mouse Genome Resources", http://www.ncbi.nlm.nih.gov/genome/mouse.
6. Z.G. Goldsmith and N. Dhanasekaran "The Microrevolution: Applications and impacts of microarray technology on molecular biology and medicine (Review)". Int J Mol Med. 13:483-495, 2004.
7. "Current Progress in the Prediction of Chemosensitivity for Breast Cancer," Daisuke Shimizu, et. al. Breast Cancer 11:42-48, 2004.
8. M. J. van de Vijver and Others "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer", N Engl J Med. 347: 1999-2009, 2002.
9. Derek M. Dykxhoorn, Carl D. Novina & Phillip A. Sharp, "Killing the Messenger: Short Rnas that Silence Gene Expression", Nat Rev Mol Cell Biol 4: 457-467, 2003.
10. "The Emerging European Health Telematics Industry". Deloitte & Touche, Feb, 2000. Health Information Society Technology Based Industry Study – Reference C13.25533.
11. "Medical practice websites enhance patient care" Ehealthcoach, December 2002.
12. M. Schmidt, G. Zahlmann "What are the benefits of Grid Technology for a health care solutions provider?", Siemens Medical Solutions, Proc. HealthGrid conf. January 2003.
13. "Next Generation Grid(s)", European Grid Research 2005 – 2010 Expert Group Report, 16th June 2003, http://www-unix.Gridforum.org/mail_archive/ogsa-wg/pdf00024.pdf.

14. J.Fingberg, et. Al. “Bio-numerical simulations with SimBio”, Physikalische Methoden der Laser- und Medizintechnik, pp. 114-120, VDI Verlag, 2003.
15. “Grid-Enabled Medical Simulation Services” GEMSS, http://www.gemss.de
16. R. McClatchey, et. Al. “The MammoGrid Project Grids Architecture” CHEP’03, San Diego March 24th 2003.
17. “The eDiamond Project”, http://www.ediamond.ox.ac.uk/
18. “Magnetic resonance imaging for breast screening (MARIBS)”, Official Web site of the project http://www.icr.ac.uk/cmagres/maribs/maribs.html
19. “myGrid: personalised bioinformatics on the information Grid”, Robert D. Stevens, Alan J. Robinson and Carole A. Goble - Bioinformatics Vol. 19 Suppl. 1 2003.
20. A. Sousa Pereira, V. Maojo, F. Martin-Sanchez, A. Babic, S. Goes, “The Infogenmed Project”, ICBME 2002: December 2002.
21. S. Nørager, Y. Paindaveine, “HealthGrid Terms of Reference”, Version 1.0, 20th September 2002

2. A Compelling Business Case for Healthgrid

Although both healthcare in general and the use of IT to support the development of effective treatment, delivery and management of healthcare are top priorities in many countries, there are many areas competing for investment. The benefits of using even basic IT to provide high quality information and decision support to clinicians and patients are intuitively very significant. In other industries – airlines, automotive, banking, defence, and manufacturing – IT has underpinned productivity, quality, security and improved product performance for many years. However, progress in even basic IT has been patchy and slow in the healthcare industry; there are few high quality, well documented business cases with results and very few for IT implementation at large scale. There are even fewer cases that demonstrate the benefits of dramatically new IT technologies (like grid) in innovative areas of healthcare such as genetics, imaging, or bioinformatics. Therefore in applying for funding and prioritization of resources to continue to develop healthgrid applications, I t is vital that a clear and highly compelling business case is created that acts on all the benefits levers of healthcare.

2.1. THE GROWING IMPORTANCE OF IT IN DELIVERING EFFICIENT, HIGH QUALITY HEALTHCARE

The advent of healthgrid applications, even at the research stage, coincides with a crucial period of investment and experimentation in IT for healthcare. The main drivers for this shift in the pace and levels of investment include:

- Increased understanding of the impact of medical errors on patient safety and the resulting cost of care. IT's basic value proposition includes the ability to regulate processes and scale information "written" once to many uses and contexts.
- Demand for healthcare is outstripping resources at all levels, driven by an ageing population in most countries, living longer but with access to an increasingly sophisticated armoury of tests, surgical interventions, medications, etc. IT has the power both to add to the armoury of clinical tools and to reduce costs through efficient operation with fewer process steps, less wasted activity (tests, unnecessary prescriptions, etc.) and better utilization of disparate resources

The coincidence of growing capability in grid technology with this increase in investment has its drawbacks. First, there are many strategic and investment plans being made at local, regional and national levels that take no account of emerging technologies like grid; even if the first truly useful healthgrid applications will not be ready for several years, this is within the planning horizons and budgeting horizons of the Public Sector. Secondly, as IT is introduced into everyday healthcare, custom and practice is changing on how care is delivered. Such change in the clinical world is very significant – for instance rationalising the outpatients' process to a single series of steps supported by sharing of electronic data, in all hospitals within a region, is a considerable change. Overlaying such

serious changes with the completely new capabilities of grid will simply add to the challenges. And in healthcare, change can take time to embed – a recent study in the USA showed an average 17 year delay in adopting widely proven practices in healthcare.

And are healthgrid technologies being anticipated in the many eHealth strategies being created around Europe? In short, the answer is "No". Very few senior health managers in Europe understand the potential or the practicalities of healthgrid; in general, they are certainly not embedding their strategies with even link points to take advantage of grid in the future. The risk, therefore, is that it will be even harder than it should be to take advantage of healthgrid capabilities over the next 5-10 years – unless the potential is understood quickly and strategies adapted accordingly.

2.1.1. Measuring success – Quality, Access, Cost

So as the business case for healthgrid is so critical, how can it be articulated in terms that senior health managers can understand? One suggestion, based on the work of the European Commission's eHealth Unit, is to define the benefits across three categories, specifically the impact on:

- Raising the ***quality*** of care. Here factors include the ability to make faster decisions or interventions; fewer medical errors; more informed decisions or diagnoses;
- Improving the ***access*** of patients to care. Sources of benefit might include the extension of lengthy or complex tests and diagnoses to a much larger number of patients through increased capacity; the provision of new tests or diagnoses that simply could not be made using traditional approaches at reasonable cost;
- Reducing the ***cost*** of care. This is a complex issue for healthgrid since it is an emerging technology creating opportunities for new procedures and tests that may initially add to short term costs; however, there may be sources of benefit from such short term investments leading to long term reductions in cost of care, e.g. as disease is identified earlier and prevented.

It is important to recognise that rarely do these three factors appear independently – for instance it may be that improving the access to care via new tests also impacts the long term cost of treating either chronic diseases or immediate palliative care.

Casting the benefits of healthgrid applications against these three factors has a great advantage in creating compelling business cases for senior health management – and politicians – because it allows them to see the benefit in the comprehensible terms of managing day to day healthcare outcomes and budgets. Creating such resonance is critical to gaining priority and share of resources / budgets.

2.2. WHY INVEST IN HEALTHGRID APPLICATIONS AND SERVICES?

Not only does the modern healthcare management team have many choices for investment of their time and money in traditional sources of patient care improvement, but they also have a bewildering array of IT support that can be purchased. So why, in such an

already complex, packed marketplace, should the relatively new, often untried, grid technologies be given any priority at all?

2.2.1. Critical opportunities for distributed computing approaches

Of course, not all healthcare informatics problems will be remotely suitable for a grid solution. There are "sweet spot" problems where the advantages of grid approaches will outweigh the potential drawbacks of a relatively untried and new technology. The characteristics of clinical problems that could have significant advantage from distributed computing-type solutions include:

- analyses that require dynamically assembled data-sets and investigation routines; for instance, genetic-related investigations where the initial analysis may raise the need for further data sets to be added to give better, more representative results from analysis;
- processes of analysis and data assembly that cross organisational boundaries, where the ability to distribute both the data and analysis without recall to the normal "data process" flows is key. Again, medical research, or in future patient-centric analyses, are probably two areas where the utility of the grid will be highest;
- huge scale analysis, that requires a scalable infrastructure to deal with the potentially massive quantities of data to be both assembled and analysed. This leads us again to imaging and genetic analysis as potential opportunities;
- dynamic grouping of healthcare professionals for review / analysis of diagnosis or research results, such that different "expert teams" can be assembled without a formal organisation structure (indeed, across organisation structures). Feedback from clinicians on existing grid health projects indicates a strong need to enhance collaboration on a daily basis between communities, removing their reliance on conferences to achieve this.
- Further benefits may be realised through the pooling of resources, whether it be the sharing of training cases to enable smaller clinics to benefit from the knowledge available in larger hospitals, or the sharing of compute resource to reduce the local investment on IT.

Therefore, in summary, there seems to be an advantage available from using grid approaches where the clinical problem requires a scalable, flexible infrastructure that can work across normal organisation and process boundaries.

2.2.2. Impact on wider patient access to care

The key value that grid approaches can bring to increasing patient access is to make possible new analyses of data, whether for individual patient care or group research, that traditional computing approaches cannot provide. The principal features of problems that suit such approaches are those involving huge quantities of data requiring iterative,

repetitive analyses – typically image diagnosis, genetic diagnosis are current problems with these features.

2.2.3. Impact on raising the quality of care

The application of grid technology could allow better analysis of patient data – by dynamically assembling data sets for comparison; by using discoverable publishing to improve access to previously difficult to find data; by allowing self-describing data sets to be more freely used, therefore raising the quality of the resulting analysis. The areas where this could have greatest benefit may include rarer instances of disease diagnosis, complex image manipulation, and even temporal comparisons of patient information to assist with determining change.

2.2.4. Impact on reducing the cost of delivering care

From all the discussions, it seems that the main, direct advantage that healthgrid could provide in its application is to create a high degree of utilisation of infrastructure and computing power, while still allowing a very flexible, scalable infrastructure to be applied that could deal with dramatically varying demand. Indirect cost advantages would derive from two main categories - first is the maintenance and effort put into IT, which in a grid solution should be, in theory at least, easier to manage since data is self-discoverable and infrastructures are managed in a more flexible way. Secondly, there are all the potential cost savings in the delivery of care stemming from the improved quality and increased access to care that grid approaches offer.

2.3. BARRIERS TO ECONOMIC, RAPID IMPLEMENTATION

While there may be some very serious advantages to be had from applying healthgrid technologies to suitable problems, there remain significant barriers to implementation. They can be summarised into 3 main areas:

2.3.1. Governance and accountability

On many levels, the healthgrid does not match current governance models and tried and tested processes. As one example, research conducted using grid approaches does not necessarily have the same degree of independent scrutiny and open accountability to which traditional peer-reviewed research is routinely subjected. In fact, the very nature of dynamically assembled, self-discoverable data sets and analyses means that such scrutiny is probably impossible. Secondly, the entire area of trust (particularly in data) is critical to the widespread acceptance of grid approaches in health. This trust issue ranges from building diagnoses or clinical evidence on data collected, maintained and shared by organisations or individuals outside of the originator's span of control to accepting that grid applications must be shared across organisations' infrastructures.

2.3.2. Quality of Service and speed

With any distributes system, where all pieces of the infrastructure (computing devices, data stores and networks) are not under a single span of control, the issue of the availability of resources, and the maintenance and reliability of such resources, is critical. Add to this reliability issue the potential contention for resources that massive data manipulation could experience, and the quality of service (guaranteed speed of response) could be frequently compromised. There are approaches for managing this problem, but most increase the cost or require heavy structured governance processes.

2.3.3. Incomplete models & technologies

Much of the grid technology has only been applied in research fields where human lives do not literally depend on it or the decisions made on its output. Before life-critical applications can be trusted, many more examples, pilots and controlled trials will be necessary. Whilst there have been significant advances in standards for the integration of healthcare systems, it is evident that further work is needed in order to take this to the dimension of 'the big joined-up healthcare' approach.

2.4. IN CONCLUSION

The healthgrid is potentially a significant addition to the armoury of tools health professionals and researchers can use to improve quality, increase access and reduce the cost of healthcare. However, significant progress is required on the governance, quality of service and operational models for grid technology before it can become a widespread tool in daily use.

3. Medical Imaging and Medical Image Processing

3.1. **MEDICAL IMAGING**

Medical diagnosis and intervention increasingly relies upon images, of which there is a growing range available to the clinician: X ray (increasingly digital, though still overwhelmingly film-based), ultrasound, MRI, CT, PET scans etc. This trend will increase as high bandwidth systems for picture archiving and communications are installed in large numbers of hospitals (currently, primarily in large teaching hospitals). More than patient data, the medical images by far represent the major amount of information collected for medical data. However, medical images are not sufficient by themselves as they may need to be interpreted and analysed in the context of the patient's medical record (that is the metadata associated with the images).

There are a number of factors that make patient management based on medical images particularly difficult. Medical data are naturally distributed over a number of acquisition sites. Physicians most often have no way to access all the medical records across all of their patients. Patient images often represent very large quantities of data (e.g. 3-D images, time sequences, multiple imaging protocols) with complex structure (clinically and epidemiologically significant signs are subtle including patient age, diet, lifestyle and clinical history, image acquisition parameters, and anatomical/physiological variations). In many cases, no single imaging modality suffices, since there are many parameters that affect the appearance of an image and complementary information is captured by different physical acquisition systems.

Medical data are used in diagnosis, continuing care, and therapy planning. For diagnosis, medical images acquired in a medical centre are usually visualised and interpreted immediately after the acquisition by the radiologist before being sent (often on films) to a physician for second viewing. These two readings normally take place in different offices and possibly even in different sites. For therapy follow-up, even more clinicians may be involved as images taken at different times may have been acquired in different radiology centres and several physicians may need to read them. For therapy planning and assisted intervention, images also need to be accessible from the intervention room.

Picture Archiving and Communication Systems (PACS) deployed in hospitals today address some of the challenges related to medical data management. However they suffer many limitations:

- Often they are disconnected from the Radiological Information System (RIS) carrying the medical records.
- They are often proprietary solutions of medical imaging companies and no open standards exist to ease communication between different PACS.

- They are usually limited to data management inside one health unit (one hospital or at best a federation of hospitals) and are not scalable on a national or international scale.

Manipulating medical data on a large scale also raises the problems of security and confidentiality of personal data. Grid technologies are expected to ease the design of distributed medical information systems in a secured environment. Although grids cannot by themselves resolve the problem of heterogeneity in data formats and communication protocols, they are expected to motivate the establishment of standards in this field.

3.1.1. From medical data acquisition to medical data storage and archiving

Although most recent medical imaging equipment produces digital images, the long term archiving of data is often performed on film only. Medical images represent enormous amounts of data: a single image can range from a few megabytes to one gigabyte or more. The total amount of digital images produced in Europe thus probably exceeds 1000 petabytes each year. The legal aspects concerning medical data archiving vary from country to country in the European Union but the actual trend is towards long term archiving of medical data (about 20 years for any data, up to 70 years for some specific data) and to make the patient the owner of its data.

To ease data storage and communications, the DICOM standard (Digital Image and Communication in Medicine) has been supported by several international bodies and industrial companies. Most recent image acquisition and treatment devices implement the DICOM standard and that eases data exchanges between imagers, post-processing consoles, and archiving systems. However, it does not include all features of RIS for data management and access, nor does it describe archiving strategies dedicated to PACS.

Medical data storage strategies can only be established when considering the access pattern that depends on the use of these data. The legal trend is for patients to have full read access to their medical records. The physicians obviously need access to the data of their own patients, however, any physician should not have access to all medical data owned by any patient. Other communities may in addition have restricted patient data access needs. For instance, researchers may need access to the core data although personal identification may not be needed in every case.

Grids provide a support for the distributed and mass storage of data. Several grid middlewares propose distributed and transparent file systems aggregating many storage resources to offer extensive storage capacity. Several aspects of grids that are still under investigation concern the implementation of data access control and security of data. While remaining internal to the hospitals, data security problems are rather easy to solve however enabling data exchanges between hospitals over wide area networks makes this matter much more complex. Medical data should always be considered as sensitive in general and identifying data should remain strictly confidential. In particular this means that data should only be accessible by authorised users (for sensitive data) or accredited users (for identifying data), often excluding service providers and system managers. Encryption (and thus anonymisation) of data on disk and during network transmission is therefore mandatory; the access to decryption keys being strictly controlled.

