Anthony Eugene Solomonides

Data, Metadata, and Workflow in Healthcare Informatics:

Mediating the triangular relationship between healthcare providers, researchers, and patients

a thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy (DPhil)
at the University of the West of England, Bristol

17th FEBRUARY 2017

Blank page

Dedicated to five former students who taught me much more than I taught them:

Tamás Hauer

Steve Jenkins

Mark Olive

Hanene Rahmouni

Kay Wilkinson

Acknowledgements

At UWE, I must acknowledge Professor Richard McClatchey, Dr. Rob Stephens and Professor Larry Bull: I am truly grateful for their friendship and support over many years. Thanks also to my many colleagues in CSM, CEMS, BIT, CSCT and especially CCCS—thanks for sharing nineteen years in the alphabet soup.

Outside UWE, I would especially like to express my gratitude to Dr. Jonathan Silverstein for his trust, friendship and support over the past five years and my sincere appreciation to Drs. Goutham Rao and Bernard Ewigman for their support and collaboration—three colleagues who created a highly conducive research environment at the NorthShore Research Institute.

Many thanks also go to my many colleagues in the American Medical Informatics Association and in the CAPriCORN network for numerous conversations and debates about these matters.

And thank you to Roseann, who has been encouraging me to get to this moment for longer than she cares to remember—relief at last!

Tony Solomonides, 17th February, 2017.

Blank page

Abstract

This dissertation considers a number of interlinked concepts, propositions and relations, and puts forward a set of design theses, to support *the role of informatics* in the overall goal of *knowledge-based*, *information-driven*, *integrated*, *patient-centred*, *collaborative* healthcare and research. This rather ambitious scope may be delimited by exclusion: the work is not concerned explicitly with *genomics* or *bioinformatics*, but it does encompass certain aspects of *translational medicine* and *personalized healthcare*, which I take to be subsumed in some sense under "knowledge-based" and "information-driven". Although I do not exclude *public health informatics*, my exposure extends only to surveillance of infectious diseases, patient engagement, and the effectiveness of screening programmes. I do take *ethical, legal, social and economic issues (ELSE)* to be included, at least to the extent that I aim at an infrastructure that encompasses these issues and aims to incorporate them in technical designs in an effort to meet ethicists', lawyers', policy makers', and economists' concerns halfway. To a first approximation, the aim has been to integrate two strands of work over the last decade or more: the *informatics of medical records* on one hand and the *distributed computational infrastructures for healthcare and biomedical research* on the other.

The papers assembled in this dissertation span a period of rapid growth in biomedical informatics (BMI[i]). Their unifying theme was not declared programmatically at the beginning of this period, but rather developed, along with individual pieces of work, as my engagement – and that of my students – with BMI became more focused and penetrated deeper into the issues. Nevertheless, I believe I have learned something from each project I have been involved in and have brought this cumulative experience to bear on the central theme of my present work. My thematic vision is of a scientifically literate and engaged community whose members – citizens, patients, caregivers, advocates – are sufficiently interested in medical progress and in their own health to take ownership of their medical records, to subscribe to a research service that informs them about progress and about current studies that may interest them, and so take responsibility for their own and the health of those close to them. This entails many things: agreements on what constitutes legitimate data sharing and when such sharing may be permitted or *required* by the patient as owner of the data. It calls for a means of recognizing the intellectual contribution, and in some healthcare economies, the economic interest of a physician who generates that record. Ethically, it requires a consenting policy that allows patients to control who may approach them for participation in a study, whether as a subject, as a co-investigator, as a patient advocate, or as a lay advisor. Educationally, it requires willingness on the part of physician-researchers and scientists to disseminate what they have discovered and what they have learned in terms that are comprehensible to the interested lay participant—but do not speak down to her.

[i] Notwithstanding its ambiguity, this acronym has gained wide acceptance in US technical literature.

Blank page

Data, Metadata, and Workflow in Healthcare Informatics

Mediating the triangular relationship between healthcare providers, researchers, and patients

Contents

Chapter 0

Publications—
The Selection of Papers

Blank page

0 The selection of papers

This is a collection of loosely interrelated papers published over a period of ten years or so, submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy by publication (DPhil). This section provides a listing of the **main papers** included with a brief rationale for its inclusion. Some related additional **subsidiary papers** are also listed in the bibliography but are not included in the submission in the interests of conciseness. The main body of the dissertation offers what may reasonably be described as a rational reconstruction of the intellectual process that led to the particular sequence of papers.

Chapters and Corresponding Papers	**Comments / Author's Contribution**
1—Overview	
Main papers form the substance of the submission; subsidiary papers are discussed relatively briefly and are used to flesh out aspects of the work or additional contributions that are related to the main claim.	This overview is based on two peer reviewed workshop contributions, the first to a panel on *The Many Meanings of Precision Medicine* at the AMIA Joint Summits in Translational Science 2016, and the second to the fourth Middlesex University Workshop on *ICT in Healthcare – Legal, Ethical and Social Challenges* also in March 2016.
2—Phase I – MammoGrid and Healthgrid Projects	
A. F Estrella, C del Frate, T Hauer, R McClatchey, M Odeh, D Rogulin, S R Amendolia, D Schottlander, **T Solomonides**, R Warren. *Resolving Clinicians' Queries Across a Grids Infrastructure* **Methods of Information in Medicine** Vol **44** No 2. 2005 pp 149-153. ISSN 0026-1270 Schattauer publishers.	In the first phase of MammoGrid, the interpretation of clinicians' and clinical researchers' requirements and translation of the languages of doctors and technologists to each other was central to my role in the project. ***Contribution*** I contributed sections corresponding to the data model and to user modelling and reviewed the paper as a whole. (Authors are listed alphabetically.)
B. R Warren, **T Solomonides**, C del Frate, I Warsi, J Ding, M Odeh, R McClatchey, C Tromans, M Brady, R Highnam, M. Cordell, F Estrella & R Amendolia. *MammoGrid – A Prototype Distributed Mammographic Database for Europe* **Clinical Radiology** Vol **62** No 11 pp 1044-1051. DOI 10.1016/j.crad.2006.09.032 November 2007, Elsevier publishers.	***Contribution*** I provided the first complete draft of this paper *ab initio*. Professor Brady read and improved the description of the Standard Mammogram Form. Dr. Warren brought the language into line with standard radiological usage. I reviewed all changes before publication. (Note that the lead author was required by the journal to be a clinician, even in a technical paper.)

Chapters and Corresponding Papers	Comments / Author's Contribution
2—Phase I – MammoGrid and Healthgrid Projects (continued)	
C. Estrella F, Hauer T, McClatchey R, Odeh M, Rogulin D, **Solomonides T**. *Experiences of engineering Grid-based medical software*. **Int J Med Inform.** 2007 Aug; **76(8)**:621-32. Epub 2006 Jun 19.	This retrospective paper reflects on (a) the applicability of software engineering techniques in the specification and implementation of a healthgrid project (MammoGrid) and shows that use-case modelling is a suitable vehicle for representing medical requirements and for communicating effectively with the clinical community; and on (b) the practical advantages and limitations of applying the Grid to real-life clinical applications and presents the con- sequent lessons learned, especially in terms of demands on the level of commitment needed from collaborating radiologists and the degree of standardization and stability of the underlying software. ***Contribution*** Seeking a convenient means for communication between clinician researchers and software engineers, I led the adoption of UML use case diagrams as a structured means of representing interactions between radiologists and the MammoGrid infrastructure. This led to my principal contribution to this paper in the precise specification of user requirements and the evaluation of the extent to which they were met.
D. Olive, M., Lashwood, A. and **Solomonides, T**. (2011). *A retrospective study of paediatric health and development following pre-implantation genetic diagnosis and screening*. In: Olive, M. and Solomonides, T. (ed.) **IEEE Proceedings of the 2011 24th International Symposium on Computer-Based Medical Systems** (CBMS 2011), pp. 32-38.	I led the EuroPGDcode project on behalf of the European Society for Human Reproduction and Embryology (ESHRE). Although funded by the Executive Agency for Health and Consumers, so not directly a "healthgrid" application, the purpose of the project was to demonstrate the possibility of codifying and automating the collection of Preimplantation Genetic Diagnosis (PGD) data across Europe in such a way as to support research. Although the project did not succeed in unifying the process across Europe, it led to a highly productive collaboration with the relevant British researchers. ***Contribution*** This paper was jointly written by the three authors, each of whom made their contribution from a different point of view. I conceived the project in this form, bringing together my PGD work with that of Mark Olive's doctoral research and Alison Lashwood's clinical expertise.

Chapters and Corresponding Papers	Comments / Author's Contribution
3—Phase II – HealthGrid White Paper and SHARE Road Map	
E. V. Breton, K. Dean, **T. Solomonides**, *The HealthGrid White Paper*, in *From Grid to Healthgrid*. **Studies in Health Technology and Informatics**, vol. 112, ISBN 1-58603-510-X, ISSN 0926-9630 IOS Press.	Following a number of successful and well-received European projects exploiting health grids, the HealthGrid Association was formed and incorporated in France by the leading investigators in the area. Vincent Breton at CNRS, France, Kevin Dean at Cisco, UK, and Tony Solomonides, UWE, Bristol, were invited by HealthGrid to solicit contributions, including their own, and to edit and publish a peer reviewed white paper that would describe both the early achievements and the potential of grid technologies in healthcare. ***Contribution*** Whilst the majority of grid research was preoccupied with data grids (rapid storage of large volumes) and computational grids (virtual parallel machines), I particularly identified and discussed the potential of healthgrids to support collaboration in the spirit of the e-Science programme in the UK.
F. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez *SHARE road map for HealthGrids: Methodology* **International Journal of Medical Informatics**, 78S (2009) S3–S12	After the HealthGrid White Paper, HealthGrid was granted an EU FP6 project, SHARE, with the explicit brief *to establish healthgrids a the infrastructure of choice for biomedical research, and subsequently for healthcare, in Europe*. The envisaged system of healthgrids would be able to serve as a web-like backbone for the sharing of research objects (data and metadata, workflows, collaboration, results, analyses, etc.) and when proved mature and secure through research, to be further deployed in the delivery of healthcare. ***Contribution*** This paper presents a methodological review of the challenges and opportunities facing the SHARE collaboration. The paper provides an account of the multi-phase process through which the ultimate road map was developed. Vincent Breton devised the earliest deployment plan while I supplied the methodological framework; Blanquer and Hernandez provided analysis through particular use cases.