3.2. **BUILDING VIRTUAL DATASETS ON GRIDS**

To enable analysis of medical images related personal and clinical information (*e.g.* age, gender, disease status) has to be identified. The number of parameters that affect the appearance of an image is so large that the database of images developed at any single site - no matter how large - is unlikely to contain a set of statistically sufficient exemplars in response to a query related to one of these domains:

- Screening programs: to study the distribution of some diseases at a pan-European scale and to correlate this information with common factors.
- Studies on rare diseases for which limited data is available on any single site.
- Assembling individualised datasets: when studying data from one patient or one particular population, one may need to assemble a comparative epidemiological dataset by selecting data with similar features at a pan-European scale (same gender, age, social category, etc).
- Alarm networks: to detect the spread of some pathologies over national boundaries.

Overcoming the problem of data distribution implies constructing a huge, multi-centre - federated - database, while overcoming statistical biases such as lifestyle and diet leads to a database that may transcend national boundaries. A distributed medical database could be used to assemble *virtual* datasets: *i.e.* datasets assembled on demand from various data sources belonging to different regions and countries for a specific purpose. For any medical condition, there would be huge gains in using virtual datasets so long as that (federated) database appears to the user as if it were installed in a single site (i.e. a single logical dataset). Such a geographically distributed (pan-European) database can be implemented using grid technology, and the construction of a prototype would enable a study of the suitability of grid technologies for distributed image analyses.

The medical image analysis community require transparent access to collections of image data that may reside in a number of locations inside and outside their hospitals and in a number of different formats. It is crucial in deploying any software solution to this community that the complexities of those technologies that support virtual datasets are hidden from the users and that the essentials of their requirements are satisfied firstly ‘in the large’. Only then will the systems analysts and designers responsible for deploying the enabling technologies gain the commitment from that user community to develop the required infrastructure to satisfy the requirements ‘in the small’. The solution offered for virtual datasets must be sensitive to the over-riding issues of data protection and ownership (by individuals, by medic and hospitals), data security, medical anonymity and ease of access to the data.

Heterogeneity of image data is one headache in constructing grid-based virtual databases of images. It will be necessary for any usable grids medical image implementation to integrate multiple datasets be they database-resident or file-resident. To this end the requirement for discovery of and interaction with heterogeneous data schema needs to be resolved, potentially through the use of high-level meta-data abstractions (possibly using ontologies) of each different dataset. Careful consideration must be given to

semantic heterogeneity too : different data systems may well refer to the same data item with different names or different items with the same name. Identification of patients on a large scale is a critical problem too: usually, each hospital internally uses its own individual identification mechanism. The need for ensuring the patient's privacy makes it even more difficult.

The issue of handling annotation is one particular problem in building virtual datasets. Annotation can be added to image data in several forms: in radiologists drawing regions of interest on medical images (*e.g.* to denote areas for further study, computer assisted detection (CADe), biopsy etc), in radiologists writing medical notes alongside images, in technicians supplying written 'conditions' under which the image was recorded, and in annotation on sets of images, on a particular study or on actual patient records. Any virtual dataset would need to cater for these different levels of annotation and allow queries to be executed against the semi-structured and/or structured annotation. Clearly there is a need for standardisation in image annotation in the medical community (if possible) to enable query resolution.

Any successful medical data system must also provide links between image data and non-image data such as biopsies, medical treatment records and patient meta-data. Furthermore links between different forms of image (PET, CT, X-ray, mammograms) also need to be resolved as do the more general data issues such as privacy, security and appropriate role-based access.

3.2.1. Database indexing

One of the most important aspects in building large-scale virtual image datasets is the ability to perform queries in a transparent and efficient manner. The most standard way to formulate these queries is to express conditions on attributes associated to images. Nevertheless, these approaches are very intensive both in terms of computational power and data manipulations. An intermediate level between direct image access and requests using only metadata consists in querying image features. This kind of queries relies on the computation of indexes describing either global properties of images or local properties of individual image regions, salient objects or topological relations between these objects. These indexes can largely contribute to the acceleration of Content-Based Image Retrieval (CBIR) since standard database operators can be used, and the direct access to raw image data can (most of the time) be avoided.

However, the indexing of medical images has not retained the attention of researchers as much as the indexing of photographic images thus far and the selection of pertinent indexing methods, adapted to different kinds of images is a difficult and a very application-dependant task. There is therefore a real need for standardising the representation of these indexes, but also the description of algorithms used for their computation. Some of the key issues that have to be solved in a widely distributed image database environment are:

- The deployment on different geographical sites of indexing algorithms / libraries, and the management of new algorithms (or of algorithm version evolution).
- The indexing policy: which algorithms have to be applied, and which parameters are adapted for the different images? When is it necessary to (re)launch the

indexing? What happens when new images/algorithms are integrated to the distributed environment?

- The “traceability” of indexes: it is crucial for having a pertinent query scheme to be able to know which algorithm, in which version, and with which parameters, was used to compute a set of indexes.
- In the case of complex processing, several stages can be chained: the data produced by a given algorithm can be used as input of another stage of processing. The distributed system must include standardised ways to describe these dependencies and must be able to launch the necessary computations in the case of insertion of new data, or when a new algorithm is made available.

The possibility of handling the security of these indexes at different levels may be needed: in the same way that personal (nominative) data have to be anonymized for certain categories of users, the image data can itself require security, particularly when it permits patient identification (*e.g.* the 3D scanner of a face). However, indexes computed from these image data can be considered as public when they do not leave the possibility of patient identification.

3.3. MEDICAL IMAGE PROCESSING

3.3.1. Image analysis algorithms

Computerised medical image analysis algorithms have been developed for two decades or so. The aim is to assist the clinicians in facing the amount of data by providing reliable and reproducible assistance to diagnosis and therapy. Indeed, the manual processing of 3D images is very fastidious and often error prone. Moreover, 3D medical image interpretation requires a mental reconstruction for physicians and is subject to large inter-operator variations.

Although image processing algorithms can provide accurate quantitative measurements (*e.g.* the measurement of the heart left ventricle ejection fraction from dynamic image sequences) or can accomplish some tasks that are not feasible by hand (*e.g.* accurate registration of multi-modal images), the reliability and the responsibility issues remain key showstoppers to their large scale development. Algorithm validation is often made difficult due to the lack of provable theory in order to compare with processing results and their development tends to be limited in scale.

Some medical image analysis algorithms are also very computing intensive (e.g. stochastic algorithms like Markovian models, Monte Carlo simulations). Therefore, some algorithms that are known to produce better results are not used in practice due to a lack of computing power. Given that a sufficient amount of computing resources is available, parallelization is often a means to significantly speed-up these algorithms.

Grid technologies will not only provide access to large amount of data for testing. It will also enable image processing communities to *share common datasets* for algorithm comparison and validation. They will offer an access to large processing power suited to

processing full datasets in reasonable time, compatible with the needs for experiencing new algorithms. They will also ease *the sharing of algorithms* developed by different research groups thus encouraging comparative studies. For all these reasons, grid technologies are expected to boost the production of medical image analysis algorithms and to facilitate their quality improvement.

3.3.2. Registration

Registration techniques have encountered considerable success in the medical image processing community not only as they permit the production of average models but also because they ease the comparison of image data coming from multiple sources. Registration may be intra-patient (when registering data coming from a same patient but acquired at different time and/or on different imagers) or inter-patient (when comparing data from different patients). It can be mono-modal (when registering images acquired using the same image modality) or multi-modal. The matching criteria used to perform optimisation depends on the kind of registration performed. But there is another categorisation of registration algorithm that has a largest impact on the optimisation procedure and its computational cost: one often differentiates between *rigid* and *non-rigid* registration algorithms.

Rigid registration algorithms concern the registration of intra-patient data: data images are considered to represent the same physical body (although it might appear quite differently in different acquisition modalities) and the registration procedure search for a rigid transformation (a composition of a translation and a rotation) to match the two images. Rigid transformations are described by 6 parameters only (3 degrees of freedom in translation and 3 degrees of freedom in rotation) and the associated optimisation process is usually reasonably tractable, unless processing very large dataset. Common extensions to rigid registration include similarity registration (7 degrees of freedom, adding a scale factor) or affine registration (12 degrees of freedom, adding anisotropic scale factors and shear factors).

Non-rigid registration algorithms concern the alignment of data acquired from different patients and representing similar but different shapes. Non-rigid registration is more complex than rigid registration as the transformation includes many more degrees of freedom (it is often a parametric transformation with variable degree of complexity or a dense transformation field). Therefore, non-rigid registration algorithms are much more costly (up to hours of computation time on today's workstations) and parallelization of some algorithms has been proposed. One of the key challenges to share non-rigid registration algorithms on a grid is the standardisation of the transformation format. Currently, transformation models as different as B-splines, NURBS, radial basis functions, or dense displacement fields are used to encode the deformation. A common framework will be needed to handle, compare and use all these models.

Image intensity correction techniques also often rely on optimisation procedures and therefore may fall in the compute intensive algorithms described in the previous section.

3.3.3. Interactive image processing algorithms

Another particularity of medical image processing algorithms is that some of them need to be executed interactively. There are two main reasons why a medical application might need to be interactive:

- To solve reliability problems: to ensure that the user gets full control of the algorithm output by interactive guidance.
- To solve legal responsibility issues: automatic processing of medical data often raises the problem of legal responsibility. A user-guided algorithm is not subject to this kind of criticism.

To ensure interactivity, an algorithm needs to be executed in a time short enough for the user to remain active in front of the screen (usually the whole process should not exceed a few minutes in the medical context). Grid infrastructures can provide the computing power needed to ensure that the execution time remains reasonable by allocating powerful computing resources for interactive jobs or by empowering parallel applications. However, porting interactive applications on a grid is made complex by the need to split the user interface (that displays the algorithm progress result on the user's screen) and the computing algorithm (that is remotely executed on the grid resources). Therefore, interactive applications have to be carefully designed in order to be ported onto grids.

A typical user-guided interactive medical application is that of segmentation algorithms. Medical image segmentation is a complex problem for which there exists no general solution. Most segmentation algorithms such as deformable models or voxel clustering algorithms are iterative. It is therefore possible to update the algorithm progress on the user screen periodically and to take into account some user input at each stage to guide the algorithm while it is progressing. Likewise, enabling interaction with grid-powered non-rigid registration algorithms would enable correction of mistakes created by local minima (especially in multi-subject brain registration) while retaining the accuracy of the automatic processing and a reasonable human computation time.

Mammograms analysis for breast cancer screening

One current example of a large-scale medical image acquisition and processing application is the automated detection of malignant tumours in mammograms developed to support breast cancer screening programs that are starting in several European countries today. Screening programs at a national scale require the reading of a huge number of images (*e.g.* one mammogram for each woman older than 40 years every 2 years) thus considerably increasing the burden of image analysis on radiologists. Grid-enabled mammogram analysis projects aim to prove the viability of the grid by harnessing its power to enable radiologists from geographically dispersed hospitals to share standardised mammograms, to compare diagnoses (with and without computer aided detection of tumours) and to perform sophisticated epidemiological studies across national boundaries. Research is currently being conducted into imaging workstation architectures, into information infrastructures to connect radiologists across a grid, and into DICOM-compliant object models residing in multiple, distributed data stores, as well as into mammogram indexing, etc. There are a number of relevant technologies that are being harnessed together to provide a distributed infrastructure to support radiologists in their work. These include mammogram analysis algorithms, grid middleware implementations, and computer-aided detection software.

However they have only just scraped the surface in matching these user requirements. Data heterogeneity is one major issue in the storage and analysis of medical images – even in a single region of a single country never mind inter-regional or international data differences. The ability to process unstructured (*e.g.* radiologists annotations), semi-structured (patients' medical history) as well as rigidly structured patient data (metadata such as age, drug treatments, etc) is essential to enable the controlled execution of epidemiological studies or other query-based analyses.

4. Computational Models of the Human Body

4.1. THERAPY PLANNING AND COMPUTER ASSISTED INTERVENTION

Beyond medical data acquisition and analysis, modelling of the human body enables specific medical treatments. The key distinguishing factor compared with image processing or image reconstruction in the same application domain is the use of computational methods for predictive purposes – providing physically accurate (within modelling accuracy) information that is not included in medical images themselves.

Enormous progress has been made in recent years (aided by the increases in performance of computing platforms) and numerical modelling is now able to provide realistic (and validated) predictions of very complex phenomena. However, there is a real need for the continued development of numerical modelling and simulation technology to address the future challenges of multi-scale, multi-physics problems that arise naturally and automatically in virtual human modelling.

Given the complexity and the computing cost of most human body models, grid technologies are a good candidate to face computation challenges arising in this area.

4.2. ATLASES

Atlases have long been used in medicine for anatomy and physiology studies. For centuries, atlases have been produced manually by experts from their knowledge of the human body. Atlases attempt to provide a 'standard' description of the human body or parts of it. They are very dependent on the designer and have been incrementally refined with the progress of medicine. They tend to be general and hardly take into account infrequent parameters.

With the advent of digital images and image registration algorithms, the production of digital atlases has become possible. Digital atlases are assembled by registering large training sets in a common frame and averaging the registered images by different means. Digital atlases prove to be much easier to produce than manual atlases. They have encountered a tremendous success and have lead to significant research progresses, especially in the domain of brain imaging. The production of atlases require the availability of training datasets large enough to be statistically representative of the population under study and of sufficient computation power for accomplishing the registration and intensity correction computations. Grid technologies promise to cover both aspects and should therefore boost the production of anatomical and functional atlases of the human body. Given a wide scale medical information system and considerable computing power, one can even imagine producing on-the-fly individualized atlases. For example a physician may want to study the brain of a 50 year-old male subject to multiple sclerosis; he could ask for the production of an atlas from a training set with matching criteria. Such an individualised atlas would prove to be much more specific and precise than a generic atlas.

4.3. NUMERICAL SIMULATIONS OF THE HUMAN BODY

The release, some years ago, of the Visible Human (VH) dataset made it possible, for the first time, to access anatomical information without compromises. This produced a significant momentum in many areas. However, after some time it became clear that, while the dissection approach used in the VH project ensured extreme quality, it also lacked physiological information that other forms of data contain. These include *in vivo* data collection, multi-subject, gender, sex, and age variations, lack of connection with functional information, no pathology, etc. Many research projects have been carried out in Europe over the last few years to try to circumvent some of these limitations. A basic feature of the VH project, lacking in all these other projects, is completeness. The VH project relates ONLY to the normal anatomy of one human subject, and provides ALL the anatomical information for that subject. The other projects focused only on specific aspects. Because of the lack of the necessary critical mass, none has dared to search for completeness.

The Living Human Project (LHP) intends to develop a world-wide, distributed repository of anatomo-functional data and of simulation algorithms, fully integrated into a seamless simulation environment and directly accessible by any researcher in the world. The objective is patient-specific bionumerics and image processing (both for pre-processing and visualisation) for the complete human body. It requires the integration of individual systems through hierarchical approaches at the algorithmic level. With the development of grid and large medical databases, one can expect the development of more specific or even individualised models. These models could be built from specific patient data and target specific pathologies or functions.

Many areas of development in numerical human modelling are already at the stage that they can be used by medical researchers as tools for investigation into cause of medical problems and treatment procedures. Research into cardio-vascular disease in particular is an area where HPC simulation software is widely used, for example to improve understanding of processes leading to illness or to failure of implants such as artificial heart-valves or stents.

The interest of the grid approach is to provide services to medical or clinical users, removing any need for them to have to handle the details of any computing systems or simulation methods. Grid technologies are also required to provide high-bandwidth to large collections of coarse-grained, distributed, non-textual, multidimensional, time-varying resources. Web services technologies are required to cope with the dynamic aspects of a digital library that provides not only data, but also simulation services, collaborative work services, interactive visualisation services, and so on.

Broadening the term “medical supplier” to include pharmaceutical industries, the acceptance of the potential benefits of using numerical simulation tools (*i.e.* actual use or willingness to investigate use) is already well established within the R&D divisions of companies. For large companies, grid offers the means to deploy simulation software across their own distributed resources. There are also established SME’s supplying services and consultancy based on numerical simulation. Future grid developments will allow them to

enter into virtual organisations with their customers (including controlled access to data sources) and to have access to external computational resources when needed.

4.4. ISSUES FOR THERAPY PLANNING

Many human body models have been developed for therapy planning. Examples of numerical simulation used by health practitioners include radio-surgery/radio-therapy planning (see section 4.3.2), electromagnetic source localisation (an inverse procedure to identify areas of disorder within the brain based on external EEG/MEG measurements), reconstructive maxillofacial surgery (see section 4.3.1), etc. Today, most developments are in the transition between research use and clinical use. Grid can be used to provide access to appropriate computational services and deliver these to medical users. Healthgrid would need large scale deployment studies to allow the evaluation of a wide range of requirements, including local deployment aspects and practical experience with production grid use. The major challenges will be to ensure that services can be delivered into the user's workplace in an appropriate, ergonomic manner and that security, policy and legal constraints related to the use of patient data are fulfilled.

A grid scenario for radiotherapy planning and treatment

From a technology point of view, radiotherapy is a highly complex procedure, involving a variety of computational operations for data gathering, processing and control. The modularity of the treatment process and the need of large data sets from different sources and nature (physics, mathematics, bio-statistics, biology, and medicine) make it a privileged candidate for healthgrid applications.

In an enlarged Europe sharing data, expertise and computational resources will be a significant factor for a successful cost containment and improved access to a high overall quality of care in radiotherapy. It is an ideal tool for harmonising the cancer treatment as well as providing a common base for research collaboration.

Presently patients are treated with standardised radiation doses. Gene profiling may enable an individualised adjustment of the dose so as to achieve tumour control in patients with a low radiosensitivity and avoid severe side effects in patients with above average sensitivity to radiation. In a first step a grid structure should allow research groups, each focusing on different molecular mechanisms, to access data in the distributed infrastructure for comparison studies. In a next step users should be able to submit the results of predictive tests for analysis to a shared software and expert platform for radiosensitivity grading.

A similar approach can be followed for other aspects clinical decision making such as the assessment a tumour's capacity for metastatic spread. For rapidly metastasising tumours, systemic (chemotherapy) treatment needs to be associated to the locally delivered radiotherapy. New tests now under development, predicting on the basis of gene profiling which tumours are most likely to metastasise, can make 60% of the chemotherapy currently administered e.g. for breast cancer, redundant. However, it takes a highly specialised team to interpret the results of these tests correctly. Grid-supported consultation of libraries of gene profiles or, alternatively, tele-consulting services offer also in these case excellent perspectives.

Tissue electron density provided by CT scanning is still needed to calculate the dose delivered by photon and electron beams. To define the planning target volume (PTV) and organs-at-risk (OAR), new imaging modalities based on MR-imaging, MR-spectroscopy and PET are far superior and become a requirement for high-precision high-dose radiotherapy. In contrast to CT scanning, the latter imaging modalities are available only in reference centres for reasons of cost and expertise. To secure access for all patients to optimal imaging for radiotherapy planning, the coordinating centre could perform a grid-mediated selection of an imaging centre, and the resulting complementary image acquisitions could be sent back through the grid. To reproduce the patient positioning and perform the complementary imaging in treatment-relevant conditions, the patient-individual immobilisation devices could be physically sent to the imaging centre. Alternatively, a retrospective registration grid service could be used to realign all the images in the relevant coordinate system.

Many tools have been developed for computer-aided definition of PTV and OAR including anatomical atlases that can be warped to the patient-individual anatomy. A grid could make such tools and their upgrades in due time available to all groups involved in PTV and OAR definition. Nodes on the grid that provide expert help for patient-related problems in defining PTV and OAR are needed.

Accuracy of Monte Carlo (MC) dose computation is excellent, provided that the computing power is sufficient to allow for enough runs to reduce the statistical noise. The grid is a natural alternative to costly parallel computers. In this way, MC dose computations could become standard for radiotherapy quality assurance (QA), planning, and plan optimisation years before individual departments could afford a local investment that is capable to support MC. Requirements needed for such deployment include the existence of a service level agreement between the departments and the grid providers by which the grid level of performances in terms of security, stability and response time is guaranteed.

Each delivery centre manages the commissioning of its own treatment units and incorporates both mechanical-physical and dosimetric parameters, including uncertainty flags, into an identity card that is accessible through the grid. This identity card will allow treatment-planning providers and computation services to establish, refine or fit their computational model of the linear accelerator. The identity card also contains the reference data so that periodical quality assurance (QA) procedures could make sure that the machine performs accordingly. One might expect that the cooperation through the grid between QA providers and delivery centres will streamline the QA procedures and harmonise the identity cards over the different accelerator types.

The quality assurance of the treatment can also benefit from the grid, even if it is patient specific: once a treatment plan has been designed, some locations are selected to measure the dose level in a physical phantom that replaces the patient during the first treatment session. In parallel, the coordinating centre consults the grid for an independent dose computation service to compute the dose in the same set of points in the phantom. The comparison of the measured dose to the computed fractional dose is performed automatically at the delivery centre and will be submitted to the coordinating centre. In case of violation of tolerances, the treatment plan will be recomputed in patient and phantom by a second dose computation service in the grid. Alternatively, the coordinating centre may consult the grid for a virtual treatment at another delivery centre.

4.5. TOWARD REAL-TIME CONSTRAINTS

Some medical applications such as surgery simulation are more demanding and require real-time computations. Real-time is a challenging problem for grid infrastructures today. Although grids can provide additional computing power, distributing computations to remote resources if often done at the price of an initialisation cost that can be significant (from minutes to hours in common batch-oriented scheduling systems). To empower real-time applications, a grid middleware would need to ensure immediate execution of real-time code. Strong network requirements are also dictated by real-time constraints. Grid services dealing with jobs as sensitive as surgery simulation and computer assisted intervention should also have the capacity to make advance reservation of resources and to cope with any emergency situations: the requested computation and networking resources must be allocated when the surgery starts and it should be possible to submit prioritised jobs in case of emergency with resource requisition and contention resolution as required.

4.5.1. Surgery simulation

Surgery simulation is the aim of many research activities today: it is a promising tool both in planning surgery and in training surgeons. Realistic surgery simulation usually involves complex biophysical models of the human body. The building of a model for surgery simulation (e.g. using finite element modelling) and its use in an interactive context have to be distinguished: building the model may require intensive and long term effort, but its final formulation should enable very fast computation for the purpose of the simulation itself (deformation of organs, evolution of physiology, etc).

Given the complexity of human body modelling, surgery simulators are often limited to a specific intervention procedure. Another constraint is the mechanical devices manipulated by the practitioner during the intervention: an endovascular intervention procedure or a laparoscopic surgery intervention are more easily simulated than open surgery since they require visual and haptic feedback devices with limited capabilities. Development of open surgery simulation tools is also limited today by the state-of-the-art in 3D rendering and full degree of freedom devices. Even considering only limited intervention procedures, the computations involved may be very difficult to achieve in real time: visual feedback is known to require an update frequency of 25 Hz and realistic haptic feedback may require much higher frequencies (up to 300 Hz for soft tissues and thousands of Hz for rigid material such as bone). While a great deal of progress in grid technologies, both in power and bandwidth, may be anticipated, there are further demands to be place on it. For example, the compositional integration of various models (mechanics, visual rendering, device interactions, etc) would be yet another requirement, if grid is to enable more realistic and broader real-time simulation tools.

4.5.2. Augmented reality and computer assisted intervention

The next stage in real-time modelling of biophysics is its coupling with intervention data in order to bring additional information that could not be observed during a medical intervention. For instance, augmented reality consists in superimposing on the

scene that the practitioner perceives additional information coming from a computerised model, usually through visual devices. This enhanced perception proves to be useful in many types of interventions: it allows a neurosurgeon to visualise the brain tumour he has to remove by projecting it on the head of the patient prior to and during the intervention, e.g. to guide its resection; or it aids a dentist to visualise the planned position and axis of drilling to place an implant; or a radiologist to guide the placement of a needle for a biopsy or a radio frequency ablation. In all these cases, augmented reality helps reduce the invasiveness of the procedure.

Many currently existing augmented reality systems rely on simplified models where only a simple calibration step is required, simply because this is computationally tractable. Indeed, more complex augmented reality applications need huge computing power for the pre-operative construction of patient-specific models and for the per-operative adaptation of these models to reality (registration, geometric deformations, etc). Going to the complete integration of a bio-physical model into a clinical augmented reality system is a challenging task where the grid could be the key. However, this would imply very strong requirements on the security and dedication of the computer and network resources in order to ensure the reliability of the real-time system.

Another way to enhance practitioner capabilities is to provide a computer assisted action, for instance though the use of robots. Even if the robot is passive (e.g. a robot-arm guided by a surgeon), it brings a large benefit such as minimizing human arm motion and filtering out any hand tremor. Active robots may provide even more benefit, for instance by compensating for organ motion to give the surgeon the illusion of working on a static structure. By decoupling perception (using augmented reality) from action (using robots), it has been possible to separate the surgeon from the patient, and remote surgery has proved to be possible through the use of high bandwidth dedicated networks. Manipulating the controls through networks from a distant location certainly raises the problem of network performance and quality of service: the data flow is critical and a guaranteed bandwidth mandatory.

4.6. REFERENCES

[1] Information on Maxillofacial surgery application can be taken from "The GEMSS Grid: An Evolving HPC Environment for Medical Applications", D.M. Jones, J. W. Fenner, G. Berti, F. Kruggel, R. A. Mehrem, W. Backfrieder, R. Moore, A. Geltmeier

[2] "Parallelization of Monte Carlo simulations and submission to a grid environment", Maigne L., Hill D., Breton V., Reuillon R., Calvat P., Lazaro D., Legré Y., Donnarieix D., accepted for publication in Parallel Processing Letters

5. Grid-Enabled Pharmaceutical R&D: Pharmagrids

The Pharmaceutical R&D enterprise presents unique challenges for Information Technologists and Computer Scientists. The diversity and complexity of the information required to arrive at well-founded decisions based on both scientific and business criteria is remarkable and well-recognized in the industry. The decisions can form the basis for multi-year multi-person multi-millions of Euro investments and can create new scientific territory and intellectual property. Thus all aspects of managing, sharing and understanding this information is critical to the R&D process and subject to substantial investment and exploration of new informatics approaches.

Pharmaceutical R&D information includes are large variety of scientific data as well as sources of critical organizational information such as project and financial management data and competitor intelligence information. This data takes some fairly unique formats as well. Examples are images, models, sequences, full text scientific reports, records of prescriptions and physician encounter re-imbursements. These sources of information consist of internal proprietary, external commercial and open-source data.

The problems range from knowledge-representation and integration, to distributed systems search and access control, to data mining and knowledge management, to real-time modelling and simulations, to algorithm development and computational complexity.

Grid technology holds out the promise of more effective means to manage information and enhance knowledge-based processes in just the sort of environment that is well established in pharmaceutical R&D.

A pharmaceutical Grid should be a shared *in silico* resource to guarantee and preserve knowledge in the areas of discovery, development, manufacturing, marketing and sales of new drug therapies [5.3] and cover three dimensions:

- a resource that provides extremely large CPU power to perform computing intense tasks in a transparent way by means of an automated job submission and distribution facility
- a resource that provides transparent and secure access to storage and archiving of large amounts of data in an automated and self-organized mode
- a resource that connects, analyses and structures data and information in a transparent mode according to pre-defined rules (science or business process based)

Pharmaceutical grids open the perspective of cheaper and faster drug development. Pharmaceutical grids should enable parallel processes in drug development, away from the traditional approach where target discovery, target validation, lead discovery, lead optimization and transition to development take on average 12 years. These parallel processes would take advantage of in silico science platforms for target identification and validation, compounds screening and optimization, clinical trials simulation for detection of deficiencies in drug absorption, distribution, metabolism and elimination.

Pharmaceutical grid for a rare disease

Infectious diseases kill 14 million people each year, more than ninety percent of whom are in the developing world. Access to treatment for these diseases is problematic because the medicines are unaffordable, some have become ineffective due to resistance, and others are not appropriately adapted to specific local conditions and constraints. Despite the enormous burden of disease, drug discovery and development targeted at infectious and parasitic diseases in poor countries has virtually ground to a standstill, so that these diseases are de facto neglected. At the same time, the efficacy of existing treatments has fallen, due mainly to emerging drug resistance.

Rare Diseases represent grave personal tragedies and *in toto* substantial health and economic burdens even for the wealthiest nations [5.4]. Nor is it always true that there is no economic driving force for the development of therapeutic interventions for rare diseases [5.5].

The unavailability of appropriate drugs to treat neglected diseases is among other factors a result of the lack of ongoing or well coordinated R&D into these diseases. While basic research often takes place in university or government labs, development is almost exclusively done by the pharmaceutical and biotech industry, and the most significant gap is in the translation of basic research through to drug development from the public to the private sector. Another critical point is the launching of clinical trials for promising candidate drugs.

Producing more drugs for neglected diseases requires building a focussed, disease-specific R&D agenda including short-, mid- and long-term projects. It requires also a public-private partnership through efficient, secure and trusted collaborations that aim to improve access to drugs and stimulate discovery of easy-to-use, affordable, effective drugs. The goal is to lower the barrier to such substantive interactions in order to increase the return on investment for the development of new drugs.

A 'pharmagrid' should create a virtual organization and collaborative environment which will motivate and gather together:

- drug designers to identify new targets and drugs
- healthcare centres involved in clinical tests
- healthcare centres collecting patent information
- organizations involved in distributing existing treatments
- informatics technology developers
- computing and computer science centres
- biomedical laboratories working on vaccines, genomes of the virus and/or the parasite and/or the parasite vector

Pharmagrid will support such processes as:

- search of new drug targets through post-genomics requiring data management and computing
- massive docking to search for new drugs requiring high performance computing and data storage
- handling of clinical tests and patient data requiring data storage and management
- overseeing the distribution of the existing drugs requiring data storage and management
- trusted exchange of IP, possibly auction-mediated

A grid dedicated to research and development on a given disease should provide:

- resources for computationally intensive search for new targets and virtual docking
- resources for massive storage of post genomics and virtual docking data output
- grid portal access to post genomics and virtual docking data
- grid portal to access medical information (clinical tests, drug distribution, etc.)
- a collaboration environment for the participating partners.

For competitive and intellectual property protection reasons, pharmaceutical Grids will predominantly be private enterprise-wide internal grids with strict control and standards. At least this will likely be the case in the near-term as more and more R&D organizations explore and become comfortable with this technology and its potential.

However, the promise of the grid to create effective virtual organizations based on efficient secure and trusted-collaborations will create the foundation for new forms of partnerships – amongst commercial, academic, government and international R&D organizations.

Figure: Concrete Structure of a Grid for Rare Diseases.

The *Basic Grid Technology* layer comprises the basic grid engine for scheduling and brokering of resources. The *Virtual Organization* (VO) layer integrates users from different and heterogeneous organizations. Access rights, security (encryption), trust buildings are issues to be addressed and

solved on this layer. The *Distributed Data Access / Information Retrieval* layer addresses one of the major challenges: the problem of semantic inconsistency between biological and chemical databases is even more urgent in the grid context. Ontology-based mediation services for data integration might provide one road to go for a grid for rare diseases; another option would be to make use of developments from other grid projects (e.g. the distributed query processor (DQP) [5.6] or the federated version of SRS [5.7]) The *Integration of Application* layer will require substantial meta-information on algorithms and input / output formats if tools are to be interoperable in the grid. Assembly of tools for virtual screening into complex workflows will only be possible if data formats are compatible and semantic relationships between objects shared or transferred in workflows are clear. Next comes the *Workflow* layer. One core element of a grid for rare diseases is the virtual screening machine including, amongst other functionalities, a generator for focused virtual libraries, high throughput docking software, different filters for pre- and post-processing of hits in the virtual screening procedure and software for the prediction of basic ADME parameters. The combination of the tools behind these functionalities in a workflow and the execution of this workflow in the grid requires a formal description as provided e.g. by WPDL [5.8] or SWFL [5.9, 5.10]. The *Ontology / Knowledge Representation* layer maintains formalized knowledge representations (ontologies). These must play a key role in any future pharmaceutical grid. A grid for rare diseases would require significant activity to construct an ontology for the disease under investigation, for genetic epidemiology aspects including the categorization of clinical phenotypes. Moreover, a pharmaceutical ontology would have to bridge from biology to chemistry as it would have to describe formally a pharmaceutical target as well as the concept of an “in silico screening hit” and its development into a “lead compound” for experimental evaluation. The *Data and Knowledge Mining Services* layer includes services for statistical approaches to data mining (e.g. in the field of epidemiology) and learning and optimization of *in silico* drug discovery approaches. Knowledge mining services will largely depend on the availability of a pharmaceutical ontology. Interoperability of statistical models as well as the issue of comparability of predictions made on the basis of these statistical models.