Chapters and Corresponding Papers	Comments / Author's Contribution
3—Phase II – HealthGrid White Paper and SHARE Road Map (continued)	
G. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez. *SHARE, from Vision to Road Map: Technical Steps.* **Studies in Health Technology and Informatics** Volume 129: ***Building Sustainable Health Systems*** - Proceedings of the 12th World Congress on Health Informatics – MedInfo 2007. IOS Press 2007	MedInfo 2007 provided a unique opportunity to address a very broad international conference and to expose the "HealthGrid" philosophy in its historical context, as it transitioned from the white paper to the SHARE Roadmap 1. ***Contribution*** I was the lead author and presenter. Once again, in this paper authors are listed by institution, with the most senior at each listed last in its group.
H. Mark Olive, Hanene Boussi Rahmouni and **Tony Solomonides** (UWE, Bristol, UK)[ii] Vincent Breton, Nicolas Jacq and Yannick Legré (IN2P3, CNRS, Clermont-Ferrand, France & HealthGrid, EU) Ignacio Blanquer and Vicente Hernandez (Universidad Politécnica de Valencia, Spain) Isabelle Andoulsi and Jean Herveg (Universitaires Notre-Dame de la Paix, Belgium) Celine Van Doosselaere and Petra Wilson (European Health Management Association, EU) Alexander Dobrev, Karl Stroetmann and Veli Stroetmann (Empirica GmbH, Germany) *SHARE: A European Healthgrid Roadmap* in **Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare** (Mario Cannataro, Ed). Chapter 1, pp. 1–27. IGI-Global, Hershey, PA. 2009.	This publication is a distillation of the full roadmap and other final reports (on technology, on ethical and legal issues, on case studies, etc.) A preliminary version of this was the subject of a two-day workshop review by more than twenty invited experts. Once approved and accepted by the EU, it was published as a glossy report by the European Commission under the title *SHARE the journey: A European Healthgrid Roadmap*. Subsequently, the paper underwent further peer review and finally appeared in Cannataro's handbook. ***Contribution*** This paper was primarily authored by Tony Solomonides with assistance from two graduate students, Rahmouni and Olive. The authors are listed by institution with the senior author from each institution listed last. From the abstract in the Handbook: *The principal goal of this chapter is to elucidate the future requirements of healthgrids if they are to become the infrastructure of choice for biomedical research and healthcare. These requirements take many forms, technical, organizational and economic, with initiatives required in the domains of ethical and legal regulation. Thus, particular objectives of the chapter are to explore and analyse each of these domains to a sufficient depth to be able to make sense of the overall picture.*

[ii] *A note on authorship of SHARE and HealthGrid papers* Some of these multi-institutional and multi-author papers were derived from longer reports to the funding body, principally the EU Framework programmes. The lead institution in each publication has its authors listed first, with the senior author last among them. Thus, Solomonides appears last on the UWE list, following Rahmouni and Olive.

Chapters and Corresponding Papers	Comments / Author's Contribution
3—Phase II – HealthGrid White Paper and SHARE Road Map (continued)	
I. **A E Solomonides**. *Compliance and Creativity in Grid Computing.* 16th World Congress on Medical Law, Bioethics Track. Toulouse 2006.	***Contribution*** Sole authorship. The ideas in this paper are a reflection of the author's views and thought alone.
4—Phase III – The Learning Health System	
J. **Anthony Solomonides**, Satyender Goel, Denise Hynes, Jonathan C. Silverstein, Bala Hota, William Trick, Francisco Angulo, Ron Price, Eugene Sadhu, Susan Zelisko, James Fischer, Brian Furner, Andrew Hamilton, Jasmin Phua, Wendy Brown, Samuel F. Hohmann, David Meltzer, Elizabeth Tarlov, Frances M. Weaver, Helen Zhang, Thomas Concannon, Abel Kho. *Patient-Centered Outcomes Research in Practice: The CAPriCORN Infrastructure.* **Studies in Health Technology and Informatics** Volume 216: *MEDINFO 2015: eHealth-enabled Health.* pp 584 – 588. IOS Press 2015. DOI 10.3233/978-1-61499-564-7-584	From the abstract to the paper: *To capture complete medical records without compromising patient privacy and confidentiality, the network created policies and mechanisms for patient consultation, central IRB approval, de-identification, de-duplication, and integration of patient data by study cohort, randomization and sampling, re- identification for consent by providers and patients, and communication with patients to elicit patient-reported outcomes through validated instruments. The paper describes these policies and mechanisms and discusses two case studies to prove the feasibility and effectiveness of the network.* ***Contribution*** This paper was written by Tony Solomonides to define how the hashing approach to the de-identification of patient data would work in practice.
K. **Anthony Solomonides**. *The Learning Patient in the Learning Health System.* (**MedInfo 2017**, submitted)	This submission is a short version of a longer paper under development on the role of "expert" patients in a learning health system. ***Contribution*** Sole authorship. The ideas in this paper are a reflection of the author's views and thought alone.

Blank page

Chapter 1
Overview—
A unifying theme

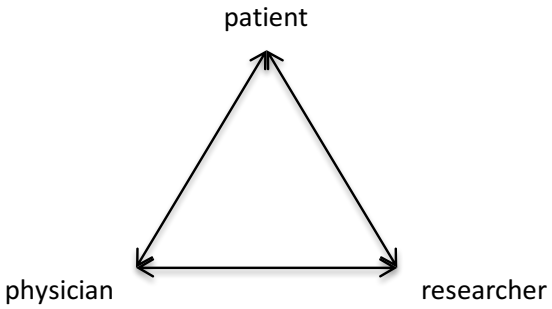
Blank page

1 Overview

The papers assembled in this dissertation span a decade and a half of work in biomedical informatics (BMI). Their unifying theme was not declared programmatically at the beginning of this period, but rather developed, along with individual pieces of work, as my engagement – and that of my students – with BMI became more focused and penetrated deeper into the issues. Nevertheless, I believe I have learned something from each project I have been involved in and have brought this cumulative experience to bear on the central theme of my present work. It may be helpful to begin from this before recapitulating the experience that led there. My thematic vision is of a scientifically literate and engaged community whose members – citizens, patients, caregivers, advocates – are sufficiently interested in medical progress and in their own health to take ownership of their medical records, to subscribe to a research service that informs them about progress and about current studies that may interest them, and so take responsibility for their own and the health of those close to them. This entails many things: agreements on what constitutes legitimate data sharing and when such sharing may be permitted or *required* by the patient as owner of the data. It calls for a means of recognizing the intellectual contribution, and in some healthcare economies, the economic interest of a physician who generates that record. Ethically, it requires a consenting policy that allows patients to control who may approach them for participation in a study, be it as a subject, as a co-investigator, as a patient advocate or as a lay advisor. Educationally, it requires willingness on the part of physician-researchers and scientists to disseminate what they have discovered and what they have learned in terms that are comprehensible to the interested lay participant but do not speak down to her.

1.1 Reciprocity

Central to this view is a triangle of reciprocal relationships between patients and their caregivers, physicians and other providers, and biomedical researchers and other scientists.



At each of the three vertices is an archetype. Each is the focal representative of a category of roles or actors. A “patient” stands also for individuals who are well and wish to preserve their health; for parents and caregivers; for patient advocates and other support groups. A “physician” is at the apex of a phalanx of fellow professionals, including pathologists, radiologists, nurses, technicians, dieticians, social workers, psychologists, and so on. A “researcher” may be a wet-lab scientist, a pharmacologist, a bioinformatician, a statistician – the possibilities are even more numerous. Behind each of these archetypes is a source of funds: an employer (or savings) for the patient; an insurer or other payer for the physician; a funding agency or a pharmaceutical

company for the researcher. Associated with each vertex is a characteristic cost: the cost of being ill or of looking after someone who is ill; the cost of running a medical office or hospital system; the cost of providing medical or nursing care; the cost of a research lab. Overlaying the entire scheme, the inevitable frictional costs of a market-based healthcare economy and the cost of public health. These constitute the healthcare economy.

Each of the three archetypes at the vertices both depends on the other two and provides something essential to them. The interdependence of physician and patient is perhaps obvious. Their relationship can be parsed in each direction: the physician provides care for the patient and adds to her experience as she delivers care. The physician takes responsibility for the patient's wellbeing and the patient repays the physician with trust and loyalty, helping to maintain the stability of her practice. Historically, it has been said that physicians used to do things *to* patients (the object model of the patient), then moved on to do things *for* patients (the consumer model of the patient), and now finally are coming to do things *with* patients (the collaborative model of healthcare). This development is mirrored in certain demographic segments where there is demand for a more active engagement in health maintenance (witness the growth of exercise and yoga cultures, and the "quantified self" movement) and in information seeking on the Internet to support or supplement, and even to question, medical authority.

The relationship between researcher and physician may be read as "translational" – the problems of physicians are at the heart of projects that researchers tackle; the knowledge that researchers establish is translated into medical or operational improvements in care. The time scale over which this relationship manifests itself is longer and the very relationship itself is less readily identified. Healthcare providers' typical focus is on providing care for patients as efficiently and effectively as possible, with just enough attention to maintenance of records, especially electronic records, to ensure continuity of care. These records often require considerable treatment before they can be used for research; for example, if they are in free text, either chart review by another expert or reliable natural language processing would have to be performed to extract discrete data that can be mined or correlated with other outcomes data towards discovery. Discrete data entry renders data more useable for research, but providers often find discrete entry systems, with their succession of menus, limited choices and cascades of screens, both more time consuming and more restrictive than free text. Thus, unless a physician has some investment in a research project, the value of their work to the researcher is at best highly mediated and at worst of no use at all. Conversely, the results of research, published often in recondite articles in a highly diverse specialist literature, cannot be translated immediately into care decisions. Typically, it filters through to physicians in "journal club"[iii], or decision support aids (cf. [3], [4]), or commissioned articles in professional (as opposed to learned) journals and newsletters.

The third relationship, that between patient and researcher, is less sharply defined, not least because it has traditionally been mediated by a healthcare provider and also because it is rapidly evolving in the face of larger changes. As the principles of evidence-based medicine have been widely adopted and translated into practice, the financial cost and the slow nature of traditional methods of knowledge creation, notably clinical trials, have underscored a need for effective

[iii] Journal Clubs are typically institutionally or departmentally organized; they are sometimes supported by journals or journal sections (see [1], [2]).

alternatives. Evidence creation – perhaps *discovery* would be a better term – now relies increasingly on comparative effectiveness research (CER) based on observational data from electronic health records (EHR) generated by physicians and other providers in the course of the delivery of care. Much of this can be done through analysis of de-identified data, bypassing the need for consent by appeal to an institutional review board (IRB) for an exemption. A complementary trend has seen increasing activism on the part of patients and patient advocacy groups, both to assert the need for more emphasis on patient-centred outcomes research (PCOR) and a willingness to engage in the formulation of research questions, programmes and proposals. Consequently, a need has arisen for consultative structures that allow patient communities (broadly conceived, as above) to engage, propose and approve research projects, fulfilling in an indirect way the *informative* requirement of the consent process without necessarily reverting to the—sometimes prohibitively difficult—old processes of obtaining consent. The entire enterprise of the Patient-Centered Outcomes Research Institute (PCORI) in the United States is dedicated to promoting, funding and disseminating this approach [5].

1.2 Obstacles to the realization of reciprocal relationships

I have posited the "triangle of reciprocity" above as an aspirational goal; each pairwise relationship provides opportunities, but is equally fraught with challenges. In this section, I shall attempt to navigate these more or less in the order physician—patient—researcher—physician, but it will be clear that they are mutually implicated and impinge on each other, so that it is necessary to keep the triangle in view throughout.

1.2.1 Relationship and Ownership

I have already noted the transition from "doing things to patients", through "doing things for patients", to "doing things with patients", respectively viewing the patient as the passive recipient of treatment, as the active customer, and now increasingly as the principal stakeholder and quasi-expert in her or his own health. In the latter, current view, predicated to some extent on limited resources, patients have significant responsibilities — as well as rights — in the maintenance of their own health. Arising out of these responsibilities is the patient's right to know what the medical records say about him or her. In this dissertation, one thread discusses how these concerns may be addressed from a technological point of view.

An inevitable issue in healthcare informatics is the question of ownership: Who is the owner and who should have custody of a patient's medical records? This has been inherited from the era of paper records, when the tension was more between providers and payers than between patients and physicians. For example, in the NHS, ownership is now explicitly attributed to the custodian organization in [6] [7]; in an extreme case a dispute between the relevant government department and a practice that has lost records through flooding may hinge on the distinction between ownership of the physical medium on which patient data were recorded and the data itself[iv]. In the US, the project *Health Information and the Law* maintains an online map [8] of the United States with links to state legislation concerning ownership of patients' records.

[iv] This was reported anecdotally in November 2004 at a London meeting of the Department of Trade and Industry with the EU Commission's Head of e-Health, Dr. Gerard Comyn, at which discussion turned to the question of ownership of health records. For the event itself, see: https://www.digitalhealth.net/2004/11/new-eu-e-health-funding-will-focus-on-integration/

In the era of EHRs, the questions multiply: Who may access those records in the course of healthcare delivery to the patient, under what circumstances and how? When, if at all, may those records be accessed for other purposes, such as public health, quality improvement, and research? When is the data subject's consent necessary for such "secondary" use of patient data? Granted patient consent, what are the privacy and confidentiality implications of any sharing or secondary use of personal medical records? This issue has been further politicized in the United States as the Trump administration has moved quickly to suspend all regulatory actions of the Obama administration, including the updated "Common Rule" [9].

Discussion of this issue in depth requires a monograph in itself. While the patient's record primarily holds (more accurately: represents) information concerning the patient's health status, it also incorporates some of the physician's intellectual work, and it includes billing data that may legitimately be claimed for the payer. If the patient links data from a health-related social or quasi-social site, such as HealthHeritage [10] or WiserCare [11], or from a wearable device through the manufacturer's linked web services, the picture becomes even more confused. Does advice or a risk score from one of these sites belong to the patient, the physician, or the originator? It is clear that to take proper account of this, a highly ramified data structure would be necessary, and one that would only be obtainable if it can be recorded automatically. No one, not the physician, not the patient, nor any administration could otherwise justify the investment in time.

It has been shown that the problem is tractable, if still somewhat expensive to implement, in the case of privacy constraints based on data provenance and a formal understanding of the regulatory framework. A series of joint papers [12] authored by my student, Hanene Rahmouni, addressed some formal aspects of these issues in a particularly elegant manner, by representing the legal framework in a declarative logic and translating them into actionable deontic logic at the operational level.

1.2.2 Identification and Consent

In many, possibly most, cases, the researcher requires access to the patient primarily to test a therapy or other intervention. Less frequently, access is needed to survey the patient about a recent illness or procedure to determine, respectively, its sequelae or effectiveness. An increasing volume of research, however, involves—at least initially—only observational data recorded in the process of health care provision. A significant research industry has built up around this activity. Identifying the right patients to study requires accurate "phenotyping", i.e. the specification of a set of criteria in the medical record that identify precisely those patients of interest. The eMerge network's Phenotype Knowledgebase [13] holds a collection of rigorously tested phenotype algorithms. More widely, the large number of PCORnet Phase I demonstration projects have generated interest and awareness of the phenotyping problem in the research community. Once the basic population of interest has been identified, a number of alternatives for research are available: a random sample may be drawn and matched controls identified by means of another algorithm applied to the same EHR. Or matched samples from different health systems where different approaches are employed may be compared prospectively for effectiveness. A major advance in the PCORnet approach to research is the emphasis on patient engagement and participation in research project formulation. This leads to a fresh set of requirements which have refocused the question, not so much on ownership, as in the navigation

of permissions to access. This is not a new area of research; for example, the Manchester group's FARSITE architecture has addressed this issue in the British context. [14] However, the need in the case of PCOR is broader: the scheme for such research involves the patient (and caregiver, advocate, etc.) as an active contributor in the design of the research project, from proposal, through question formulation, determination of primary and secondary goals, hypothesis formation, target population and recruitment process, to the collection, analysis and interpretation of results. There is an implied transition from subjecthood through what may be described as "co-design" to full "co-production". Ainsworth and Buchan (and co-discussants) [15] have very recently extended and deepened the argument for "combining" health data uses towards *health system learning* (i.e. the necessary prerequisite for a *learning health system*) in a way that complements the argument made here for the *learning stakeholder* in the learning health system. I remain agnostic as to the right verb for this convergence: "combined" conveys the right sense of *economy*—non-redundancy—but I also want to reflect the diversity of viewpoints. I freely confess that a paragraph here will not suffice to do justice to their work; a natural next step in my research would be to seek to marry their insights with the concept of the learning stakeholder.

Consent to participate in research is a deceptively complex concept. The strict legal requirement in most settings is *informed* consent for a *specific* study. The first qualifier, "informed", entails an explanation to the potential research subject, in relatively plain terms, of what the study entails, what its goals are, and what risks it may impose, as well as to assert the freedom to withdraw at any time. On the specificity restriction, an extension of the study, or even a variation of the protocol, requires the researcher to return to the patient for further consent, and that may yet need to be approved, and so mediated, by a healthcare provider. Once consented and enrolled, the patient—study-subject—has only one principal sanction available, to withdraw from a study. It is possible to envisage a different form of mediated (*media*ted) relationship in which the patient consents to be involved in research, to offer to participate in studies and to be kept informed of progress, especially when publications are available. The practice, adopted already by some journals, of publishing a "patient's summary" of research findings would make this even more potent as a means of engaging the patient fully. There are several issues to be addressed, including the immature researcher's tendency to aim for immodest goals and the problem of research subjects over-identifying with the researchers' desire to see their project succeed. I am currently working with a member of the African American community on the "South Side" of Chicago and a colleague at the University of Denver to formulate an education programme termed "Boot Camp Translation" in research methods for patients. [16, 17]. The concept of taking the patient-subject into the researchers' confidence and allowing her to make a meaningful intervention in a research programme may therefore best be broken into phases and forums where, through formal roles and formal settings, roles may be differentiated and unbiased engagement be made possible. A clear prerequisite for this is education that explains and justifies the means of goal setting, the "methodology"—a difficult notion about which experts disagree as much as any lay discussants—by which results will be obtained, and the interpretation of results into action. In the discussion of the goals and aims of PCORI [5], clear criticism has been voiced of research that leads to non-actionable results and even outcomes that may be good from a population health point of view but of less value to the individual patient. This is not to agree with this view, but to highlight the need for the broader scene-setting education for those who wish to impact research proposals.

1.2.3 Mediation, Neutrality and Common Ground

Human interaction and education are necessary, but not sufficient. There is a need for the relationship to be smoothly mediated between patient, physician and researcher. If, perhaps when, patient-*managed* electronic records are more widely adopted, it may be possible to transcend the questions of custody, ownership and control. The issues already discussed in terms of ownership are echoed here in a different form: must the physician share every hunch, every concern, with the patient? Conversely, must the physician write nothing in the record that may offend the patient? Will the patient's right to correct his record mean removal of anything he does not like—e.g. "morbid obesity" in his problem list? (cf. the "fat acceptance movement" [18])[v]. Imaginative solutions are no doubt possible for many such problems: e.g. non-prejudicial private notes by the physician to herself may remain private, but may none the less be subject to scrutiny by the quality assurance process in the institution. The patient may object to unwarranted entries, but "morbid obesity" is a technically defined term; he has the choice to move to another doctor, if he can find one who would not consider his BMI to be a problem. Experiments in sharing information with patients have so far proved promising. For example, six years since its inception, the Open Notes initiative has made health records available to over 12 million patients at 46 medical centres against considerable initial scepticism from the medical community [19, 20].

Notwithstanding, many problems remain. The banking system is occasionally used as an analogue of what may be implemented in an Electronic Health Bank. In October 1997, Dr. Bill Dodd, a Scottish GP, gave an interview to the British Journal of Healthcare Computing and Information Management in which he set out his proposal for a health banking system in three organizations: a Health Information Bank and a Health Information Academy, both non-profit, and a "commercially oriented Health Information Corporation". [21] The proposal envisaged competition, so that the patient would have a choice of banks, much as he does in the financial sector. The idea lay dormant in the UK but had been apparently independently conceived and had begun to be developed in parallel by Marion Ball and others at IBM in the US [22]. Denis Protti, of the University of Victoria, Canada, invited to advise the ill-fated English National Programme for IT (NPfIT), revived Dr. Dodd's proposal, first in an internal publication for Connecting for Health (CFH) and subsequently, in 2008, in the Canadian journal Electronic Healthcare [23]. It is not clear whether this idea transmuted into the subsequent proposal for Care.Data—which would have supported a different kind of data banking—but Professor Protti's enthusiasm for Dr. Dodd's idea is clear.

1.2.4 The Research Dimension

The idea of a health bank has since been taken up more widely, with several prominent academics championing the cause in the present decade. Prominent among these are Dr. Amnon Shabo (Shvo) at Haifa who made this a central plank of his keynote address at MedInfo 2015 as well as in a series of high profile publications [24]. Dr. Patricia Flatley Brennan, also used the

---

[v] I have some personal experience of this through my engagement with the stakeholder group for the PCORI study *Short and Long Term Effects of Antibiotics on Childhood Growth*. Parents have asserted a strong preference for phrases such as "a child has obesity", i.e. suffers from a medical condition, rather than "a child is obese". The contrast with the Fat Acceptance movement is acknowledged, but getting the language right is felt to have priority.

platform of a keynote at MedInfo 2015 to extend her eloquent argument [25, 26] for personal health records (PHRs) into a broad vision. In these presentations, as in the most recent papers, the question of research is broached, albeit somewhat obliquely. [27–31]. First, research is called for on how PHRs are used; this is a step towards evidence-based policy and would indeed provide some essential background knowledge to optimize the use of such records. The attribution of responsibility and costs goes hand in hand with the assertion of rights: how—and what part of—the physician's record of an encounter become part of the PHR? How much of the patient's PHR must be revealed to a provider? I have already discussed issues of ownership and custody, intellectual property and ethical disclosure. There is potential for the PHR to be used by provider organizations to market additional services to patients when it is clear that they are receiving such services elsewhere. Where must the line be drawn—a line that may in any case be deeply embedded in a "black box" technology?

The generation of knowledge from medical records created in the course of healthcare delivery has been advocated by researchers for some time, but took a definite form and gained impetus from the Institute of Medicine's (as the National Academy of Medicine then was) embrace of the concept of the Learning Health System (LHS) [32]. In a series of publications, different aspects of such a system were analysed and debated, ultimately culminating in the formation of a non-profit organization, the Learning Health Community.