5.1. REFERENCES

[5.1] http://www.pharmaGrid.com/
[5.2] http://www.prismforum.org/
[5.3] Rene Ziegler, proceedings of the HealthGrid conference, Clermont-Ferrand, January 2004
[5.4] http://www.rarediseases.org/
[5.5] http://www.unicorntodoublehelix.com/
[5.6] http://www.ogsadai.org.uk/dqp/
[5.7] http://www.tm.uka.de/~fuhrmann/Publications/fuhrmann03overlaySRS.pdf
[5.8] http://www.wfmc.org/
[5.9] http://www.cs.cf.ac.uk/User/Yan.Huang/GridWF/SWFL.htm
[5.10] http://www.wesc.ac.uk/projects/swfl/

6. Grids for epidemiological studies

Conventional epidemiology requires extensive collections of data concerning populations, health and disease patterns, as well as environmental factors such as diet, climate and social conditions. A study may focus on a particular region or a particular outbreak, or it may take as its theme the epidemiology of a condition across a wide area. The range of data required will, therefore, vary with the type of study, but certain elements persist: a degree of trust in the data is essential, so its 'provenance' has to be assured and the standards of clinical practice under which it was obtained have to be above a certain threshold. Where the data has been gathered under different clinical regimes, it must be possible to establish their semantic equivalence, to ensure that aggregation or comparison of datasets is legitimate. Ethical issues may also arise if data collected in the first place in the course of individual health care is to be used for research.

The analysis of aggregated data requires the construction of complex models and the use of sophisticated statistical tools. This has necessitated collaboration between physicians and statisticians, and the rise of epidemiology as a discipline. The impact of genomic analysis will extend the kinds of variable under study and the range of expertise to be applied.

The technology to allow federation of databases stored locally in hospitals has existed for some time. It is possible for these databases to be queried for epidemiological purposes while preserving patient anonymity. Such distributed queries may be managed and supervised by the hospitals with primary responsibility for the data, ensuring compliance with ethical and legal regulatory frameworks. None the less, the political difficulties inherent in the integration of information systems are well known and this has plainly not happened to the degree that it is possible despite major government efforts.

Grids supervene mere integration of databases. They can enforce the interoperability of tools and analysis services and they may also enforce common standards and semantic clarity about database content and tool input / output. Indeed, the Grid-based federation of retrieval systems provides a significant alternative to federation of databases. We may not see the latter for quite some time: federation of databases requires – in case the databases should be interoperable – clear semantics and standards based on conventions about semantics. Attempts to use semantics-based mediators have not been particularly successful so far.

In contrast to bioinformatics, where at least two major systems for data integration are in use (ENTREZ at the NCBI and SRS at EBI), no such integration layer exists in the field of medical informatics.

One road to go for the integration of medical data would be to adopt Grid strategies for data integration developed for bioinformatics. In SIMDAT, an Integrated Project funded in the course of the FP6 IST programme, federation of the data integration system SRS is one of the major R&D goals defined for this project.

If such an approach is adopted, the cost and effort for establishing completely new databases in the field of clinical research / genetic epidemiology would be significantly limited, thus paving the way for smooth and rapid implementation of first demonstrators.

The proposed adoption of federated SRS as a data integration platform for medical (phenotype) data should not at all prevent a HealthGrid community in the field of genetic epidemiology from doing their homework on standards. Any type of interoperability requires a broad and common understanding of data types and applications. Therefore, domain-specific meta-data will play a crucial role in Grids for genetic epidemiology (as much as in all other HealthGrid scenarios) to enable interoperability of analysis methods and comparability of data and results.

6.1. DATA SEMANTICS IN GENETIC EPIDEMIOLOGY

Standardised semantics will be essential for genetic epidemiology. Although a significant portion of developments done in the context of the semantic web will be relevant and partially re-useable for biomedical Grids, domains such as genetic epidemiology will need dedicated initiatives for clarified semantics carried on by experts in the field. Unified naming of phenotypes and standardised acquisition and recording of clinical parameters have to be supported by a Grid for genetic epidemiology. One of the central services in a Grid for genetic epidemiology studies has to be a clinical annotation service for clinical phenotype descriptions. Such an annotation service has to be user – friendly, easy to use by non-computer-experts and it has to make use of widely accepted naming concepts in the domain of genetic epidemiology (if they exist at all). One possible solution to the problem of a Grid-based annotation service for clinical phenotypes would be an ontology-based annotation service which would allow navigation through controlled vocabularies and selection and linking of defined concepts to entries in existing databases for phenotype recording.

6.2. IMAGE ORIENTED EPIDEMIOLOGY

The specific requirements for the use of Grid technology related to imaging have been discussed in chapter 3. Here we will only address the specific issues related to the use of images in epidemiological studies.

Patient management (diagnosis, treatment, continuing care, post-treatment assessment) is rarely straightforward; but there are a number of factors that make patient management based on medical images particularly difficult. Often very large quantities of data, with complex structure, are involved (such as 3-D images, time sequences, multiple imaging protocols). In most cases, no single imaging modality suffices, since there are many parameters that affect the appearance of an image and because clinically and epidemiologically significant signs are subtle. Among the many relevant factors are patient age, diet, lifestyle and clinical history, image acquisition parameters, and anatomical and physiological variations. Thus any database of images developed at a single site– no matter how large – is unlikely to contain a large enough set of exemplars in response to any given query to be statistically significant. Overcoming this problem implies constructing a very

large, federated database, while controlling for statistical biases such as lifestyle and diet almost certainly leads to a database that must transcend national boundaries. Realizing such a geographically distributed (pan-European) database necessitates so-called Grid technology [4], and the construction of a prototype would push emerging Grid technology to its limits.

The MammoGrid project:

The MammoGrid [5] project is providing a collaborative Grid-based image analysis platform in which statistically significant sets of mammograms can be shared between clinicians across Europe. The applications to be implemented can be thought of as addressing three main problems:

- Image variability, due to differences in acquisition processes and to differences in the software packages (and underlying algorithms) used in their processing.
- Population variability, which causes regional differences affecting the various criteria used for the screening and treatment of breast cancer.
- Support for radiologists, in the form of tele-collaboration, second opinion, training, quality control of images and a growing evidence-base.

In practical terms, the project will:

- evaluate current Grids technologies and determine the requirements for Grid-compliance in a pan-European mammography database;
- implement a prototype MammoGrid database, using novel Grid-compliant and federated-database technologies that will provide improved access to distributed data;
- deploy versions of a standardization system (SMF – the Standard MammoGram Form [6]) that enables comparison of mammograms in terms of tissue properties independently of scanner settings, and to explore its place in the context of medical image formats; and
- use the annotated information and the images in the database to benchmark the performance of the prototype system.

The European dimension of the MammoGrid consortium, including hospitals in north and south Europe, provide the first opportunity for statistical studies of breast cancer to be conducted and analyses to be made on geographical, cultural, environmental and temporal influences on cancer development. MammoGrid should provide statistically significant numbers of exemplars even for rare conditions of cancer development and will therefore enable more diverse epidemiological studies than hitherto have been possible. The project will develop standard data formats and strict automated quality checks, which will lead to improved and normalised breast screening procedures. Such a secure, efficient and standardised storage of medical knowledge in an EU-wide federated database will also provide an ideal educational tool for training radiographers and radiologists. Standardisation on data formats will control the variation in the quality of images and diagnoses in European healthcare.

6.3. **BUILDING POPULATION-BASED DATASETS**

A European Grid for Genetic Epidemiology would open completely new perspectives for gathering data on large populations and – as a consequence – would allow stratification

of large cohorts for large scale European Genetic Epidemiology studies. One possible problem that we foresee in this context is that there are regional, legal and cultural differences that may obstruct the building of pan-European, population-based datasets. As a consequence, we propose to complement any type of HealthGrid activity that could possibly encounter problems of this type is supplemented and accompanied by research activities in the field of ethical, legal, and cultural aspects that might impact future healthgrids.

The current situation in Europe is quite heterogeneous. Initiatives to build large population-based datasets have been started in Iceland [9], the UK [10], and in one Baltic state, Estonia [11]. These national initiatives are driven by a different rational: whereas in Iceland it was a private-public partnership between DECODE genetics and the government of Iceland in the UK and in Estonia the initiatives are based on governmental scientific research programmes. In how far commercial aspects will interfere with the goals of a pan-European initiative to build population-based datasets remains unclear, however, it is clear that large population-based datasets (and associated sample collections) are not only interesting for basic science but also for the pharmaceutical industry.

Even though we foresee problems as discussed above, the chances that come with large scale studies and pan-European population-based datasets will exceed the risks of potential abuse of genetic information by and large. Currently, genetic epidemiology studies suffer from low numbers of samples, inconsistent acquisition of bio-parameters and complex genetics.

6.4. **STATISTICAL STUDIES**

Built on population-based datasets statistical studies on the influence of allelic predisposition, behavioural aspects, nutrition habits, regional or national healthcare management and many other parameters will be possible. A central task for a Grid project for genetic epidemiology would be to enable and to promote interoperability of statistical analysis tools. Similar to initiatives in the field of systems biology an exchange service for statistical models based on a common understanding and classification scheme of statistical approaches would be needed. A point to start with would be a “tool box” of statistical models including relevant meta-information on algorithms, modelling strategies and constrains, application scenarios and possible equivalence or variations of statistical models. As a Grid service this tools box would allow easy exchange of methods and improve interoperability of statistical models and data mining capabilities on the side of the users of the Genetic Epidemiology Grid.

6.5. **PATHOLOGIES EVOLUTION IN LONGITUDINAL STUDIES**

The study of pathologies follow-up would include information related to regular hospital visits, home-care monitoring of signs and symptoms, recording of interventions and drug effects, environmental issues etc. However, these studies are usually fragmented and non-uniform, thus, cannot result in common conclusions. One can see this issue from two standpoints: a) how pathology follow-up or the setup of clinical trials can be supported, and

b) how the results of clinical trials can be better utilized in a manner that feeds medical knowledge and clinical practice.

The main obstacles that have to be overcome towards the evolution of pathologies into longitudinal studies, in order to provide enhanced medical knowledge and procedures, are:

- Clinical protocols are not always standardized and widely accepted
- Measurements, devices, computational overhead as well as data, may vary
- Variability in populations participating in the clinical trials
- Conception of diagnosis and treatment may also vary

Accordingly, the requirements arisen for effective longitudinal studies are:

- Large studies leading to better statistics and understanding of mechanisms
- Multi-center approaches that take into account environmental and other factors
- Availability of evidence-based medicine
- Sophisticated statistical analysis and modeling
- Facilitate cooperation among healthcare professionals
- End-up with protocols, data descriptions, measurement descriptions and models

Adoption of a Grid-based approach in developing pathology follow-up studies may provide:

- Support and improvement of existing databases import/export facilities
- Transparent access to data from the user viewpoint, without knowledge of the actual data location
- Authorization policies allowing anonymous and private login for access to public and private databases
- Provision for the privacy of medical information and fulfilment of legal requirements in terms of data encryption and protection of patient privacy
- A wide range of analysis tools, and contribution to the comparison-benchmarking of software applications, as well as to the combination of methods supporting clinical practice
- Access to tools and services that support the clinical trials, e.g., real-time processing tools, alerting tools for the clinicians, educational services for patients, etc.
- Establishment of common protocols for homogenizing data originated from distributed and heterogeneous databases, based on common semantic mechanisms
- Methods for fetching data based on similarity measures, for example, supporting diagnosis in ambiguous cases
- Common calibration methods for measurements, thus, mechanisms dealing with measurements' variability and ensuring a common understanding of measurements and devices

Grid on nosocomial infections

Nosocomial infections are among the three most costly and deadly infectious diseases. The growth in these has continued unabated for nearly two decades, despite many measures – such as shorter hospital stays – which can reasonably be expected to have had an attenuating effect.

A major reason for this growth has been the emergence of antibiotic resistant bacteria. There are now bacterial strains which are resistant to all but one known antibiotic. It is widely argued that the only sustainable defense against this danger is greater vigilance, public education and a significant reduction in 'antibiotic pressure' in the community.

Greater vigilance and preparedness are also the only possible defenses against two other modern plagues: bioterrorism and various economically catastrophic animal diseases – in the United Kingdom, BSE and FMD being cases in point.

There are several scientific and technical challenges in the design of a Grid epidemiological information system. The typing, i.e. the identification, of bacterial strains is a problem for several reasons, among which the multiplicity of typing methods and the difficulty in communication in the absence of a universal coding system are significant. Projects to define a common language often rely on one particular method, but there is a need to continue to accommodate new techniques which promise greater discrimination. It is argued that typing of bacterial strains, with the need to search for and reconcile fuzzy information across a large number of reference locations, is in itself a suitable Grid problem.

Any strategy to combat antibiotic resistance based on epidemiological insights will have to take account of the impact of such factors as levels of antibiotic prescription and of what is known about patterns of disease evolution. [7] In both these areas, provided information is gathered – e.g. about the volume of pharmacy-dispensed antibiotic prescriptions – the evidence base on which to determine best practice would itself continue to evolve and improve.

A grid collaboration in the epidemiological control of antibiotic resistant pathogens would require at least the following:

- partnership and integration of knowledge from projects such as EURIS and EARSS;
- a plausible solution to strain identification as an information problem;
- coordination of computational efforts to identify and predict patterns of disease propagation.

6.6. **DRUG ASSESSMENT**

On the biological and pharmacological side, the determination of allelic frequencies of drug target genes in European population is one important application field for a genetic epidemiology Grid with large population-based datasets. A second application scenario concerns aspects of drug safety; again an aspect that is highly relevant for public health and the pharmaceutical industry. Adverse drug effects depend – amongst other factors – on cytochrome gene polymorphisms and one of the first large scale study done on a Grid for genetic epidemiology could be a project on cytochrome allelic variability in patients with e.g. resistance to a certain class of compounds.

A third application scenario could strive to unravel the genetic basis of drug insensitivity which is not based on allelic variation of acute response detoxification genes. As an example we might think of the insensitivity of a huge percentage of multiple sclerosis patients to treatment with Interferons. Another scenario would concern the insensitivity of a significant portion of the European population to treatment with glucocorticoids.

From the Grid research perspective, drug related epidemiological studies require a tight integration of knowledge coming from heterogeneous disciplines, namely pharmacology and genetics. Currently, knowledge representations (ontologies) for pharmacology are missing by and large; we therefore expect that a Grid on genetic epidemiology that addresses aspects of drug action will have to include an activity on ontology construction for the domain of pharmacology. A "pharmacology-ontology" would also help to formalise and to standardise the description of clinical parameters measured in the course of large scale studies. As drug assessment comprises all aspects of pharmacodynamics, special attention will have to be paid to appropriate representation of dynamic processes (e.g. changes of drug serum concentration over time); sharing of mathematical / statistical models for the analysis of drug effects and drug stability will be essential for pan-European studies.

6.7. **GENETIC EPIDEMIOLOGY**

The genetic basis of complex diseases provides a real challenge to any information system for genetic epidemiology and for a Grid for genetic epidemiology in particular. Complex diseases are characterized by the high number of parameters to be recorded and by an "intrinsic fuzziness" of the conceptual definition of clinical phenotypes (e.g. "depression"). Genetic epidemiology studies in this field require much larger cohorts of patients to produce significant results.

A Grid for genetic epidemiology could have several effects:

- Homogenisation of the selection of clinical parameters to be measured for the analysis of the genetic basis of complex diseases
- Interoperability of data at both, the data acquisition level as well as the database and data management level through structured knowledge representations
- Broadening of the statistical basis through expansion of relevant cohorts from regional or national scale to pan-European scale
- Interoperability of statistical models and efforts to enrich meta-information on analysis tools, algorithms and modelling approaches

Genetic epidemiology studies try to establish links between genetic variation (polymorphisms / allelic variance) and individual risk that have an impact on the quality of life (including major diseases).

Genetic epidemiology studies have a direct impact on decisions on health quality standards, disease management and risk assessment. Unfortunately, the prospects of Europe-wide genetic epidemiology studies have not yet been fully explored; even though

significant effort has been undertaken in the course of national projects, data from different studies are not easily comparable and data access is very limited.

A Grid – based system for genetic epidemiology will actually promote the development and / or adoption of standards in this field. It will also greatly improve interoperability of statistical analysis methods used for the analysis of genetic epidemiological data and it will probably allow for new ways to perform data mining approaches in a distributed (data) environment. The requirements of Grid – based systems for interoperability, clear semantics of data and applications, secure data handling of medical data and administration of virtual organisations are extraordinarily high.