1.2.5 Study Designs: from Pragmatic Trials to Observational Studies

The LHS approach to knowledge creation contrasts sharply with the traditional understanding of (randomized control) clinical trials. Yet as far back as 1967, Schwartz and Lelouch introduced a distinction between explanatory and pragmatic trials to deal with the "real world" dimensions that must be taken into consideration in making comparisons:

> *... Suppose, for example, we require to compare two analgesics and assume first that the two are chemically very alike, differing only in a single radical. The biologist may then be interested to know whether the drugs differ in their effects when they are administered on an equimolecular basis. This is the explanatory approach.*
>
> *On the other hand, assume that the two substances are chemically quite unrelated. Each will presumably have an optimal level of administration, having regard to its side- effects, and the problem of interest is now to compare the two drugs administered at these optimal levels. This is the pragmatic approach.* [33]

A discussion of causality *per se* is beyond the scope of this dissertation, but it is noted here that a plausible explanatory process (e.g. metabolic) is a requirement in assessing whether correlation is indicative of causation. The article [33] was reprinted in 2009 and given fresh impetus to work to support appropriate clinical trial design [30], including the creation of a tool PRECIS-2 [34] to help trial designers in their task. number of limitations: recruitment may be constrained by ethical permissions, so that the subjects may be less sick than the real population to be treated; recruitment may also be limited by the location of experienced investigators willing to recruit, resulting in small sample sizes and in lack of adequate diversity. Notwithstanding, they are also rather expensive to conduct. Observational studies relinquish some elements of control in the interest of addressing some of these issues: data may be collected from a much larger population "in the wild"—with all issues of compliance and monitoring that raises—yet with representative diversity and distribution of health status.

Observational studies may also be the only method possible when randomization is either impossible or would be unethical. Examples include surgery [35] and the current PCORnet Obesity Observational Study: Short- and Long-term Effects of Antibiotics on Childhood Growth [36] in which I am currently involved. The challenge in this study, as in many similar ones, lies in adequate control for confounders in the population to be studied. The population of interest is all children in the data warehouses of participating institutions (no cluster randomization is implied here) and the endpoints are weight at 5 and 10 years old, with data concerning antibiotic exposure in the first two years of life, reason for this, comorbidities (e.g. asthma with possible use of corticosteroids), and available demographics and social indicators.

1.2.6 Patient and Provider Engagement – Co-Design and Co-Production

A recent discussion paper from the National Academy of Medicine [27] discusses studies such as this in terms of the knowledge to be derived from "best care". Although a major focus of the article is PCORI activity, there is little emphasis on patient engagement; rather progress is anticipated from top-down pressure from executive suites:

> *As described above, within the current health context, the two activities of clinical operations and research operate in largely separate environments with different (and at times competing) players, funding streams, incentives, and priorities. The authors believe that research can move more quickly once research interests are aligned with operations. Likewise, operations will be more evidence based and thus the quality of care improved once operations stakeholders are engaged in the development of research priorities and their needs and strategies are reflected in the research agenda.*
>
> *To build these relationships requires that health executives promote the benefits of integration, including ideas related to seamless integration of research and practice, and that they create structures, funds flow, and processes, and allocate time and resources, to those collaborations.*

This is, perhaps, a pragmatic response to the issues that PCORI now confronts. How to persuade senior executives that its projects—and more importantly, its defining mission—are worth preserving? How will PCORnet "keep the lights on" after PCORI's "sunset" in 2019? Concerns about what kinds of projects the Institute should be engaged in were expressed early on. [37]

The idea of patient engagement in research was still novel when PCORI was established. One of its creative moves in this direction has been the funding of Patient-Powered Research Networks (PPRNs) alongside the Clinical Data Research Networks (CDRNs), which were cast in a more traditional mould and resembled the academic collaborations set up under the Clinical and Translational Science Awards (CTSA) programme. Returning to the issue of health banking, it is now evident that neither major players in the social media space (Microsoft, Google) nor governments are yet trusted by patients with their data, witness the failure of HealthVault, GoogleHealth and, in the UK, Care.Data. Patient-run organizations may finally prove to be the catalyst that makes this happen. Apart from faith-based organizations (in the US, at least), patient-run organizations appear more likely to be trusted than any current alternative. Health banks built around a particular condition common to their members would have the additional advantage of singular focus, the ability to tap into relevant research, and to keep their members engaged. However, they are unlikely to scale up to a viable size and would be vulnerable to the fatal attraction of financial support from commercial entities.

1.3 The role of informatics

Having reviewed the obstacles, I turn to the potential of *informatics* in its broadest sense, encompassing both the abstract qualities, contexts and nuances of information and the technologies that can be used to manage it—from acquisition through representation, transformation, analysis, précis and dissemination, as well as from protection, through secure storage, encryption, censorship, to aggregation and obfuscation as means of preserving privacy and confidentiality. Informatics is fraught with the same questions, or mirrors of them, that we have already encountered, but as a young and optimistic discipline appears to offer more hope of solutions. The counter-danger lies in the tendency to settle on a "technical fix", i.e. a purported solution that appears to solve the problem but does not address the underlying issue.

1.3.1 Electronic Patient Records – Provision of Care[vi]

Medical records have long been an issue of concern: historically, in virtually all national health systems, they have been held in independently maintained, unconnected filing systems, both paper and electronic ("silos"). As the priorities of healthcare practice have changed, electronic patient records (EPR) have been increasingly adopted, giving rise to a number of questions:

Should a health system maintain a single integrated record for each patient? If the claim that modern medicine is holistic is sustained, an assertion of the "obvious" value of an integrated record appears to follow immediately. However, this is readily countered by the observation that each healthcare professional needs to know only that part of the EPR that is relevant to their specialty (a restriction termed *patient confidentiality*) as well as by fears about unauthorized access to the entire record by others (posing a threat to *patient privacy* —even to the extent of *identity theft*— and, in some systems, compromising their ability to get health insurance coverage). Notwithstanding these valid concerns, there is considerable evidence that integrated patient records support better healthcare, so that the effort necessary to address them is amply justified. It will be argued that appropriate annotation of the EPR can address these challenges when coupled with adequate security of access to the systems [12].

May computed data be included in the EPR alongside observed data? Electronic records make it possible not only to record symptoms, signs and observations, but also to analyse patterns that may be suggestive of other possibilities. For example, a series of readings of high blood pressure, taken independently at different locations and on different occasions, may suggest that the patient should be screened for hypertension and an appropriate alert be entered into the system. This may be controversial for a number of reasons and may pit the physician's professional responsibilities against the patient's values concerning his or her own health.

A further paradox arises in conditions, such as cancer, where a suitably curated patient registry is mandatory, considerable information about a patient's status (e.g. staging information) may not be imported back into the record unless a physician reviews it anew, since it would become

[vi] There are plausible distinctions between electronic medical records, electronic personal records and electronic health records. Most of the time these are ignored and the terms are used interchangeably. I differentiate EPR and EHR in this section to emphasize the focus – the person, or the panel of patients.

actionable the moment it is (re-)instituted in the record. This information therefore remains siloed, despite the best intentions of the institution.

What kind of access to her or his EPR should a patient have? Even today, in most settings this is a controversial question, but by and large governments have mandated, or at least adopted measures to encourage, the sharing of the EPR with the patient. In such settings, even if initially access to the record is "read only", it will be necessary to allow patients to comment on and, indeed, sometimes to correct their record. Moreover, technically, there are many occasions and ways in which the patient's record may be shared with the patient: online as a means of review and reflection, concurrently with the physician perhaps on a second screen during a consultation visit, as a means of health and wellbeing maintenance when using "apps" to interface monitoring or fitness devices such as blood glucose meters, pedometers, or gym equipment.

The principle of "shared decision making"—doing things with the patient—also sometimes paints the electronic record as a jointly maintained chronicle of the patient's health status, but poor design of interfaces has limited ways in which patients and providers can work together around the record. It is not difficult to imagine a dual screen, so that the patient and physician are facing each other and looking at some fragment of the record that is relevant to the current interaction.

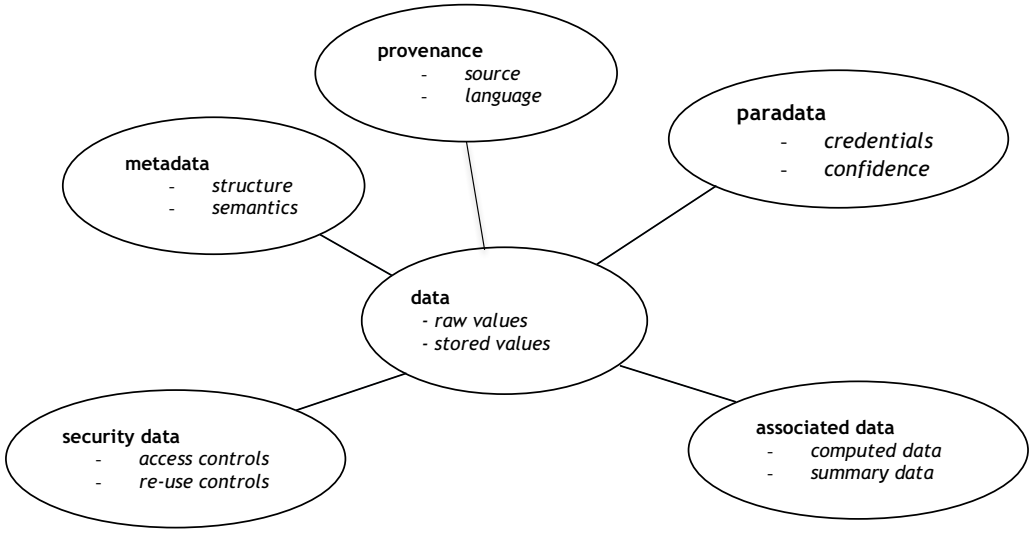


Fig 1: The data manifold Data is characterized not only by its values, but also by what is loosely termed its "metadata", which can be analysed into metadata proper, provenance data, paradata, security data and various computed summaries, etc.

In relation to the second and third questions above, I argue that it is necessary to differentiate data in the EPR according to its source and method of derivation (*provenance*), its form (*metadata*), its reliability (*paradata*), and other relevant characteristics. It will be argued that this enrichment of the data can be exploited to manage its use and reuse. (See Fig. 1, The Data Manifold.)

1.3.2 Electronic Patient Records – Research Issues

A number of European projects addressed technical and workflow issues arising from the use of mixed patient and imaging data. In the MammoGrid project, requirements were identified through a common graphical language that enabled physicians and developers to agree on a common specification, which was then gradually realised through successive refinements. While in MammoGrid the ostensible purpose was support for European collaboration in healthcare across borders, in Health-e-Child, the focus was research. Although this problem was tackled systematically at a later stage, in the Health-e-Child project *ad hoc* protocols had to be devised that satisfied clinicians, researchers and ethicists, proving that data were being shared with due regard to the project's regulatory framework (for research protection), with due consents (for patient protection) and without disrupting the clinical process.