Based on the general considerations outlined above, a Grid for genetic epidemiology would have to address the following aspects:

- clear semantics for data acquisition methods
- standards for the selection and description of patient collectives
- standards for patient collective size and statistical power with respect to patient collective size
- an ontology for technologies used in genetic profiling (an ontology similar to the microarray ontology generated by the MGED consortium)
- an ontology for phenotype descriptions based on a relevant controlled vocabularies
- a dedicated, Grid enabled annotation service for genetic epidemiology
- data security aspects of biomedical data handling, in particular paying tribute to the different European regulations for the handling of patient data
- interoperability of data analysis methods, in particular a means for declaration of statistical methods used
- capturing of statistical rational applied to patient collective selection
- capturing of rational for candidate gene selection
- capturing of rational for the selection of chromosomal regions
- declaration and brokering of statistical analysis services
- Grid based statistical modelling and data mining
- Grid based evaluation of existing relevant literature (including electronic patient records) by means of automated information extraction methods (text mining).

Substantial effort on open standards, capturing and formalisation of statistical considerations relevant for patient collective selection and controlled vocabularies / ontologies is needed. The scientific benefit of such effort, however, would be paramount:

- Data from national and European genetic epidemiological studies would be comparable at different levels, ranging from sample acquisition and sample

treatment protocols to the rationale for patient stratification and suitable statistical analysis approaches

- Standards for the description of clinical parameters would be established; the semantic relationship between parameters would be clear and consequently comparability of genetic epidemiological studies based on conceptual equivalence at different levels would be possible
- Interoperability of statistical models and analysis methods would be greatly enhanced; rational capturing for statistics would become a routine procedure
- Conclusions drawn from genetic-epidemiological studies could be re-analysed and re-tested with each new (equivalent) study.
- Parameters influencing e.g. the prevalence for certain tumour types in certain regions within the EU could be identified with a much higher chance. Effects influencing genotype-phenotype associations such as nutrition habits, behavioural differences, quality of health services and so forth could probably be quantified with much better significance.
- Variability of associations between genes and phenotypes could be assessed at the European level, which means that the genetic heterogeneity within Europe would open new perspectives to define "control groups" in statistical meta-analyses.

For a Grid for genetic epidemiology we foresee a key role for Grid services that refer to established controlled vocabularies and ontologies.

A problem particular to this field is that it suffers from the complicated and very complex phenotype descriptions necessary to describe e.g. depression in terms of quantitative parameters. This problem is very serious; current discussion of future trends in genetic epidemiology of complex diseases already foresees that this field of science is running the risk to become too expensive to be continued in the way this science has been done in the past. [8] A Grid for genetic epidemiology will provide a first means to make data and tools interoperable at the European level; ultimately such dedicated Grid will help to limit the costs of genetic epidemiology research in the field of complex diseases.

Examples of epidemiology Grids are:

- Genetic epidemiology Grids for the identification of genes involved in complex diseases
- Statistical studies: work on populations of patients. One example is the tracking of resistance to therapeutic agents. This is most notable in relation to antibiotic resistance in common bacteria in nosocomial and community settings
- Drug assessment: drug impact evaluation through populations analysis
- Pathology follow-up: pathologies evolution in longitudinal studies
- Grids for humanitarian development: Grid technology opens new perspectives for preparation and follow-up of medical missions in developing countries as well as support to local medical centres in terms of tele-consulting, tele-diagnosis, patient follow-up and e-learning.

6.8. **REFERENCES**

[1] Cancer Research UK *Breast Cancer Factsheet* (2003); *Scientific Yearbook 2001-02* (2002). See http://www.cancerresearchuk.org/; http://science.cancerresearchuk.org/.

[2] E.J. Feuer and L.M. Wun, *DEVCAN: Probability of Developing and Dying of Cancer*, Version 4.0, National Cancer Institute, Bethesda MD (1999). See summary at http://imaginis.com/breasthealth/statistics.asp.

[3] E.L.Thursjell, K.A.Lernevall and A.A.S.Taube Benefit of independent double reading in a population based mammography screening program, Radiology, 191, page 241 (1994).

[4] I. Foster, C. Kesselman & S. Tueke, *The Anatomy of the Grid – Enabling Scalable Virtual Organisations*, Int. Journal of Supercomputer Applications, 15(3), 2001.

[5] The Information Societies Technology project: MammoGrid - A European federated mammogram database implemented on a Grid infrastructure, EU Contract IST-2001-37614.

[6] SMF : Mirada Solutions' *Standard Mammogram Form*™ See http://www.mirada-solutions.com/smf.htm.

[7] Grenfell, BT, et al, Unifying the Epidemiological and Evolutionary Dynamics of Pathogens, Science 303, 327-332 (January 2004)

[8] Merikangas and Risch, *Genomic Priorities and Public Health*, Science 302, 599-601 (October 2003)

[9] see http://www.decode.com/

[10] see http://www.ukbiobank.ac.uk/science.htm

[11] see http://www.geenivaramu.ee/index.php?show=main&lang=eng

7. Genomic Medicine and Grid Computing

The full realization of the *Genomic Medicine* concept, in which genomics and proteomics are used to empower healthcare, requires the integration of knowledge from worlds traditionally apart, specially biology and medicine. To harness effectively the wealth of information available in research centres and care facilities, a new framework of computer methods and tools must be in place, bridging medical and bio informatics.

In such an approach, all levels of information – from the molecule to the population, through the cell, the tissue, the organ and the patient – and the most appropriate techniques and methods would be used. Some would come from bioinformatics and others from medical informatics or even public health or epidemiological informatics (cf. Table 1).

7.1. DEVELOPMENTS IN GENOMICS AFFECTING CARE DELIVERY

The completion of the Human Genome Project (HGP) is seen for medicine as a source of new knowledge to understand the relationships between the structure of human genes, environmental factors and physiopathological processes [1]. In the post-genomic era, the possibility of studying all the genes, all the proteins or a high number of mutations in human cells paves the way to hitherto infeasible research methods to understand the molecular basis of complex diseases and so to facilitate the development of new diagnostic and therapeutic solutions [2].

Genomic medicine will impact care provision in different ways:

Clinical diagnosis: New high-performance research devices (biochips) make it possible to monitor simultaneously a large number of parameters that can be used as diagnostic markers. Genetic analyses are used to identify individuals who are likely to contract a disease, as well as to confirm a suspected mutation in an individual or a family, before any associated symptoms appear [4]. Proteomics will also offer new markers of interest for patient monitoring [5].

Disease reclassification: Comparison of different gene expression profiles between healthy cells and those that come from a diseased tissue allows in some cases the identification of different molecular shapes and the proposal of new classifications for the diseases, which will allow an improvement in their diagnoses and prognoses.

Pharmacogenetics and Pharmacogenomics: In the last few years, successful technological methods have been developed to study and apply individual variations on a molecular scale. New technologies that aid our understanding of the role of genes in diseases are providing the industry with substantial opportunities of more powerful medicines, safer drugs and better vaccines (*pharmacogenomics*) [6].

Genetic epidemiology and Public Health: The use of new genetic information technologies will make it possible to perform cost-effective screening (genetic tests) at the population level [7]. To transfer genomic knowledge to the field of public health and epidemiology, it will be important to develop efforts in associative genetics, in

genotype-phenotype population studies, and in programmes to disseminate genetic information and to train health workers.

Current research on genomic medicine is producing an enormous volume of data, requiring distribution resources to make it available worldwide and advanced computational tools to analyse it [8].

7.2. THE CONVERGENCE OF BIO- AND MEDICAL INFORMATICS

The term 'biomedical informatics' is increasingly being used in conferences and articles, indicating the space where the disciplines of medical informatics and bioinformatics meet and interact.

State of the art methods in bioinformatics include internet data banks, from which the whole scientific community can benefit. However, present informatics tools appear to lack the necessary methods and features effectively to link genetic and clinical information and, beyond those, existing genetic databases and their possible health applications [9].

Information management tools are necessary to convert the enormous amount of data that geneticists and molecular biologists can obtain at their labs in information that physicians and health workers can use. The challenge now is to find the appropriate technologies to transform biomedical breakthroughs into shared knowledge, facilitating diagnostic and therapeutic solutions.

Though it is currently difficult to predict the health problems that a single gene or protein mutation can produce and how to translate that knowledge into new clinical procedures, it is clear that genes interact with many other genes and environmental factors. Only combined studies of gene interactions in humans and other animals and large epidemiological studies from many different populations can reveal the complex pathways of genetic diseases.

Progress in the understanding of the genetic code, gene products and functions, is elucidating the mechanisms underlying diseases. The holistic view of a person's health is built up from the integration of different sources of knowledge, combining both clinical and genetic information. Biomedical information resources available to researches and practitioners include patient data and conditions, genome and sequences, protein sequence and structure, mutations, genetic diseases, genetic tests, terminology and coding systems, patient counselling resources, and more.

Navigating between *phenotype* and *genotype* in clinical settings means that genetic assessment will be integrated in patient investigations. This vision requires the design and implementation of computer methods and tools to deliver effective platforms for seamless biomedical data association. The integration of biomedical knowledge resources brings up a new problem domain with some specific challenges to be addressed:

- There are many different sources of information spread over the web; the relevant information needs to be modelled, discovered, accessed and retrieved.
- Data integration is difficult since databases can present a wide range of formats and different semantics. In addition, public information resources are often only available through web interfaces, not easily interrogated by computer applications.

- Coding and terminologies are not unified, so that it is sometimes difficult to discern quality and link related concepts. Gene naming, for example, is far from being unified.
- Medical coding systems are not ready to manage the emerging genetic information.
- Intellectual property rights, privacy and confidentiality issues and protection of the ownership of valuable data may hinder the exchange of contents.
- Results are often published in natural language formats (scientific bibliography), requiring mining techniques to recover the knowledge in computer ready representations.
- The amount of data available and being produced is tremendous, requiring high-performance computer storage, processing power and networking infrastructures to ensure that it is effectively communicated, managed and exploited.

Health Information level	Classical health informatics applications	New genomic data and information	New health informatics applications
Population	• Public Health & epidemiology databases • Technology assessment, outcomes research	• Genome epidemiology • Genetic Screening	• Genome epidemiology databases and network (CDC-HuGeNet)
Disease	• Disease classification systems • Computerized clinical practice guidelines (CCPGs) • Information systems in clinical trials	• New classification of disease based on its molecular causes • Genetic-based decision making • Clinical trials in pharmacogenetics	• Decision-making support tools • Molecular classification of disease • CCPGs including genetics tests and therapy follow-up based on genetic data • Pharmacogenetics databases
Patient	• Computerized patient health record (CPHR)	• Genetic individual profiles (SNPs, mutations)	• Genetic data in the CPHR
Tissue, organ	• Pathology lab systems, medical image processing	• Physiological genomics • Genetic networks	• Tumour databanks • Disease models
Cell	• Imaging in Cytogenetics, histology • Microbiology lab information systems	• Gene expression profiling • Proteomics	• Molecular imaging • Information systems in pharmacogenomics (drug R&D)
Molecule	• Biochemistry and genetic tests and laboratory information management systems	• DNA and protein sequences • Macromolecular structures	• Facilitating integrated and guided access to relevant genomic databases to health professionals

Table 1: Synergy between medical informatics and bioinformatics to build broader views and raise opportunities in health informatics (cf. [10])

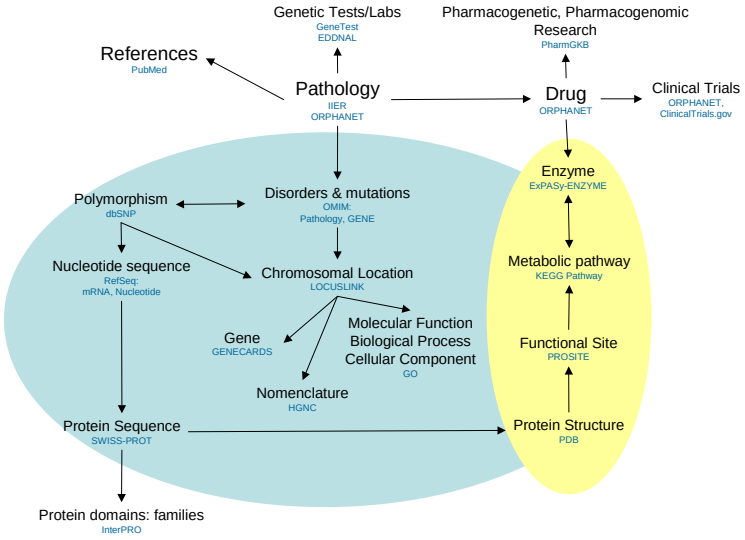

Figure 1: A conceptual framework for the study of genetic disorders.

This figure illustrates a possible protocol to guide a researcher or practitioner on obtaining pertinent information on a disease, as follows: A professional would start by searching by **pathology** name. This search could be performed on the *OMIM* database, publicly available on the Internet. The pathology is due to a **mutation** (information available at *OMIM*) or to a **polymorphism** or **SNP** (information available at *dbSNPs*). SNPs are within a **nucleotide sequence** (*RefSeq*) which in turn is in a gene (Genecards). This **gene** has a **chromosomal localization** (*LOCUSLINK*), an approved name (*HGNC*) and a **molecular function** found within Gene Ontology (*GO*). The gene codes for a **protein**, a **sequence** of amino acids (SWISSPROT). The sequence determines the **structure** of the **protein** (*PDB*). The protein is classified into **protein domains** (*InterPRO*) and has a **functional site** (*PROSITE*). Proteins have **enzymatic** properties (*ExPASY-ENZYME*) in **metabolic pathways** (*KEGG*). Drugs are **chemical compounds** (*Orphanet*) that are developed through **pharmacogenetic research** (*PharmGKB*) and validated in **clinical trials** (*ClinicalTrials.gov*). Most of the entries described can directly link to bibliography in life sciences (*PubMed*).

7.3. SEMANTIC INTEGRATION OF BIOMEDICAL RESOURCES

Biomedical resources are usually unrelated to each other, though the contents they hold are strongly and semantically connected. Bringing together such knowledge is a complex task, since it is difficult automatically to make the semantic connections.

The semantic integration of such resources would be one of the enabling factors to promote the deployment of novel biomedical applications involving research-oriented competence centres, specialized core facilities and laboratories (such as micro-chip array, mass spectrometry, etc.), and health centres where clinical guidelines are applied, such as hospitals. The main goals of semantic integration of biomedical resources are:

- to allow coherent access to biological, biomedical, bioinformatic, medical and clinical resources, especially data sources, such as bioinformatics data banks (e.g. SwissProt, Protein Data Bank - PDB) and Electronic Patient Record systems (EPR) [11];

- to facilitate the discovery and exploitation of intra- and inter-data source semantic relationships (e.g. a protein sequence in SwissProt is related to a protein secondary structure in PDB or a 3D shape of a protein in PDB can be bound to a drug compound of a ligand database).

The semantic integration of biomedical resources can benefit from existing standards, applying emerging knowledge management and modelling methodologies and technologies, such as Data and Text Mining, Document and Content management systems, ontologies, relational databases, semi-structured databases, modelling languages, and metadata management. The main services that compose semantic integration framework include:

- semantic modelling of different biomedical concepts and resources using ontologies (such as GeneOntology [12]) and metadata;
- semantic annotation of biomedical resources, to allow a continual knowledge exchange between data sources and users (researchers, doctors, physicians, etc.);
- discovering, browsing and querying of biomedical resources, offered both to human users and to computer programs, driven by semantic concepts other than keywords;
- semantic modelling of medical documentation through different types of metadata: media-type dependent, content-descriptive, content classification, document composition, document history, document location.

Recent advances in grid technology are in line with semantic integration needs. Emerging grid infrastructures include:

- Web services that allow the discovery, invocation and execution of distributed services, and could be used to implement some basic biomedical services and applications;
- Grid-based DBMSs and metadata management systems. In order to provide a secure, efficient, and automatic data source management in a Grid environment a new concept can be introduced: the Grid-DBMS [13].
- Support for Virtual Organization clusters through basic Grid services, such as security, and tools and platforms for cooperation.

Semantic integration involves both modelling and technology. While the former allows for the deployment of high level semantic services and applications, the latter can enhance performance and efficiency on distributed and Grid environments.

7.4. BIOMEDICAL GRIDS FOR HEALTH APPLICATIONS

Many research and development areas of informatics are needed to support genomic medicine, including the development of models and digital simulations, molecular imaging, global scale data access and association, etc. [14]. Grid technology is among these and can contribute to the development of some key areas by (1) supplying high computing power, (2) enabling seamless access and integration of complex and distributed data sources, and (3) establishing collaborative Virtual Organizations in order to enhance human-to-human interactions [15] [16].

Expected contributions of grid technologies to the realization of genomic medicine include:

1. ***Computational genomics and proteomics*** in the identification of genes and proteins, automatic annotation and characterization of genetic individual variations (e.g. virtual laboratories of genetic information)
2. ***Technologies to store large amounts of phenotype, genotype and proteotype data*** in meta-relational databases.
3. ***Support to the development of clinical trials.***.
4. ***Provision of personalized healthcare services*** through genetic profiling of patients, understanding heredity, coherent clinical observations, epidemiological studies, and statistical analysis.
5. ***Development of models and digital simulations of cells and diseases.*** Link gene expression patterns with disease models to uncover pathogenic pathways related with the patient's clinical condition, life-style, nutrition, and genetic disposition. Ubiquitous access to the whole history of health of a person, independently of the centre where there has been gathered information of the clinical episodes. 3-D models (of the body, cells, etc), combining anatomic and functional parameters, can be built to implement metabolic pathways and processes, linking structural information with cell assembly information. With the appropriate computer resources, gene sequences, functions, pathophysiological processes and clinical manifestations could be progressively integrated in a unified abstraction. This functional model could provide biomedical researchers and health educators and professionals with a reference for their routine work. These systems will be used in the assessment of the effects of a toxic agent or of the action that a given drug triggers in the cellular response against a disease. (e.g.: [17]).
6. ***Providing tools to support physicians' training and to improve biomedical knowledge management.*** Most physicians have only a rudimentary understanding of genetics and genomics. E-learning tools may be decisive by introducing an easy and rapid means to adopt new methods and new perspectives in routine work and the adoption of genomic medicine. These collaborative e-learning tools would share computational resources such as data files and simulations and are themselves candidates to exploit grid technology, e.g. to integrate and share features. Thus a goal must be to provide e-health portals, oriented towards the resolution of problems by use of distributed applications.

7. ***Molecular imaging.*** The new field of functional and molecular imaging arises from the combination of medical imaging technologies with genomic approaches. This area can increase the diagnostic arsenal by means of *in vivo* visualisation of cellular and genetic processes. Molecular imaging developments pursue quantitative and non-invasive studies of diseases at the molecular level. Grid can provide the processing power needed in this area.
8. ***Genetic epidemiology.*** Population studies may be undertaken in which the influence of environmental and genetic factors in particular diseases are explored. The information

sources needed to perform such studies are spread in different and remote sites. Grid infrastructures can facilitate seamless access to all these resources.

9. ***Development of Pharmacogenomics.*** Drug design can be revolutionized through the a new reasoned approach using gene sequence and protein structure function information rather than a traditional trial-and-error method. A new generation of data models and repositories will be needed to handle the complex spectrum of information sources needed in these approaches (laboratory measures, clinical findings, human genetic variation, chemical compounds, and metabolic pathways). Grid offers services that assist in the management of this diversity of information sources.
10. ***Developing tools that support clinical decision making***, combining multiple relevant information sources (genetic, clinical and environmental). In a genomic medicine framework, medical practitioners will access biological information and integrate it with data included in computerized patient records or departmental systems in large hospitals. Grid could help to integrate all the data used in decision-making and to build the computing power needed to run real time, complex interactive systems.
11. ***Integrating databases and knowledge between the clinical world and that of genomic research.*** Biomedical research is a collaborative science, in which multidisciplinary teams join skills and resources. Often, this research comprises multiple institutions and sets up virtual organizations. Partners engaged in biomedical research need a computational infrastructure that can support this kind of collaboration and sharing of information systems, often 'legacy' systems, heterogeneous and decentralized. In addition, progress in life sciences depends on the ability to develop common representations (ontologies, integrated vocabularies, etc.) to model and describe heterogeneous information. The challenge is to adapt existing systems or to develop new ones that allow the exchange and integration of data. Grid, enhanced with semantic integration services, can help not only in the sharing of computer resources, but also to integrate genetic data obtained from functional and comparative (individual) genomics into clinical information systems.

7.5. REQUIREMENTS AND ARCHITECTURES OF BIOMEDICAL GRIDS

The way data at different levels of the grid can be effectively acquired, represented, exchanged, integrated and converted into useful knowledge is an emerging research field known as "*Grid Intelligence*" [19]. In particular, ontologies and metadata are the basic elements through which Grid Intelligence services can be developed [20]. Using ontologies, Grids may offer semantic modelling of user's tasks/needs, available services, and data sources to support high level services and dynamic services finding and composition. Moreover, data mining and knowledge management techniques could enable novel services based on the semantics of stored data. *Semantic Grid* focuses on the systematic adoption of metadata and ontologies to describe grid resources, to enhance and automate service discovery and negotiation, application composition, information extraction, and knowledge discovery [21]. *Knowledge Grids* [22] offer high-level tools and techniques for distributed mining and extraction of knowledge from data repositories available on the grid, leveraging

semantic descriptions of components and data, as provided by Semantic Grid, and offering knowledge discovery services.

Biomedical Grids must be able to produce, use and deploy knowledge as a basic element of advanced applications and will be mainly based on *Knowledge Grids* and *Semantic Grids*. Leveraging their high level services, it will allow delivery of information, knowledge, medical guidelines, and research results in an applicable form to the right user, in the right setting. The Cancer Biomedical Informatics Grid (*caBIG*), a cancer-based biomedical informatics network developed by the National Cancer Institute (www.nci.nih.gov), goes along this direction. *caBIG* will connect cancer related data sources, tools, individuals, and organizations, and will help redefine how research is conducted, care is provided, and patients and participants interact with the biomedical research enterprise (cabig.nci.nih.gov/caBIG/overview/).

Biomedical Grids may help in storing, integrating, and analysing the data produced or used (e.g. provided by public databases) in the experiments and research activities. Moreover, they will support the modelling, designing and execution of workflow experiments (e.g. "in silico" experiments), by using standard modelling techniques such as UML, ontologies, and workflow languages. Main conceptual layers of Biomedical Grids include:

- *Data sources and modelling layer*. The data sources, comprising data produced during experiments (e.g. mass spectrometry, microarray, and so on), data provided by public databases (e.g. PDB, SwissProt), and data coming from clinical practice, need to be modelled using well established and novel knowledge management methodologies, such as UML and ontologies. Data sources need to be integrated and federated to allow easy access to specific information or to data semantically correlated. Main tasks of this layer are: ontology-based modelling of biological/biomedical databases; modelling of distributed biomedical applications, such as in-silico experiments. The modelling should comprise all phases of experiments, such as sample preparation, data generation, data pre-processing and filtering, images analysis, bioinformatics analysis, bio-medical analysis, results visualization [23].
- *Application composition and enactment layer*. This workflow composition layer makes it possible to realize complex bioinformatic and biomedical applications (e.g. *in silico* experiments) by composition of basic (open source) bioinformatics tools, that will be executed on the grid, exploiting the resources and data provided by research centres forming different Virtual Organizations. Useful software tools need to be classified in the modelling layer of the platform, with respect to technology and use aspects. Key issues of this layer are: domain ontologies to model (open source) bioinformatics software components, and public available biological databases; ontology-based querying and browsing on domain ontologies for the discovery, selection, and location of bioinformatics and biomedical resources (data and software components), to be used in the composition of applications; workflow-based modelling and scheduling of distributed applications on the Grid; extensive use of Open Source software components and components provided by the research centres.
- *Data analysis and knowledge extraction layer*. In this layer advanced data analysis tools, composed using the workflow technologies, allow the extraction of knowledge

useful for prosecuting experiments. This layer should comprise a set of data analysis plug-ins using different methodologies and approaches, for example: statistical analysis and data mining; survival analysis and other temporal data analysis; visualization of multidimensional data; classification of data, and so on (e.g. KNOWLEDGE GRID [22], PROTEUS [24]).

7.6. THE ROAD AHEAD FOR GRID-ENABLED GENOMIC MEDICINE

Grid is an emerging technology, still in its infancy. The road ahead is uncertain, but it is possible to set up a very general roadmap for its successful application in the area of genomic medicine. Some of the required steps include:

1. Developing the specific semantic grid services required for a knowledge integration environment.
2. Deploying and testing the first grid middleware prototypes for the health sector (research and care provision).
3. Developing, deploying and testing the first grid genomic medicine applications.
4. Fostering and promotion of the grid culture by means of the education and training of the physicians, scientists and other staff involved in genomic medicine.

7.7. REFERENCES

[1] F. S. Collins and V. A. McKusick (2001) "Implications of the Human Genome Project for medical science", JAMA, (285): pp. 540-4.

[2] D. J. Weatherall (2003) "Genomics and global health: time for a reappraisal", Science, (302): pp. 597-599.

[3] A. Tefferi, M. E. Bolander, S. M. Ausell, E. D. Wieben and T. C. Spelsberg (2002) "Primer on medical genomics. Microarray experiments and data analysis", Mayo Clinical Proceedings, 77(9): pp. 972-940.

[4] J. R. Nevins, E. S. Huang, H. Dressman, J. Pittman, A. T. Huang and M. West (2003) "Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction", Human Molecular Genetics, (12): pp. 153-157.

[5] K. K. Jain (2002) "Role of proteomics in diagnosis of cancer", Technological Cancer Research Treatments, 1(4): pp. 281-286.

[6] J. Licinio and M.-L. Wong (Eds.) (2002) Pharmacogenomics: the search for Individualized therapies.

[7] J. S. Ross, G. P. Linette, J. Stec, E. Clark, M. Ayers and N. Leschly (2004) "Breast cancer biomarker and molecular medicine", Expert Revisions in Molecular Diagnosis, 4(2): pp. 169-188.

[8] A. Bayat (2002) "Science, medicine, and the future: Bioinformatics", British Medical Journal (BMJ), 324.

[9] A. S. Pereira, V. Maojo, F. Martin-Sanchez, A. Babic and S. Goes (2002) "The INFOGENMED project" In ICBME 2002, Singapore.

[10] F. Martin-Sanchez, V. Maojo and G. Lopez-Campos (2002) "Integrating Genomics into Health Information Systems", Methods of Information in Medicine, 41: pp. pp. 25-30.

[11] R. Sokolowski (1999) "Expressing Health Care Objects in XML" In EE 8th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, Palo Alto, California, pp 341-342.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000) "Gene Ontology: tool for the unification of biology", Nature Genetics, 25(25-29).

[13] G. Aloisio, M. Cafaro, S. Fiore and M. Mirto (2004) "The GRelC Project: Towards Grid-DBMS" In Parallel and Distributed Computing and Networks (PDCN) - IASTED, Innsbruck, Austria.

[14] BIOINFOMED (2003) Synergy between Medical Informatics and Bioinformatics: Facilitating Genomic Medicine for Future Healthcare, BIOINFOMED Study, Report White Paper.

[15] I. Foster, C. Kesselman and S. Tuecke (2001) "The Anatomy of the Grid: Enabling scalable virtual organizations", International Journal of High Performance Computing Applications, 15(3): pp. 200-222.

[16] I. C. Oliveira, J. L. Oliveira, F. Martin-Sanchez, V. Maojo and A. S. Pereira (2004) "Biomedical information integration for health applications with Grid: a requirements perspective" In HealthGrid 2004, Clermont-Ferrand, France.

[17] G. Berti, S. Benkner, J. W. Fenner, J. Fingberg, Lonsdale, S. E. Middleton and M. Surridge (2003) "Medical Simulation Services via the Grid" In Nörager, S., Healy, J.-C. and Paindaveine, Y. (Eds) 1st European HealthGrid Conference, Lyon, France.

[18] M. Cannataro and D. Talia (2004) "Semantic and Knowledge Grids: Building the Next-Generation Grid", IEEE Intelligent Systems (ISSI-0095-1203) - Special Issue on E-Science, 19(1): pp. 56-63.

[19] N. Zhong and J. Liu (Eds.) (2004) Intelligent Technologies for Information Analysis, Springer Verlag (to appear).

[20] T. R. Gruber (1993) "A translation approach to portable ontologies", Knowledge Acquisition, 5(2): pp. 199-220.

[21] D. d. Roure, N. R. Jennings and N. Shadbolt (2003) "The Semantic Grid: A future e-Science infrastructure" in Grid Computing: Making The Global Infrastructure a Reality, (Eds.) Berman, F., Hey, A. J. G. and Fox, G., John Wiley & Sons:pp. 437-470.

[22] M. Cannataro and D. Talia (2003) "KNOWLEDGE Grid An Architecture for Distributed Knowledge Discovery", CACM, 46(1): pp. 89-93.

[23] C. F. Taylor (2003) "A systematic approach to modelling capturing and disseminating proteomics experimental data", Nature Biotechnology, 21: pp. 247-254.

[24] M. Cannataro, C. Comito, F. Lo Schiavo and P. Veltri (2004) "Proteus, a Grid based Problem Solving Environment for Bioinformatics: Architecture and Experiments", IEEE Computational Intelligence Bulletin, 3(1): pp. 7-18.

8. Healthgrid Confidentiality and Ethical Issues

In healthcare, patients' sensitive personal data is recorded and used. This implies a need for strict confidentiality and enforced protection of privacy. These requirements have not previously been dealt with in grid technology, as a consequence of the fact that in High Energy Physics, the root of much grid technology, elementary particle data needs no privacy protection, unlike humans in a modern society.

Biomedical data often includes very sensitive information about a subject and although generally used for the benefit of the community, this information is still prone to abuse. There is appropriate concern about the proper treatment of sensitive data. Incidents of abuse have been previously reported in the public media [L03], proving that the threat is genuine. Consider, for example, the impact on society if banks, insurance companies and employers, could access healthcare data about their customers, revealing past, current, and probable future health status. Indeed, abuse of medical data can affect all of us, as at some point in life practically everyone has to complete loan, insurance or job applications.

It is clear that privacy protection directly impacts personal well-being as well as society as a whole. Indeed, some go as far as to believe that failure to protect privacy might lead to our ruin [C03]. Privacy is recognized as a fundamental human right. Public authorities are sharply aware of these repercussions, and they are putting considerable effort into privacy protection legislation [EU95][EU02]. Because of the possibilities opened up by modern grid technology (such as trans-border processing of sensitive data), studies regarding legal constraints in a healthgrid are of great importance (see Chapter 9).

Medical practice and research have always adhered to strict ethics. These domains are accustomed to supervision by (ethical) institutional review boards which enforce such requirements as obtaining informed consent from patients [M01]. Scientists and technicians developing grid technology are often unfamiliar with concerns about the proper treatment of information, but healthcare professionals are very conscious of this requirement. The privacy and legal issues raised by healthgrids mainly arise through the transparent interchange and processing of sensitive healthcare information, resulting from the aim of removing the line between local and remote resources with grid technology. These problems are certainly not entirely new to medical informatics. It is therefore of utmost importance that experts share their experience on security and privacy related issues in healthcare, in order to avoid that these become barriers for the realization of the healthgrid.

8.1. PRIVACY PROTECTION, SECURITY AND THE HEALTHGRID

8.1.1. Grid Security Technology

From the very start, the grid community has put a lot of effort into the design of security measures [W03]. **Authentication** and **authorization** mechanisms are the main point of focus of these developments, as they are the most basic of security measures. Integration at the level of the lower middleware allows security mechanisms to be uniform

(developer APIs) and interoperable (cf. [GLOBUS]). Implementation is still at an early stage. It is important to realize that the further development of security technology is key to the acceptance of the healthgrid concept.

Avoiding unauthorized access to sensitive data is the first level of confidentiality protection. In healthcare, state-of-the-art security solutions have always been used. An equal level of protection will be demanded from a grid environment. Any healthgrid initiative should therefore be aware of the latest security developments in the grid community. Development of basic services, such as for example integration on a lower middleware level of fine grained access control (e.g. provided by CAS or VOMS grid solutions), should be encouraged by the biomedical community.

A specific healthgrid initiative should enable the further development and testing of these security mechanisms, beyond the point where classical grid developers may stop, believing that for their application sufficient measures are already in place.

The security technology currently present in the grid community might even offer a sufficient solution for the first and most obvious healthcare applications: computational problems in healthcare. Deployment of computational grids in healthcare is a reasonable first step towards a true healthgrid – though it is only a first step. The problems faced there are similar to the ones encountered in more classical grid domains.

Unlike many other areas of healthcare, confidentiality in such cases is usually of secondary importance. The nature of the application itself reduces the risk of disclosure of sensitive information. Computational challenges inherently segment the processed data and typically only deal with non-identifiable data related to complex computational models. Thus, the similarity with classical grid applications persists also in the security domain, there is no real need for specialized 'information security'.