A more demanding project in the field of electronic patient records was EuroPGDcode, concerning collaboration between European assisted reproduction clinics in tracking outcomes from pre-implantation genetic diagnosis (PGD). Although the data was ultimately pooled for statistical purposes, the complexity of assisted parenthood resulted in each pregnancy having a large and variable number of distinct fields to be tracked. With a view to accurately tracking the health of children born following PGD, it was necessary to track multi-parented children, whose siblings might well be differently multi-parented, with the possibility that there might not even be a one-to-one relationship between fertilized oocytes, implanted embryos and live births.

In work undertaken in my last position at NorthShore University HealthSystem, I led a team that sought to address several of these issues against the limitations of the underlying EHR. This has been done through Structured Clinical Documentation Systems (SCDS), a means of recording discrete data from patient-physician encounters in a way that allows both the capture of precise data in compliance with the principle of "one source of truth", and stores that data discretely, rather than embedded in a text note, so that it can be analysed at a later stage.

1.3.3 Electronic Health Records

While personal patient records are primarily of value in the care of the individual, aggregated records may provide valuable information for public health, for quality improvement, and for research into particular conditions, comorbidities, treatments, and other questions of *evidence* in the quest for science-driven, evidence-based practice. I may use the term electronic health record (EHR) to refer to the aggregation of EPRs. Further questions arise from such aggregation:

What are the legitimate (secondary) uses of EHR? A better form of this question is: *What legitimizes secondary uses of EHR?* In this form, the question can be analysed further.

—Have the patients whose records have been aggregated in the EHR consented to its secondary use? The patient's consent is often required to be "informed", i.e. given in the light of full disclosure of the purpose of the proposed secondary use, so that – in principle – the patient may object even on the grounds of ideological difference from the implicit or explicit goals of any research or policy analysis based on data that includes his or her own.

—*Does the EHR hold de-identified data? Anonymized* (i.e. irreversibly de-identified) or *pseudonymized* (reversible under closely controlled conditions), and to what standard? It has often been asserted that de-identified data may be reused without further permission from the patient. However, there is a significant body of work showing that intelligent fusion of the data with public sources (such as the census or electoral rolls) may be exploited to re-identify the record, at least with a given degree of certainty. [38]

In conjunction with the PhD work of my former student Hanene Rahmouni, I have shown that suitable annotation of the data with privacy-related metadata can be used to manage data exchanges. The argument will be extended to the management of the data more generally, including such issues as control of storage allocation and duplication in distributed platforms such as grids and clouds.

This requirement bridges across from work funded by the EU and conducted in the UK to current work funded by PCORI and developed in Chicago. Both cases touch on data and information and connect to the technology and to policy making.

1.4 Organization of this Dissertation

Chapter 1 has provided some of the history and motivation for the work that has been assembled in this dissertation. As already confessed, there was no grand programme at the outset which I set out to realize. Rather, beginning with a technical background and a growing interest in biomedical informatics, at least after my very first engagement in the mid-80's, I have taken such opportunities as have arisen to deepen and to broaden my involvement. Throughout the late 1980's and the 1990's, I was engaged in small-scale individual projects, some of very considerable interest in their own right, but without much connection to the professional world of BMI. This changed in the early 2000's, when the opportunity to work in highly connected projects arose. The next three chapters trace this evolution through its stages, roughly, the European healthgrid technology projects, then the European healthgrid policy projects, and last the largely American Learning Health System period which continues to this day.

Chapter 2 takes up the story of MammoGrid, a project that originated in deep and deeply informed technology. Considerable expertise was brought to bear on the problem of standard-ization of mammography, the provision of remote annotation services, and the facilitation of tele-consultation with fellow radiologists. What my team from UWE, working at and through CERN, brought to the project was knowledge and experience of complex databases, while I personally happened also to be conversant with the biomedical field. MammoGrid was by all accounts a great success and was followed in spirit and in technology by a number of other projects, notably Health-e-Child and neuGRID, each extending the scope and reach of earlier projects either in the complexity of diseases covered or of services offered.

Chapter 3 takes up the next phase of development. The European Commission's e-Health Unit, led by the highly energetic Jean-Claude Healy, favoured grid technology as "the infra-structure of choice" for biomedical applications, and by extension biomedical research. They therefore encouraged the formation of an association, HealthGrid, incorporated in France in 2004, and then invited HealthGrid to draw up a programmatic "white paper".

The White Paper project was duly launched at the 2004 HealthGrid Conference in Clermont-Ferrand, and by the following year, at HealthGrid 2005 in Oxford, the white paper was published under the names of its three editors, Vincent Breton, Kevin Dean and Tony Solomonides. While the paper was discussed and generated considerable debate about the extent of "automation" of medicine that might be envisaged—a prominent scientist declared after one particularly futuristic keynote that she "would not wish to live in a fascist state"—a more concrete proposal was put forward, to be funded by the EU's Framework VI Programme, for a project to create a "road map" for healthgrids. This became the SHARE project in which I was fortunate to play a major part.

Chapter 4 shifts focus from Europe to the United States. The Institute of Medicine, as the National Academy of Medicine was then known, launched a series of workshops around the theme of the "Learning Health System". The first of these was held in 2007 and I participated for the first time in 2010 at a workshop focused on infrastructure. That year also saw the enactment of the Patient Protection and Affordable Care Act which, among many other major changes to US healthcare, introduced the Patient-Centered Outcomes Research Institute (PCORI). PCORI funded a number of small to medium-sized projects, some with a focus on methodology, until 2013 when it launched a major initiative to create "Clinical Data Research Networks" and "Patient Powered Research Networks" (CDRNs and PPRNs). I was actively in the Chicago CDRN proposal that eventually was funded as the CAPriCORN network. This has provided a vehicle for practical work towards the ultimate goal of a Learning Health System. This work is currently continuing.

Chapter 5 offers some reflections on the project to bring about or re-engineer the current chaotic system into a Learning Health System. It touches on two issues: first, engineering, comparing the healthgrid model with the ultra-large-scale systems (ULSS) model; and second, the engagement of stakeholders other than researchers and gatekeeper-physicians in research. Both healthgrid and ULSS are conceived as means of bringing together autonomous systems—distributed computational power, virtual organizations, federated databases—in a way that does not deprive them of their relative independence, and also by means of something like a systemic negotiation: they have to have the means to "under-stand" each other, and better still to learn from and about each other towards such under-standing. The second major point is more assertive, almost polemical: writing the first column of *Patient Voice* in the magazine Cancer World, in 2004, Anna Wagstaff asserted in the title of her piece "Nothing about us without us". This, together with the notion that in healthcare—not yet in research—we have progressed from treating patients as objects of no opinion, to economic agents with consumer power, to now see them as possessing both valuable knowledge and the capacity to enhance or obstruct their own healthcare depending on their motivation. The not-very-fine point my closing contribution makes is that this is also true of research, albeit with more variety of inputs and possible outcomes.

Blank page

Chapter 2

Phase I - MammoGrid & Health-e-Child; EuroPGDcode

Blank page

2 Phase I – MammoGrid & Health-e-Child; EuroPGDcode

2.1 Grid Computing and Health

The concept of grid computing was motivated by a variety of unmet technical requirements in distributed computing and by a sense of opportunity in the world of "big science". It coalesced in the late 1990's from various strands of technical research, empirical analysis of network and computing capacities, and developments in research policy. By the turn of the century, the need was abundantly evident in the physical sciences. Both particle physics and astronomy were on the point of launching experiments that would generate very large volumes of data at an unprecedented rate. Conventional architectures would not be able to sustain performance, nor provide the effective capacity necessary for storage. Meanwhile physics simulations in preparation for the experiments were already demanding increased performance for experimental studies. Meanwhile, with the advent of genomics, by the mid-1990s the field of bioinformatics had shifted focus from the study of information processes in biological systems to the narrower sense of analysis of the genome and thence the process of translation of DNA to protein and beyond—the eponymous "proteome" and thereafter the "metabolome". The number and complexity of comparisons required for search, matching and alignment in genomic sequencing and in the discovery of specific genes also demanded an order of magnitude increase in the computational power available to researchers. Once the extension to life sciences had been realised, possible applications to medicine and healthcare were a likely next step.

Technically, it had long been observed that the unused compute cycles in idle workstations represented what one informatician described as "the inverse tragedy of the commons" [39]. High performance computing (HPC) applications were perennially short of processing infrastructure, while workstations on researchers and other employees' desks were sitting idle, not only when they were not being used, but even when carrying out ordinary processing tasks. Myron Livni's Condor project [40] and David Anderson's SETI@home [41] demonstrated the possibility of harnessing spare cycles either for local *ad hoc* distributed computing or for wide-area Internet-based computing. Ian Foster, Carl Kesselman and Steven Tuecke [42, 43] pulled these ideas together into a coherent narrative that introduced the concept of "virtual organizations", thus tying the social (and potentially, the economic) organization of big science into the design of the technical infrastructure. Virtual organizations (VOs) were to be loosely affiliated groups of institutions, researchers and projects that might come together perhaps only for a short time to address a specific issue, but they could equally be, or become, longer term collaborations. Foster and Kesselman edited a seminal collection of foundational papers [44] that became in effect the *de facto* definition of, a potent manifesto, and a blueprint for a grid.

There was a marked difference in approach to, and even uses of, grid computing between scientific communities and continents, partly reflecting attitudes to funding. In the UK, scientists benefited both from the EU's pronounced trend towards large multinational collaborations and from the UK's own tightly controlled and targeted *e-Science* programme. On becoming Director General of the Research Councils in 1998, Dr. John Taylor launched the e-Science programme as the flagship of his tenure in this commanding position. He argued persuasively that scientists perform the "role of middleware" as they manually transport (or worse, re-key) data from one laboratory apparatus to another or collaborate by patching data and software into emails in order to exchange ideas. The UK e-Science programme would use appropriately designed infra-

structure for the missing middleware, to support machine-to-machine interoperability and scientist-to-scientist collaboration.

The medical field in which the idea of collaboration made immediate sense was Radiology. Images are stored in large files, there is often a need for second opinion or to distribute work where radiologists may be less busy (or provide a 24/7 service for a fee), and the users already have significant exposure to technology. It was indeed in Radiology that one of the most interesting and ambitious early grid projects was conceived.

2.2 MammoGrid

MammoGrid, and its sister project eDiamond, were the brainchild of Professor Michael Brady at the University of Oxford, the former in collaboration with Professor Roberto Amendolia on secondment to CERN from the University of Sassari in Italy [45]. eDiamond was funded by the UK e-Science programme and—largely because of external industrial interest—was kept very much apart from MammoGrid, the only exceptions being two comparative publications. MammoGrid was proposed and funded more in keeping with the spirit of open science supported by EU Framework programmes, although here too there were some commercial interests to be protected. The goal of MammoGrid was to demonstrate remote synchronous and asynchronous collaboration between breast cancer screening clinics, including provision of a validation service, a second opinion service, standardization and automatic annotation of mammograms.

Responsibilities among the different partners in the collaboration were distributed as follows: Clinical (Addenbrooke's and Udine) to specify, use and evaluate the system; Informatics (CERN, Division of Technology Transfer and UWE, Bristol) to capture requirements, design data structures, determine workflows, and implement the grid infrastructure; Medical Technologies (Oxford Medical Vision Laboratory, Mirada Solutions and University of Pisa) to deploy and adapt their respective mammogram standardization and annotation services.

In the first phase, my contribution focused on capturing requirements through use cases. The language and diagrams of UML use cases, with suitable explanation and interpretation to begin with, proved a remarkably smooth intermediate language between physicians and software engineers. This early work was reported in:

A. F Estrella, C del Frate, T Hauer, R McClatchey, M Odeh, D Rogulin, S R Amendolia, D Schottlander, **T Solomonides**, R Warren. *Resolving Clinicians' Queries Across a Grids Infrastructure* **Methods of Information in Medicine** Vol **44** No 2. 2005 pp 149-153. ISSN 0026-1270 Schattauer publishers.

My interest grew in the Standard Mammogram Form (SMF™) [46] that had been devised by Ralph Highnam as part of his Oxford DPhil and then spun off into a company, Mirada Solutions. I undertook a Master's course in Radiology and wrote an exposition of the method as one of two final assignments. The other assignment, an extended essay on MammoGrid, provided the first draft of the paper that eventually was published as:

B. R Warren, **T Solomonides**, C del Frate, I Warsi, J Ding, M Odeh, R McClatchey, C Tromans, M Brady, R Highnam, M. Cordell, F Estrella & R Amendolia. *MammoGrid – A Prototype Distributed Mammographic Database for Europe* **Clinical Radiology** Vol **62** No 11 pp 1044-1051. DOI 10.1016/j.crad.2006.09.032 November 2007, Elsevier publishers.

A further paper presents a retrospective review of the MammoGrid project from a largely technical point of view. This reflective piece was published after the project had effectively been completed. I made a significant contribution to this paper, especially in the discussion of the workflow, in particular on the mediation through use-case models of the interaction between physicians and technologists and the interpretation of radiologists' needs to capture an adequate set of user requirements.

C. Estrella F, Hauer T, McClatchey R, Odeh M, Rogulin D, **Solomonides T**. *Experiences of engineering Grid-based medical software*. **Int J Med Inform.** 2007 Aug; **76(8)**:621-32. Epub 2006 Jun 19.

Other publications about MammoGrid provide additional information and are mentioned here as subsidiary papers. These include a. the earliest announcement of the project in a publication (at MIE 2003), explaining the novelty and design of the project. I made a significant contribution to this paper and presented it at the conference.

The second subsidiary paper, b., is a companion clinical paper to B. above. It proved the epidemiological value of an infrastructure like MammoGrid's by demonstrating that breast density was a risk factor in its own right, not just as an impediment to good imaging. My contribution to this paper was of an editorial nature, making sure that the concepts and language corresponded to the first paper.

Subsidiary Papers

a. S. Roberto Amendolia, Michael Brady, Richard McClatchey, Miguel Mulet-Parada, Mohammed Odeh and **Tony Solomonides**. *MammoGrid: Large-Scale Distributed Mammogram Analysis* in *The New Navigators: from Professionals to Patients* (Proceedings of Medical Informatics Europe 2003) Robert Baud, Marius Fieschi, Pierre Le Beux, Patrick Ruch (Eds.). **Studies in Health Technology and Informatics**, Vol 95 (2003) IOS Press

b. R Warren, D Thompson, C del Frate, R Highnam, C Tromans, I Warsi, J Ding, F Estrella, **T Solomonides**, M Odeh, R McClatchey, M. Bazzocchi, S R Amendolia & M Brady. *A Comparison of Some Anthropometric Parameters Between an Italian and a UK Population : "Proof of Principle" of a European Project using MammoGrid.* **Clinical Radiology** Vol **62** No 11 pp 1052-1060. DOI 10.1016/j.crad.2007.04.002 November 2007, Elsevier publishers.

Closing the loop from MammoGrid back to the work on grids that made it possible, the grid infrastructure for the project was directly borrowed from one of the particle experiments at CERN. Here, grid computing was under active development in anticipation of the voluminous data that would flow from the Large Hadron Collider (LHC) and the computational power that would be necessary even to triage the data into potentially useful or not. In the event, MammoGrid used *AliEn*, the resource broker devised for the Alice experiment [74] as the basis for the necessary grid infrastructure.

2.3 Health-e-Child and other Healthgrid Projects

The healthgrid projects that followed MammoGrid proved more significant in having added to the evidence that the approach could work than in breaking new ground. As additional support for the concept of the grid as a collaboration medium for a certain kind of biomedical project, they provided some of the concrete examples necessary to justify a systematic policy and strategy project, thus paving the way to the SHARE project which is the subject of the next chapter.

2.3.1 Health-e-Child and neuGRID

Health-e-Child was an attempt to take the lessons of MammoGrid and scale them up to three otherwise unrelated paediatric conditions, each requiring a different mode of imaging, to four European centres of excellence in these conditions, and to collaboration with an industrial giant in Siemens. The three conditions were brain tumours (gliomas), cardiac malformations (right ventricular overload cardiomyopathy), and juvenile idiopathic arthritis. The twin goals of the project were to provide knowledge for decision-making and to support clinical studies. In terms of clinical science, the cardiac study proved highly productive, giving rise to further projects. In decision support, I was involved in a simple approach to bring ontologies to bear on the interpretation of patient data and to present information to the physician in the course of decision-making. I worked with my Research Fellow, Tamás Hauer, on the introduction of ontologies and semantic reasoning, mainly to characterize tumours by location and type and visualize their prevalence. This resulted in a rudimentary system that was reported in two minor publications, one of which is appended as a subsidiary paper:

c. Tamás Hauer, Dmitry Rogulin, Sonja Zillner, Andrew Branson, Jetendr Shamdasani, Alexey Tsymbal, Martin Huber, **Tony Solomonides**, Richard McClatchey. *An Architecture for Semantic Navigation and Reasoning with Patient Data - Experiences of the Health-e-Child Project.* in The Semantic Web – Proceedings of the Seventh International Semantic Web Conference (ISWC 2008). **Lecture Notes in Computer Science** Vol. 5318; pp. 737-750. Springer 2008.

The greatest sophistication in imaging-related healthgrid projects came with neuGRID [75]. Its goal, in the words of its website, was "to become the 'Google for Brain Imaging', i.e. providing a virtual imaging laboratory that can be accessed by any scientist with a PC and web browser. This new environment will allow researchers in the field of Alzheimer's disease answer complex neuroscientific questions." My engagement with neuGRID focused more on its implications for healthgrid policy development and was reflected more in the work on the SHARE project.

2.3.2 EuroPGDcode

EuroPGDcode was a very different kind of project. It was funded by the European Agency for Health and Consumers, so that it was by definition a project to deliver a particular innovation for real use, not just as a demonstrator project. The nominal goal of the project was to bring about some order in the terminologies and codes that were in use in the field of pre-implantation genetic diagnosis (PGD) in assisted reproduction. However, in reality, it focused much more on the collection of data and the design of the associated data structures to facilitate faithful representation of complex information.

The European Society for Human Reproduction and Embryology (ESHRE), through its PGD Consortium, has collected statistics on the use of PGD from clinics in Europe and worldwide for nearly two decades. Their initial goal, as stated in [47], was to "undertake the first systematic and long-term study of the efficacy and clinical outcome of PGD." The field is controversial, not only because of ethical objections to "designer babies"—hardly justified in the case of PGD—but also because of methodological uncertainties. PGD requires DNA from the embryo; in traditional, early PGD, this required harvesting a cell at the four-cell stage of embryo development. Subsequent work has enabled polar body biopsy and ultimately free circulating DNA in blood to be used. Ethical objections are more frequently raised against pre-implantation genetic screening (PGS) and confuse the issue. PGD is targeted at genes that the parents are known to carry, and so is aimed at ruling out specific serious, usually fatal, conditions. PGS, on the other hand, entails a broad sweep across a number of serious genetic conditions in the absence of any particular reason to suspect that they may be present. It is in this context that ESHRE's PGD Consortium has collected evidence of PGD over the years. The initial proposal did not envisage either data collection or a healthgrid application, but on harmonizing terminologies. In the face of the partners' pragmatic needs, it was reoriented to easing the collection process. Resources did not allow a true healthgrid to be deployed, but an internet-based collection service was established. The most interesting part of this project lay in the extreme complexity of the data to be collected.

A number of publications arose from this work, the most salient of which is:

D. Olive, M., Lashwood, A. and **Solomonides, T**. (2011). *A retrospective study of paediatric health and development following pre-implantation genetic diagnosis and screening*. In: Olive, M. and Solomonides, T. (ed.) **IEEE Proceedings of the 2011 24th International Symposium on Computer-Based Medical Systems** (CBMS 2011), p. 32-38.

Of the many projects reported in this dissertation, EuroPGDcode stands apart as an exemplar of the need and the value of the "data manifold" model. If ever a full cloud-based application is developed, it will require all the elements, including confidence data (e.g. the level certification of the laboratories involved), provenance (e.g. details of the methods used), metadata on the semantics of data submitted from different participating clinics, as well as, more obviously, security and privacy controls. Had this case study been available at the time of the SHARE road map (see Chapter 3), it would have provided an interesting challenge for a healthgrid solution.

Blank page

Chapter 3

Phase II - HealthGrid White Paper & SHARE Roadmap

Blank page

3 Phase II – HealthGrid White Paper and SHARE Road Map

3.1 HealthGrid

The first wave of "healthgrid" projects included MammoGrid. By the time the second wave of such projects was funded, including Health-e-Child, the European Commission, which had always encouraged collaboration between projects, went a step further and recommended the formation of an association to support and promote projects with the common theme of using some variety of grid computing as the infrastructure on which the work would be done, where the results would be stored and shared, and through which collaboration might proceed. Critical to these ambitions were the subscription model of participation and the concept of "virtual organizations" (VOs). As noted in Chapter 2, grid computing consolidated certain ideas about resource sharing and exploitation of surplus or redundant capacity in scientific networks. The subscription model meant that a node wishing to benefit from the vast resources it would gain access to, had, in return, to allow its own spare capacity to be used by other nodes on the network. The model was successful precisely because need was not constant at any node. However, in the medical world, certain features of this "sharing" could be problematic: data might be moved for processing, if not for storage, to locations outside the strict boundaries dictated by their source regulatory framework. When there was sufficient local processing power, one proposed solution to this was to move the algorithm to the data, rather than the data to the algorithm. So long as we were dealing with academic demonstrator software, this was not a problem, but as soon as the necessary software required commercial licensing, this "solution" was no longer viable.