8.1.2. Healthgrid Security Requirements

Healthgrid will not restrict to the use of grid technology for distributed computing only. Eventually, healthgrid should offer a generic platform for all e-health actors. Hence, thye sharing of large amounts of distributed heterogeneous (on various levels) data is also an important issue.

It is clear that linking several distributed data sources bound to a single individual on a data grid opens of up a range of privacy risks. The (virtual) federation of a large amount of personal medical data is not the only risk at hand. Grid technology will undoubtedly further stimulate the use of genomic data in research. However, this particular type of data has a number of specific characteristics related to privacy which are not found in any other type of (medical) information:

- Genetic data not only concerns individuals, but also their relatives. A person's consent to release his or her genetic information constitutes a *de facto* release of information about other individuals, i.e. his or her relatives. In the case of genomic medicine, there is a complex interaction between individual rights and collective requirements.

- Medical data deals with the past and current health status of persons, but genetic information can also give indications about future health or disease conditions.

- An individual person's genotype is almost unique and stable; hence it can become the source of an increasing amount of information.
- The full extent of the information included in the genomic data is not known yet; hence it is difficult to assess the full extent of disclosure.
- Genomic data is easily wrongly interpreted by non-professionals; 'susceptibility' to diseases can easily be mistaken with certainty of illness.

The above clearly indicates the need to reconcile two seemingly conflicting objectives: on the one hand, the maximization of healthcare opportunities and of medical research productivity and efficiency in data handling; on the other, the protection of the human (privacy) rights; this is the challenge at hand.

A couple of basic approaches to safeguarding confidentiality have been identified in the past in healthcare practice. The first approach focuses on the creators and maintainers of the information, prohibiting them from disclosing the information to inappropriate parties. Basically, this comes down to the deployment of classical security measures (**access control, authorization**). A healthgrid initiative is ideal for the further development (and actual implementation) of grid security technology, because of the strict requirements in healthcare. A first task within the healthgrid context could thus be performing an in depth analysis of the new and specific risks and threats that arise.

8.1.3. Privacy Enhancing Technology

Technology which is specifically designed to safeguard privacy is generally referred to as Privacy Enhancing Techniques or Technologies (PETs). According to one author, PETs can be described as [B01]:

> *'A coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the information system.'*

Privacy Enhancing Technologies are fairly new – the concept has only been around since the '90s – and have been extensively researched in both the USA and in Europe.

In healthcare, PETs are mainly used for privacy protection of persons included in medical data collections. The goal of these PETs is to guarantee anonymity of data subjects while making information available for clinical practice and research. The use of such techniques in healthcare has been demonstrated in several research projects [DC02] and solutions are already commercially deployed, in clinical trials, disease studies, for the exchange of research data and for the daily handling of sensitive data. PETs such as anonymization have already been considered for standardization (introduced as a working item in CEN/TC251).

For healthgrid, access to large amounts of useful, personal information can be unlocked though the use of privacy protection techniques (mainly de-identification methods) [DC04].

8.1.4. Grid Integration of PETs and Security

Security and privacy protection techniques are closely linked. Emphasis of the latter however lies on limiting the identifiable information content of the data rather than on merely restricting access to the data itself. Although the strict difference between the two is not always clear, Privacy Enhancing Technology and security technology should be regarded as complementary in safeguarding the confidentiality of personal information.

The question whether these specific security techniques and privacy protection measures should be integrated in the healthgrid itself, is a valid one. It is beyond doubt that all healthgrids need to take into account the stringent data protection requirements of the healthcare sector. However, these measures could be implemented completely separately from the grid nature of an application. In that case there would be little difference with current *ad hoc* solutions (privacy-aware health data collection unrelated to grid technology).

On the other hand, the integration of specific privacy protection solutions into grid services could offer considerable advantages. Integration is not only logical because of the close relationship with classical measures (which are largely part of the grid middleware), but can also stimulate the use of privacy protecting technology leading to data protection 'by default' in each healthcare related grid application. Integration of PETs into the lower middleware level should probably be limited (in that context, see further, policy management). Lower middleware (such as Globus) aims at providing a broad generic toolbox for grid development. Specific biomedical informatics security and privacy are not a primary objective for middleware developers.

Just as in several data integration initiatives, healthcare specific security and privacy solutions could be offered at an upper middleware level, combining the advantage of still being generic (at the disposal of a wide community), but not overloading the toolset for other areas of research which do not need such strict measures.

The main part of privacy protection measures will, at least in the beginning, be situated at the application level. This does not imply that development is beyond the scope of a healthgrid initiative. On the contrary, next to the fact that stringent data protection is a prerequisite for healthcare IT, standardization of PET technology can be encouraged by the development of specific grid services, such as a policy-driven pseudonymization service which allows centres automatically to de-identify their databases through a grid service (guaranteeing use of the latest technology) before exchanging information with another site.

As developments and pilot projects progress, it will become clear which piece of technology should be implemented at what level.

8.1.5. Healthgrid Issues

In order to illustrate the need of specific research in any healthgrid initiative, some typical problems due to the strict requirements of the medical world will be given. The examples presented here are fairly straightforward and thus have been identified before [GK02]. However they have not been adequately dealt with. With the introduction of a healthgrid, the need for confidentiality and data protection is more pressing than ever.

The grid promises access to heterogeneous resources, so that in a healthgrid remote resources will be storing and processing sensitive personal data. These resources should

thus be trusted by the end-user. But who can be the judge of 'trustworthiness' of a grid resource? A simple and straightforward solution is to use 'closed' systems, which means that any resource in the grid is well known and specified in advance. This however conflict with the vision of a dynamic grid, in which links are established as necessary.

Solutions should rather be sought in the area of policy advertising and negotiation. Resources should be able to inform a candidate user on how the data will be treated, which policies are applied, what PETs are used, who can have access to the data, etc. These methods are sometimes said not to be genuine PETs, since they do not limit collection of personal identifiable data and do not give any guarantees about the actual processing. A resource can claim to adhere to strict rules, but in practice this can not be verified.

The first steps in the direction of policy management have already been taken by grid developers. The development of standards such as WS-Privacy, WS-Policy and Enterprise Privacy Authorization Language (EPAL) is an effort in that direction, but implementation to date is rather limited, and the full possibilities of the technology will not be researched unless it is in the healthcare area – the main application domain. A healthgrid would be the ideal environment where such PETs could be tested and further developed.

These considerations directly impact typical grid mechanisms, such as data replication. Replication mechanisms automatically copy data on a resource in order to increase efficiency (e.g. to avoid transfer delays). With medical data, this may not be permitted. The site on which the data will be replicated should at least be as trustworthy as the data source and should adhere to the same strict policies. A healthgrid should be able to handle such cases autonomously in order not to lose its dynamic nature (and efficiency).

Another example is delegation. Delegation of rights is fundamental in a grid environment, but in the medical world, this is far from obvious. If one passes on rights to others (resources), one becomes liable for actions performed on one's behalf. In a healthcare environment this has serious implications in terms of liability. Restricted proxy certificates offer a path to a solution suitable for medical applications, but clearly need to be extended.

Policy management will be an important topic in healthgrid, both for security (e.g. authorization policies) as for data protection (privacy policies). A difficult problem in this context is the one of policy enforcement and assurance.

Equally important and closely related to this subject, is the implementation of **auditing mechanisms**. All actions in a medical context should be logged in a trustworthy way. Non-repudiation combined with a legal framework could help solve liability issues in healthcare.

Next to the areas of interest mentioned in this text, there are several other healthcare needs for grid applications which could be developed at, e.g., upper middleware level for the benefit of a large community within a healthgrid context. Among these are encrypted storage for medical data (a far from obvious problem) and trustworthy federation of research databases – virtual federation of small 'cells' of de-identified data (e.g. geographical area or hospital) can decrease the re-identification risk (by increasing the anonymity set). Finally a range of PETs which are well suited to distributed environments is emerging – Private Information Retrieval and Storage (PIRS) which includes privacy-preserving data mining, processing of encrypted data, and other related technologies. However the road to an advanced generic privacy preserving framework for e-health is still long and littered with technical difficulties which will have to be tackled one at a time.

It is however a fact that grid technology can only be successful in a biomedical environment if the ethical guidelines and legal requirements are adequately met by technological solutions which are continually evaluated and updated as new needs arise.

8.2. REFERENCES

[G96] Goodman KW. Ethics, Genomics, and Information Retrieval. Comput. Biol. Med. 1996; vol 26, no.3:223-229.

[M02] Martin-Sanchez F. Integrating Genomics into Health Information Systems. In: Methods Inf Med 2002; 41:25-30.

[LCG] Website: http://lcg.web.cern.ch/lcg/

[L03] Lazarus D. A tough lesson on medical privacy: Pakistani transcriber threatens UCSF over back pay. San Francisco Chronicle Wednesday, October 22, 2003

[C03] Caloyannides M. Society Cannot Function Without Privacy. IEEE Security & Privacy, May-June 2003 (Vol. 1, No. 3).

[EU95] Directive 95/46/EC of the European Parliament and the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[EU02] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)

[M01] Mehlman MJ. The effect of Genomics on Health Services Management: Ethical and Legal Perspectives. Frontiers of Health Services Management; 17;37:17-26. 2001.

[W03] Welch, V., Siebenlist, F., Foster, I., Bresnahan, J., Czajkowski, K., Gawor, J., Kesselman, C., Meder, S., Pearlman, L. and Tuecke, S., Security for Grid Services. in 12th IEEE International Symposium on High Performance Distributed Computing, (2003).

[GLOBUS] Website: http://www.globus.org/

[IBM04] Martin-Sanchez F et al. Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. J Biomed Inform. 2004 Feb;37(1):30-42.

[HG] Website: http://www.healthgrid.org/

[B01] Borking J, Raab C. Laws, PETs and Other Technologies for Privacy Protection. The Journal of Information, Law and Technology (JILT), 2001.

[DC02] De Meyer F, Claerhout B, De Moor GJE. The PRIDEH project: taking up Privacy Protection Services in e-Health Proceedings MIC 2002 'Health Continuum and Data Exchange'. IOS Press, 2002, p. 171-177.

[DC04] De Moor GJE, Claerhout B. Privacy Protection for Healthgrid Applications (Accepted for Methods Inf Med 2004)

[GK02] Guy L, Kunszt P, Laure E, Stockinger H, Stockinger K. Replica Management in Data Grids. Technical report, Global Grid Forum Informational Document, GGF5, Edinburgh, Scotland, July 2002.

9. Healthgrid from a Legal Point of View

The introduction of grid technology in the health care sector may appear to be only of technical significance and, in any event, without any legal relevance. It appears only to concern a new computing technology participating in the provision of healthcare services and in scientific research, mostly by providing huge computing and memory resources, possibly internet based. The first projects deal with medical imaging, medical tele-assistance, medical or pharmaceutical research, human genomic studies, and the creation of databases for therapeutic, scientific, statistical or epidemiological purposes.

However these projects are ruled by radically different legal contexts. Indeed, distinct legal rules govern the practice of medicine, scientific and pharmaceutical research, epidemiological studies, even if all these disciplines contribute to medical progress.

Hence there is no unique answer to the determination of the legal framework in which healthgrid technology may be implemented and used. In reality, the answers are multiple and depend on the context of each project as well as on the considered legal viewpoints. Healthgrid technology must conform to the legal context specific to each project aiming at its implementation.

Nevertheless describing the different legal contexts in which healthgrid technology might be implemented is not sufficient. The adequacy of the legal context coupled to the characteristics of this particular technology should also be evaluated. In other words, one should question whether certain rules should not (have to) be adapted with respect to healthgrid technology.

9.1. HEALTHGRID TECHNOLOGY'S STATUS

Technologies must frequently comply with precise technical norms with a view to their legal utilization. The same assertion is also valid for the health care sector. It is therefore important to define the content of the technical norms relevant to each project.

In this matter, some technical norms have been harmonized at an international or European level. With respect to this, it is useful to note that the European Committee for Standardization has issued a very interesting study entitled "*European Standardization of Health Informatics – Results of the mandated work by CEN/TC 252*" (CEN TC 251/N01-024 – 2001-06-17).

The European Union has also adopted several rules concerning medical devices:

- Council Directive 90/385/EEC of 20 June 1990 on the approximation of the laws of the Member States relating to active implantable medical devices.
- Council Directive 93/42/EEC of 14 June 1993 concerning medical devices.
- Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on *in vitro* diagnostic medical devices.

It is hence required in each project to:

- determine the technical norms applicable to healthgrid technology in the project under consideration, depending on the national legal orders likely to rule it ;
- verify the adequacy of these technical norms.

The Council of Europe states that the improvement of human life quality and the respect of human rights should prevail when dealing with new technologies. It namely recommends in this regard that the precise evaluation of any technology should as much as possible rely on the following criteria (cf. Recommendation (90) 8 of 29 March 1990 on the impact of new technologies on health services, particularly primary health care):

- Validity of outputs,
- Validity of data capture,
- Ability to fit within the framework of primary health care,
- Social acceptability,
- Ethical acceptability,
- Professional acceptability,
- Reliability,
- Capacity for continuous assessment,
- Safety for providers, consumers and the environment,
- Cost effectiveness compared to older technologies,
- Availability of full information on the technology and experience in implementing it,
- Protection of confidentiality,
- Ability to be integrated smoothly into existing systems,
- Availability of adequate resources.

This evaluation should consist of appropriate studies giving conclusive results, and should be carried out prior to the general introduction of any new technology.

9.2. STATUS OF THE PROCESSED PERSONAL DATA

Most of healthgrid technology-related projects imply personal data processing for therapeutic purposes or scientific research (e.g. medical imaging, tele-assistance, medical or scientific research, human genomic studies, creation of healthgrid databases).

However personal data processing is subject to numerous regulations. Indeed, these data are particularly sensitive and consequently require high protection. Furthermore, because of the therapeutic or scientific stakes, personal data processing must be reliable, or it may lead to medical errors or erroneous scientific results.

On the international level many norms govern personal data processing (including the processing of personal data related to health).

Article 8 of the Convention for the Protection of Human Rights and Fundamental Freedoms is particularly to the point in this respect.

In the case M.S. v. Sweden of 27 August 1997 (74/1996/693/885) (§ 41), the European Court of Human Rights vigorously stated that "(…) *the protection of personal data, particularly medical data, is of fundamental importance to a person's enjoyment of his or*

her right to respect for private and family life as guaranteed by Article 8 of the Convention. Respecting the confidentiality of health data is a vital principle in the legal systems of all the Contracting Parties to the Convention. It is crucial not only to respect the sense of privacy of a patient but also to preserve his or her confidence in the medical profession and in the health services in general. The domestic law must afford appropriate safeguards to prevent any such communication or disclosure of personal health data as may be inconsistent with the guarantees in Article 8 of the Convention. (Case Z. c Finlande of 25 February 1997, 1997-I, p. 347, § 95)."

Article 7 of the Charter of Fundamental Rights of the European Union similarly confirms the right to privacy while Article 8 establishes the right to the protection of personal data.

The Council of Europe has issued important norms relative to personal data processing. Its Convention for the protection of individuals with regard to automatic processing of personal data (28 January 1981) (Treaty n° 108) represents a significant source for all member states.

The Council of Europe has also adopted specific recommendations concerning personal data processing involved in projects implementing healthgrid technology:

- Recommendation (83) 10 of the Committee of Ministers on the protection of personal data used for scientific research and statistics, adopted on 23 September 1983.
- Recommendation (90) 8 of 29 March 1990 on the impact of new technologies on health services, particularly primary health care.
- Recommendation (97) 5 of the Committee of Ministers to Member States on the protection of medical data, adopted on 13 February 1997.
- Convention for the protection of Human Rights and dignity of the human being with regard to the application of biology and medicine: Convention on Human Rights and Biomedicine (Treaty n° 164) (4 April 1997).
- Recommendation (97) 18 concerning the protection of personal data collected and processed for statistical purposes, adopted on 30 September 1997.
- Recommendation n° R (99) 5 of the Committee of Members to Member States for the protection of privacy on the Internet – Guidelines for the protection of individuals with regard to the collection and processing of personal data on information highways, adopted on 23 February 1999.
- Recommendation 2/2001 on certain minimum requirements for collecting personal data on-line in the European Union, adopted on 17 May 2001.

The Council of Europe recommends that specific models designed to ensure confidentiality of patient information should be developed in relation to the application of information technology to health care systems (cf. R (90) 8 of 29 March 1990, *op cit*, point 8 of the Guidelines).