A better solution lay in a recursive or nested structure of (VOs of) VOs with appropriate regulatory frameworks at each level. I bracket the nested structure in this way to signify that there is always a bottom layer: this might be a VO of a hospital system under the authority of a single trust (in the UK) or "covered entity" (in the US), which can hold data with the highest regulatory privileges. Next might be a national or state-level super-VO integrating many local VOs; this would be subject to national or state legislation. Then at the supranational or federal level, a hyper-VO of super-VOs would operate under the most restrictive regulatory regime. With the idea of market economics being introduced into the healthcare space, it was also apparent that the grid could potentially provide a "marketplace" for competition between certain services. This was discussed explicitly in MammoGrid, where, e.g., different image annotation services might compete. This was still more evident in NeuGRID, where different data sets, services, image processing algorithms, and so on, could be available to choose from in the same ambient grid. Implicit in all this lay the question whether the ultimate goal was a single Healthgrid, like the (capitalized) Internet, or a multitude of healthgrids, each with a limited scope, but possibly able to interact with others and so still form *ad hoc* VOs. The former appeared the more elegant solution, albeit fraught with regulatory issues, the latter the more pragmatic and realizable option, especially as it allowed for regulatory frameworks to be reconciled as VOs are formed "bottom-up".

These ideas—more accurately, debates—were in the air at the second conference on healthgrids (in 2004 at Clermont-Ferrand) where the incorporation in France of a new association, to be designated *HealthGrid*, was decided upon and its first set of officers elected. Vincent Breton at the CNRS, France, took a leading role and suggested the formation of a working group to

establish the vision and mission, goals and prospects of the new association. A preliminary paper by Breton, Solomonides and McClatchey was published in the same year [48] and this led to the formation of an editorial group comprised of Breton, Kevin Dean (then of Cisco) and Solomonides which was tasked to bring together a larger definitive and more comprehensively representative work. The result of this was the HealthGrid White Paper [49].

An equivalent debate did take place in the United States, fuelled in part by massive funding for high profile projects such as The Cancer Biomedical Informatics Grid (caBIG) [50]. This project was finally judged to have been at best a partial success [51], but discussion of bottom-up grid development of a similar scope to that of the HealthGrid White Paper did take place; for example, the 2008 proposals for health information sharing in Utah [52] bear a marked resemblance to ideas from HealthGrid.

3.2 The White Paper

The HealthGrid White Paper [49] was remarkable in its day for three reasons: its timing was fortuitous, its fundamental assumptions had been tested in earlier proposals and were generally accepted, and its scope was almost comprehensive.

At its meeting in Lisbon in March 2000, the European Council had included the mandate, as part of its economic agenda for Europe,

> *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.*

Strategic thinking among HealthGrid members, overoptimistic though it proved to be, was remarkably convergent with this EU goal. Although by November 2004 the report by Wim Kok's Review [53] had noted a failure to work effectively towards the goals of the Lisbon Agenda, the EU Commission was pressing ahead with programmes to promote economic growth. With increasing integration in mind, and mobility among the goals, a move towards infrastructures that would allow seamless healthcare delivery and progress in biomedical research was perceived to be an important goal. Commission officers who had encouraged the formation of the association HealthGrid in the first place, were equally encouraging of the development of a White Paper to flesh out the vision with rationale, principles and, above all, examples.

A clear principle was high connectivity. In a remarkable departure from traditional thinking, this was also linked to the idea of linking concepts and models from different levels of biosocial organization, informatics practices, and pathologies. This was nicely captured in the BioInfoMed [54] hierarchy:

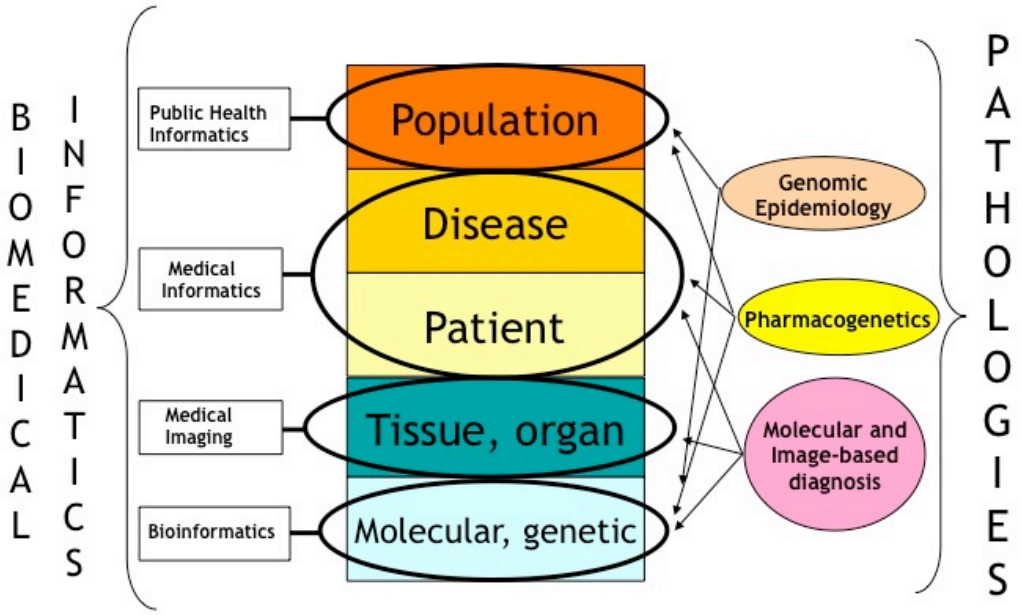


Fig 2: The BioInfoMed Schema Levels of biosocial organization, from the molecular through to the population, correspond, on one hand to subdisciplines of biomedical informatics and on the other to biomedical disciplines. Neither correspondence is precise, but suggests possible links.

Credit: Adapted with permission from a presentation by Fernando Martin-Sanchez

---

The implicit—and disruptive—notion here is that the connectedness that would be fostered by the grid would not only bring different disciplines together, breaking down traditional academic boundaries, but by the same underlying means also allow scientists working at these different levels to integrate their models, so that, for example, a molecular model of tumour development might be coupled with a tissue model to characterize tumour growth. It is possible to claim this diagram as the progenitor of what eventually became the concept of the Virtual Physiological Human.

The breadth of applications was also considerable: sandwiched between explorations of the business case for the grid and its ethico-legal dimensions, are studies of imaging, computational models of human biology, pharmaceutical research and development, epidemiology, and genomics. This range was necessary to make the case that the grid could really become a "healthgrid", the theme of the third HealthGrid Conference in whose proceedings the White Paper was published.

Another theme emerges here also that will recur in this analysis of the healthgrid concept. Both as editor and especially as author I had focused on familiar applications, such as imaging and epidemiology, and on ethical issues of privacy and confidentiality. I had highlighted, in the context of MammoGrid, the potential of the grid to be a marketplace platform, where services might compete on quality, performance, and cost. However, I had neglected aspects of cost sharing and the contractual dimension that would prove necessary in the full commercial exploitation of the paradigm. As "grid" transmuted into "cloud", these economic matters came

to the fore, so that it was possible for someone somewhat superficially to characterize cloud computing as "grid computing with a business model". Did the cloud, as an economics-aware infrastructure, achieve the planned goals of HealthGrid more quickly than could ever have been achieved without market forces?

E. V. Breton, K. Dean, **T. Solomonides**, *The HealthGrid White Paper*, in *From Grid to Healthgrid*. **Studies in Health Technology and Informatics**, vol. 112, ISBN 1-58603-510-X, ISSN 0926-9630 IOS Press.

3.3 SHARE Methodology

The successes of early healthgrid projects and of the first HealthGrid conferences led to the White Paper. This was well received by the community, but could not be defended as a rigorous scientific study; the need for a thorough examination of the issues was to be addressed through a formal "Specific Support Action" in the language of the European Union, the SHARE project [55]. The project took the White Paper as its starting point and sought to determine the steps necessary to establish an integrated, effective healthgrid infrastructure. The methodology recognized "non-functional" business and ethical requirements and "functional" technical developments that would have to be accomplished before a healthgrid could be realized and deployed. The steps, which are described in this IJMI paper; an exposition closer to the time of execution was presented and selected in the "best paper" category at MedInfo 2007.

F. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez *SHARE road map for HealthGrids: Methodology* **International Journal of Medical Informatics**, 78S (2009) S3–S12

G. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez. *SHARE, from Vision to Road Map: Technical Steps.* **Studies in Health Technology and Informatics** Volume 129: ***Building Sustainable Health Systems*** - Proceedings of the 12th World Congress on Health Informatics – MedInfo 2007. IOS Press 2007

The SHARE Collaboration presented numerous papers in the process of refining its proposed road map. Significant ones from the point of view of my contributions were:

d. Vincent Breton, Ignacio Blanquer, Vicente Hernandez, Nicolas Jacq, Yannick Legre, Mark Olive and Tony Solomonides. *Roadmap for a European Healthgrid.* **Studies in Health Technology and Informatics** Volume 126: ***From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences*** - Proceedings of HealthGrid 2007 pp.154–163. IOS Press 2007

which represents the first attempt at a complete road map, albeit lacking in detail, and a further publication,

e. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez, Isabelle Andoulsi, Jean Herveg, Petra Wilson. *SHARE Roadmap 1: Towards a debate.* **Studies in Health Technology and Informatics** Volume 126: ***From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences*** - Proceedings of HealthGrid 2007 pp.164–173. IOS Press 2007

which sought to define the outline of a debate around the first road map. This debate took place at a succession of venues, including workshops at the annual meeting of the European Health Management Association, at a meeting of the EGEE collaboration and another at the Open Grid Forum, at CCGrid, and CBMS Conferences, and on numerous occasions as *ad hoc* seminars at various academic settings. The two articles d. and e. above are listed here as subsidiary papers.

3.4 SHARE Road Map

There was no single path from the first road map to the ultimate product that was delivered some 18 months later. One track led through technical requirements and the ways in which development and deployment could be phased so as to make up a realistic project plan for the delivery of a functioning infrastructure, including security aspects. A second track undertook an in-depth analysis of ethical, legal and social issues, including privacy and confidentiality, organizational changes in workflows and in knowledge flows, possible impacts on reporting and control structures, and issues of liability. In retrospect, its economic analysis, at least compared with the in-depth study of legal issues, was not very deep. Nevertheless, a third strand of work, through case studies, investigated the applicability of these ideas to innovative medicine, including drug discovery and development, to epidemiology and public health surveillance, to collaboration in the biosciences and coordination of tertiary care.

These three elements—technology, regulatory considerations, and case studies—were developed independently and could not readily be integrated into a coherent road map. I led the integration effort from UWE with the assistance of Hanene Rahmouni and Mark Olive. The collaboration provided the crucible in which our ideas were tested until it was possible to make a first draft of the road map public to a group of about thirty experts, twenty or so of whom were invited to a day-long meeting in Brussels to critically evaluate its proposals. The road map underwent its final amendments and improvements and was submitted to the European Commission as its final deliverable. The commission requested an abridged version for publication. Once again, working on behalf of the collaboration, my two students and I assembled the “short” SHARE Road Map which was published as a booklet by the Commission and after further editing and peer review by the editorial board, republished as the opening chapter of Cannataro’s *Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare* [56].