In the extent of its attributions, the European Union has adopted special norms relative to personal data processing, namely:

- Resolution of the Council and of the Representatives of the Governments of the Member States, meeting within the Council, of 29 May 1986, concerning the adoption of a European emergency health card.
- Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
- Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

The European Group on Ethics has adopted an important opinion concerning the processing of personal data related to health (cf. Opinion of the European Group on Ethics in Science and New Technologies to the European Commission, Ethical issues of healthcare in the information society, n° 13, 30 July 1999).

The World Medical Association has issued several documents of interest to some healthgrid projects:

- Declaration on the patient's rights (World Medical Association Declaration on the Rights of the Patient, adopted by the 34th World Medical Assembly Lisbon, Portugal, September/October 1981 and amended by the 47th General Assembly Bali, Indonesia, September 1995);
- Guidelines concerning the practice of Telemedicine (World Medical Association Statement on Accountability, Responsibilities and Ethical Guidelines in the Practice of Telemedicine, adopted by the 51st World Medical Assembly Tel Aviv, Israel, October 1999);
- Declaration on Ethical considerations regarding Health Data Bases (adopted by the WMA General Assembly, Washington 2002);
- Declaration on Ethical Principles for Medical Research involving Human Subjects (adopted by the 18th WMA General Assembly Helsinki, Finland, June 1964 and amended by the 29th WMA General Assembly, Tokyo, Japan, October 1975 35th WMA General Assembly, Venice, Italy, October 1983 41st WMA General Assembly, Hong Kong, September 1989 48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996 and the 52nd WMA General Assembly, Edinburgh, Scotland, October 2000 Note of Clarification on Paragraph 29 added by the WMA General Assembly, Washington 2002).

National norms on personal data processing must comply with this international framework, although a certain margin is generally allowed to member states in their implementation. This may cause some disparity in national norms in this matter, adding to

the existence of national norms for which no international rules exist and upon which member states are free to decide.

In any case it is of prime interest to qualify correctly any operations carried out on personal data when using healthgrid technology and to define the role of each person involved (health care practitioners, service providers, patient, etc.).

From a technical viewpoint, PETs (see chapter 8) offer very strong support to the security and the confidentiality of the processed personal data. They aim to reduce the processing of personal data and to suggest appropriate measures to secure data processing.

9.3. HEALTHGRID SERVICES' STATUS

Some projects aim at providing services to health care professionals or to scientists. These services must be qualified according to the norms applicable to 'information society' services.

An information society service is any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services.

- ***"At a distance"*** means that the service is provided without the parties being simultaneous present. Services provided in the physical presence of the provider and the recipient, even if they involve the use of electronic devices are not provided "at a distance".
- ***"By electronic means"*** means that the service is sent initially and received at its destination by means of electronic equipment for the processing (including digital compression) and storage of data, and entirely transmitted, conveyed and received by wire, by radio, by optical means or by other electromagnetic means. Services that are not provided via electronic processing/inventory systems are not services provided "by electronic means" (e.g. telephone/fax consultation of a doctor).
- ***"At the individual request of a recipient of services"*** means that the service is provided through the transmission of data on individual request.

Information society services also include services consisting of the transmission of information via a communication network, in providing access to a communication network, or in hosting information provided by a recipient of the service.

Activities which by their very nature cannot be carried out at a distance and by electronic means, such as medical advice requiring the physical examination of a patient are not information society services.

The taking up and pursuit of the activity of an information society service provider may not be made subject to prior authorization or any other requirement having equivalent effect (art. 4.1 of D 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market – Directive on Electronic Commerce). The service provider must therefore comply with a number of special rules when offering information society services.

This provision of services may result from a contractual relationship. The latter must be analysed on an individual basis in each project. In case of an international situation, when providing information society services, one should preliminarily examine what are the

competent jurisdictions before defining the law applicable to the contractual obligations of the parties.

Several international instruments can be mentioned in this regard:

- Convention on the law applicable to contractual obligations opened for signature in Rome on 19 June 1980 (80/934/EEC).
- Directive 1999/93/EC of the European Parliament and of the Council of 13 December 1999 on a Community framework for electronic signatures.
- Directive 2000/35/EC of the European Parliament and of the Council of 29 June 2000 on combating late payment in commercial transactions.

9.4. END-USER'S STATUS

The use of healthgrid technology by health care professionals raises special questions. On one hand, is the end-user legally authorized to use the healthgrid technology? Is the use of healthgrid technology permitted in medical practice or in scientific research? The answer lies in the rules governing the professional activities of the end-user.

Concerning some projects, it is useful to remember that the European Union has adopted the Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of member states relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use.

On the other hand, in case of medical tele-expertise, medical tele-consultancy, or medical tele-assistance, involving healthcare practitioners from different member states, the question is to know if the health care practitioner in charge of the patient is legally authorized to seek the assistance of a foreign healthcare practitioner, and, if positive, under which conditions.

Simultaneously this foreign healthcare practitioner should also find out whether he is legally authorized to provide assistance to a healthcare practitioner located in another country.

Beyond the determination of the persons liable in case of medical accident or fault, one must define the status of the health care practitioner participating to the provision of health care in another member state, and the status of the healthcare practitioner having asked his assistance. This problem is far beyond the simple question of medical qualification equivalency.

In the same way, the cooperation between health care practitioners inside a same member state or from different member states raises the very delicate question of the legal framework of this cooperation.

9.5. PATIENT'S STATUS

Implicitly or explicitly all the healthgrid projects aim to participate in the search for medical progress as well as in its preventive and curative aspects. Hence the patient is very much at the heart of the implementation of healthgrid technology.

The Council of Europe is clear on the patient's interest in his active participation in his own treatment (cf. Recommendation R (80) 4). The legal qualification of the parties involved in the processing of the patient's personal data, including the place of the patient, is likely to highlight some tensions underlying the medical relationship.

9.6. LIABILITY ISSUES

The question of the determination of the persons liable in case of medical accident or fault relative to the use of healthgrid technology when providing health care to a patient is crucial but delicate. In case of an international situation, the question is far more complex. With respect to this, one should take into account several factors which are not necessarily likely to be under complete control.

The first element of uncertainty results from the determination of the possible jurisdictions likely to recognize the case. With respect to this, the European Union has recently adopted the Council Regulation (EC) No 44/2001 of 22 December 2000 on jurisdiction and the recognition and enforcement of judgments in civil and commercial matters. The determination of the jurisdiction will permit to determine the law applicable to the case.

The European Union has adopted some norms relative to the matter of liability:

- European Convention on Products Liability in regard to Personal Injury and Death (Council of Europe, Treaty n° 91, adopted on 27 January 1977);
- European Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the member states concerning liability for defective products.

It has to be remembered that the European Union has also adopted special rules concerning the resolution of disputes:

- Council Decision 2001/470/EC of 28 May 2001 establishing a European Judicial Network in civil and commercial matters. Its objectives are to improve effective judicial cooperation between member states and effective access to justice for persons engaging in cross-border litigation;
- Council Regulation (EC) No 1206/2001 of 28 May 2001 on cooperation between the courts of the member states in the taking of evidence in civil or commercial matters.

Mention should also be made of alternative dispute resolution and on-line dispute resolution.

9.7. IPR AND COMPETITION ISSUES

The creation and the use of healthgrid technologies may raise important Intellectual Property Rights (IPR) questions. Indeed, healthgrid technologies are sometimes created like patchworks. This poses the question of the IPR relative to the constitutive elements of the 'patchwork' under consideration.

The European Union has adopted several Directives concerning IPR issues:

- Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs;
- Council Directive 92/100/EEC of 19 November 1992 on rental right and lending right and on certain rights related to copyright in the field of intellectual property
- Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights;
- Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases;
- Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society;

Usually projects aiming at implementing healthgrid technology bring together several partners into consortium. Their behaviour also has to comply with competition law (Monopolistic positions, abuse of dominant position, concerted practices).

White Paper Contributors

Giovanni Aloisio
Center for Advanced Computational Technologies/ISUFI & Dept. of Innovation Engineering
University of Lecce, Via per Monteroni 73100 Lecce, Italy
giovanni.aloisio@unile.it

Siegfried Benkner
Institute of Scientific Computing, University of Vienna, Nordbergstrasse 15
A-1090 Vienna, Austria
sigi@par.univie.ac.at

Howard Bilofsky
University of Pennsylvania, School of Engineering and Applied Science
Computer and Information Science, Center for Bioinformatics
1416 Blockley Hall, 423 Guardian Drive, 19104-6021 Philadelphia, PA, USA
bilofsky@pcbi.upenn.edu

Ignacio Blanquer
Universidad Politecnica de Valencia, Camino de Vera /n, 46022 Valencia, Spain
iblanque@dsic.upv.es

Sir Michael Brady
FRS FREng, Dept. of Engineering Science, Oxford University, Parks Road, Oxford OX1 3PJ
jmb@robots.ox.ac.uk

Vincent Breton
CNRS-IN2P3, LPC, Campus des Cézeaux, 63177 Aubiere Cedex, France
breton@clermont.in2p3.fr

Mario Cannataro
Magna Graecia University of Catanzaro, School of Bioinformatics and Biomedical Engineering, Campus di Germaneto, Viale Europa Germaneto, 88100 Catanzaro- Italy
cannataro@icar.cnr.it

Ioanna Chouvarda
Aristotle University, The Medical School, Lab of Medical Informatics - Box 323
54124 Thessaloniki, Greece
ioanna@med.auth.gr

Brecht Claerhout
Custodix NV., Verlorenbroodstraat 120, Bus 14 B-9820, Merelbeke (Belgium)
Brecht@custodix.com

Kevin Dean
Internet Business Solutions Group, Cisco Systems
9 New Square, Bedfont Lakes, Feltham, Middlesex, TW14 8HA
kevin.dean@cisco.com

Georges De Moor
Research in Advanced Medical Informatics and Telematics
UZ Gent, 5K3, De Pintelaan 185, B9000 Gent,Belgium
georges.demoor@UGent.be

Wilfried De Neve
Department of Radiotherapy, Building P7, Ghent University Hospital
De Pintelaan 185, B-9000 GENT, Belgium
wilfried@krtkg1.rug.ac.be

Carlos De Wagter
Department of Radiotherapy, Building P7, Ghent University Hospital
De Pintelaan 185, B-9000 GENT, Belgium
Carlos.DeWagter@UGent.be

Sandro Fiore
Center for Advanced Computational Technologies/ISUFI & Dept. of Innovation Engineering
University of Lecce, Via per Monteroni, 73100 Lecce, Italy
sandro.fiore@unile.it

Kinda Hassan
Laboratoire LIRIS, Université Lumière Lyon 2, Campus Porte des Alpes
5, avenue Pierre Mendès France, 69676 Bron, France
khassan@dionysos.univ-lyon2.fr

Germaine Heeren
European Society for Therapeutic Radiology and Oncology (ESTRO)
av. E.Mounierlaan 83, 1200 Brussels, Belgium
germaine.heeren@skynet.be

Vicente Hernández
Universidad Politecnica de Valencia, Camino de Vera /n, 46022 Valencia, Spain
vhernand@dsic.upv.es

Jean A.M.Herveg
Faculté de Droit de Namur – FUNDP, Centre de Recherches Informatiques& Droit
5, rempart de la Vierge, 5000 Namur, Belgium
jean.herveg@fundp.ac.be

Martin Hofmann
Department of Bioinformatics, Fraunhofer Institut for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany
martin.hofmann@scai.fhg.de

Chris Jones
CERN, 1211 Geneva 23, Switzerland
chris.jones@cern.ch

Vassilios Koutkias
Aristotle University, The Medical School, Lab of Medical Informatics - Box 323
54124 Thessaloniki, Greece
bikout@med.auth.gr

Sharon Lloyd
Oxford University Computing Laboratory, Wolfson Building, OX1 3QD Oxford, UK
sharon.lloyd@comlab.ox.ac.uk

Guy Lonsdale
C&C Research Laboratories, NEC Europe Ltd. Rathausallee 10, D-53757 Sankt Augustin, Germany
lonsdale@ccrl-nece.de

Victoria López Alonso
Medical Bioinformatics Department, Institute of Health "Carlos III"
Ctra. Majadahonda a Pozuelo, Km.2, 28220 Majadahonda, Madrid, Spain
victorialop@inia.es

Nicos Maglaveras
Aristotle University, The Medical School, Lab of Medical Informatics - Box 323
54124 Thessaloniki, Greece
nicmag@med.auth.gr

Lydia Maigne
LPC-CNRS-In2p3, 24 av. des Landais, Campus des Cézeaux, 63177 Aubière cedex, France
maigne@clemont.in2p3.fr

Andigoni Malousi
Aristotle University, The Medical School, Lab of Medical Informatics - Box 323
54124 Thessaloniki, Greece
andigoni@med.auth.gr

Fernando Martín-Sánchez
Medical Bioinformatics Department, Institute of Health "Carlos III", Ctra. Majadahonda a Pozuelo, Km.2.- 28220 Majadahonda, Madrid, Spain
fmartin@isciii.es

Richard McClatchey
University of the West of England, Coldharbour Lane, Frenchay, Bristol, UK
BS16 1QY Bristol, UK
richard.mcclatchey@uwe.ac.uk

Enzo Medico
Dept. of Oncological Sciences, University of Torino, c/o Institute for Cancer Research and Treatment, s.p. 142, km 3,95 - 10060 Candiolo (TO), Italy
enzo.medico@ircc.it

Serge Miguet
Laboratoire LIRIS, Université Lumière Lyon 2, Campus Porte des Alpes
5, avenue Pierre Mendès France, 69676 Bron, France
serge.miguet@univ-lyon2.fr

Maria Mirto
ISUFI/CACT Center for Advanced Computational Technologies, Innovation Engineering Department, Engineering Faculty, Univ. of Lecce, Via per Monteroni, 73 100 Lecce, Italy
maria.mirto@unile.it

Johan Montagnat
CNRS (I3S laboratory), ESSI, 930 route des Colles, BP 145
06903 Sophia Antipolis Cedex, France
johan@i3s.unice.fr

Sofie Nørager[2]
European Commission -DG Information Society and Media, Office BU 31 6/41, B-1049 Brussels
sofie.norager@cec.eu.int

[2] All opinions expressed in the white paper are those of the authors and not necessarily those of the Commission.

Kazunori Nozaki
Osaka University, 1-8 yamadaoka, Suitashi, 565-0871 Osaka, Japan
kazunori@dent.osaka-u.ac.jp

Ilídio Castro Oliveira
University of Aveiro/IEETA, Campus Universitario de Santiago
3810-193 Aveiro, Portugal
ioliv@ieeta.pt

Xavier Pennec
INRIA Sophia - Projet Epidaure, 2004 Route des Lucioles BP 93
06902 Sophia Antipolis Cedex, France
xpennec@sophia.inria.fr

Yves Poullet
Law Faculty /University of Namur - 5, Rempart de la vierge, B.5000 Namur, Belgium
yves.poullet@fundp.ac.be

Juan Pedro Sánchez Merino
Medical Bioinformatics Department, Institute of Health "Carlos III"
Ctra. Majadahonda a Pozuelo, Km. 2., 28220 Majadahonda, Madrid, Spain
jpsanchez@isciii.es

Tony Solomonides
University of the West of England, Bristol, CEMS, Coldharbour Lane, Bristol BS6 6TH, UK
Tony.Solomonides@uwe.ac.uk

Irina.G. Strizh
Plant Physiology Department, Faculty of Biology, M.V. Lomonosov Moscow State University, Leninskie Gory, d.1, k.12 -119992 Moscow, Russia
irina.strizh@mail.ru

Michel Taillet
European Society for Therapeutic Radiology and Oncology (ESTRO)
av. E.Mounierlaan 83, 1200 Brussels, Belgium
michel.taillet@estro.be

Clive Tristram
ETS TRISTRAM Clive (Futur-Dessin), Les Rives, 86460 Availles Limouzine, France
clive.tristram@free.fr

Nikolay Tverdokhlebov
Institute of Chemical Physics, Kosygina, 4 - 119991 Moscow, Russia
nickhard@chph.ras.ru

Pierangelo Veltri
Magna Graecia University of Catanzaro, School of Bioinformatics and Biomedical Engineering, Campus di Germaneto, Viale Europa Germaneto, 88100 Catanzaro- Italy
veltri@unicz.it

René Ziegler
Novartis Pharma AG, WSJ-210.721, Lichtstrasse 35, 4056 Basel, Switzerland
rene.ziegler@pharma.novartis.com