This article is included as a principal publication in this dissertation.

H. Mark Olive, Hanene Boussi Rahmouni and **Tony Solomonides** (UWE, Bristol, UK); Vincent Breton, Nicolas Jacq and Yannick Legré (IN2P3, CNRS, Clermont-Ferrand, France & HealthGrid, EU); Ignacio Blanquer and Vicente Hernandez (Universidad Politécnica de Valencia, Spain); Isabelle Andoulsi and Jean Herveg (Universitaires Notre-Dame de la Paix, Belgium); Celine Van Doosselaere and Petra Wilson (European Health Management Association, EU); Alexander Dobrev, Karl Stroetmann and Veli Stroetmann (Empirica GmbH, Germany). *SHARE: A European Healthgrid Roadmap* in **Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare** (Mario Cannataro, Ed). Chapter 1, pp. 1–27. IGI-Global, Hershey, PA. 2009.

3.5 Grid vs. Cloud – a brief digression

Much of our discussion of "grid computing" readily translates to the language of "cloud computing". However, the two are not synonymous and in view of the pervasive success of cloud computing in the world at large, I will briefly explore the clear differences and comment on the apparent failure of the SHARE road map to address the economics of highly distributed infrastructures in a convincing way.

The distinction—or, depending on one's point of view, the similarities—between clouds and grids became a contentious issue because the later technology, clouds, appeared to be usurping certain features of grids without acknowledgement. Grids had emerged laboriously over many years of academic work, culminating in a diversity of systems and protocols. Clouds appeared to be the adaptation, perhaps annexation, of grid principles to large distributed infrastructures motivated by commercial interests. The term "cloud", as a descriptor of a distributed computational architecture, was certainly in use by the end of the SHARE project and was contrasted with "grid" at the gathering of experts to evaluate the SHARE roadmap[vii]. It soon became clear that in parallel with the SHARE project, the EU had also funded a similar study on cloud computing, motivated largely by a perceived need to bridge the gap from academia and the sciences to industry and "production" systems.

In his 2008 blog, *There's grid in them thar clouds*, Ian Foster identifies many elements of grids that have found their way into clouds [57]. In that same year, at the Grid Computing Environments Workshop, a highly technical workshop on grid computing, we encounter the paper *Toward a Unified Ontology of Cloud Computing* [58]. One easily forms the impression that, at this stage, the grid community is looking to understand and respond to the phenomenon that has upstaged it. Viewed from a little further afar, it would appear now that the benefit of cloud technologies went beyond the kind of resource sharing that grids adopted and used the infrastructure to address a number of requirements, some arising from vendors' interests and some from those of their clients. Chronologically first among the advantages of clouds was, I believe, "application service provision", i.e. the ability to manage software, such as office applications, remotely. This benefits the software vendor, since it makes it possible effectively to control the licensing of the software to authorized users. The arguable benefit to the client lies

[vii] Fabrizio Gagliardi, a member of the review panel who had moved in 2005 from directing the EU-funded EGEE project at CERN to a senior technical position at Microsoft, posed the question, "What is the next thing in distributed high performance computing, is it the cloud?"

in the remote management of services, with a corresponding reduction in application administration costs, but at some risk of losing data that is only accessible through an application that is no longer under their control. Second, in many successful grid collaborations, there was already a background conversation on the cost of services, fair charging models, the need to load balance, perhaps through a market mechanism, or through planning and batch processing—as though the grid world was inadvertently reinventing the economic system arguments of the nineteenth and twentieth centuries. Clouds made elasticity and scalability through virtualization their principal claim in the economics of business information systems.

SHARE identified uses for data grids and computational grids, two prime paradigms that pre-existed the project. Taking its cue from the UK e-Science programme, it also advanced the "collaboration grid" as a third paradigm, with a future envisioned "knowledge grid" as the culmination of all these types, though one that could not yet be realized. This typology bears only the most superficial resemblance to the cloud distinctions of Infrastructure, Platform or Software as a Service (IaaS, PaaS, SaaS or, collectively, XaaS—anything as a service). The succinct assertion that "clouds are grids with a business model" provides a useful contrast. SHARE had recognized economic constraints, had factored organizational and regulatory concerns into its designs, and had asserted the possibility of its electronic network serving as a marketplace. It was therefore closer in spirit to cloud computing than any comparable proposal from the same domain. Perhaps—to address what I still consider a puzzle—this was the reason why we failed to recognize the cloud as a genuinely new paradigm.

3.6 Data Sharing, Secondary Use and Regulatory Frameworks

In closing the chapter on SHARE and the road map, I wish to address certain issues that provided a focus of concern at the time of the project and led to some ideas that I have come to suggest since then.

One of the persistent issues in a project in which potentially identifying data will be shared is the protection of that data. Regulatory frameworks[viii] envisage such instruments as legally binding Data Sharing and Data Use Agreements, where two parties intend to collaborate, and Business Associate Agreements, where one party serves the other under contract. One of the research problems I formulated and proposed to my student Hanene Rahmouni was the translation of legal rules (expressed in some form of declarative logic) to operational rules in an infrastructure (expressed in some form of deontic logic of permissions and obligations). It seems a common experience that researchers find regulatory frameworks rather restrictive. Policy makers and technologists do not join in a single conversation, and law and policy tend to be handed down to technologists without space for negotiation. I was, in a sense, asking, could we, as technologists force a dialogue by showing that we could meet the regulatory framework halfway, build some significant part of it into our technology?

Hanene showed that data sharing may be viewed as a transaction of the form

<preconditions> Conditional Actions <postconditions>.

[viii] This is the regulatory language in the United States. "DSAs", "DUAs" and "BAAs" are often necessary before research can begin in earnest.

The preconditions may be that valid agreements are in place and that the patient's consent has been given and applies to this transaction. The postconditions relate to what the receiving party must undertake to do or not to do. Dr. Rahmouni's solution to this problem has been published in a number of journal papers. My part in this work was advisory in relation to the technical content and in extensive co-authorship in terms of interpretation, validation, explanation, and dissemination. I presented the problem as a trade-off between creativity and compliance in a refereed presentation to the 16th World Congress on Medical Law.

I. **A E Solomonides**. *Compliance and Creativity in Grid Computing.* 16th World Congress on Medical Law, Bioethics Track. Toulouse 2006.

I developed this theme a little further in a collaboration with two colleagues from Middlesex University, presented to the Second International Conference on Medical Imaging and Medical Informatics in Beijing in 2007. The paper was subsequently re-reviewed and invited as a journal contribution. It is listed as a subsidiary paper:

f. Penny Duquenoy, Carlisle George, **Anthony Solomonides**. *Considering something 'ELSE': Ethical, legal and socio-economic factors in medical imaging and medical informatics*. **Computer Methods and Programs in Biomedicine: Medical Imaging and Medical Informatics (MIMI)** Vol 92:3, 2008, pp. 227–237.

3.7 Institutions and Norms, Argumentation and Agents

Following the principle to automate as much as possible in any data sharing interaction, I was naturally led to consider the use of agents in settling matters under different jurisdictions. I developed this theme at a 2010 keynote address to the PRIMA conference and developed it into a refereed paper that appeared in 2012.

g. **Tony Solomonides** *Healthgrids, the SHARE project, medical data and agents: retrospect and prospect* in **Lecture Notes in Artificial Intelligence (LNCS 7057)** ***Principles and Practice of Multi-Agent Systems*** Revised Selected Papers from the 13th international conference on Principles and Practice of Multi-Agent Systems (PRIMA 2010) pp. 523-534 Springer-Verlag Berlin, Heidelberg 2012

At this conference I learned of the work of Frank Dignum, Henry Prakken and others of the Amsterdam school and began to formulate my ideas in terms of norms and institutions. I took norms to be breakable rules (in the spirit of "break the glass") where the breach had to be justified after the fact in some way. Institutions would then be abstracted as the collection of norms that apply in their interactions.

In 2012, I was invited to give a keynote address to the HealthGrid and Life Science Grids conference in Amsterdam. I developed an example in the spirit of this theoretical framework, based on an exchange of data as might have taken place in MammoGrid:

> *Dr. House in the UK wishes to share a series of mammograms with Dr. Casa in Italy. The purpose of this is to get a second opinion. The patient has given consent that allows her images to be shared provided the purpose is delivery of care—which, indeed, it is. Dr. House lets Dr. Casa know that the only (post)condition he must impose is that the image must be destroyed or deleted once it has been used. Dr. Casa points out that her professional insurance requires her to keep all images on which she gives an opinion for a period of at least two years.*

This contradictory situation need not be the end of the collaboration. If the rule in England is treated as a norm, an exchange can take place along these lines:

> ***House*** *– So, my only condition is that you must delete the mammograms after you have given your opinion.*
>
> ***Casa*** *– But you know I cannot do that. I will keep the images for at least two years, provided no controversy arises. If any does, of course, I will need to keep them even longer.*
>
> ***The exchange of data takes place****.*
>
> ***House*** *– OK, you have "broken the glass", but in these circumstances ("for good cause") I am empowered to modify the condition. You must undertake to delete these images at the earliest time consistent with your situation at the time. Until then, your record [where?] will show that you have broken the glass.*

This suggests that a logic of argumentation may be best suited for this type of interaction. I am currently seeking to further develop this concept to explore the limits of confidential patient data sharing.

Blank page

Chapter 4

Phase III - The Learning Health System

Blank page

4 Phase III – The Learning Health System

4.1 Developments in the United States

The Institute of Medicine, as the National Academy of Medicine was known in 2006, had instituted a Roundtable on Evidence-Based Medicine whose vision statement began:

> *The Institute of Medicine's Roundtable on Evidence-Based Medicine has been convened to help transform the way evidence on clinical effectiveness is generated and used to improve health and health care. We seek the development of a* learning healthcare system *that is designed to generate and apply the best evidence for the collaborative healthcare choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in health care.*

The first report, *The Learning Healthcare System: A Workshop Report* [59], was published in 2007 and set in motion a number of different trains of activity, including studies of the necessary digital infrastructure, on citizen engagement, on data commons, and on costs—indeed, with a redesignation of the meetings to *Roundtable on Value & Science-Driven Health Care*.

Dr. Jonathan Silverstein, at that time Associate Director of the University of Chicago / Argonne National Laboratory Computation Institute, had been aware of my group's work on healthgrids and had encouraged the formation of HealthGrid.US to promote relevant activities in the United States. Through Dr. Silverstein, I received an invitation to speak at the roundtable's workshop on *Digital Infrastructure for the Learning Health System* [60]. This was an exceptional opportunity: whereas in Europe our work had appeared as one of the multitude of threads that the EU was spinning through its Framework Programmes, the IOM initiative appeared to promise an engagement both with the need for Practice-Based Evidence (else, how would the system "learn"?) and the need to secure data re-use in a well-regulated process. In other words, the SHARE programme, the work with Hanene Rahmouni on data annotation and automated compliance, and the fledgling work with Mark Olive on "PBE for EBP" could all be brought together in one movement. The presentation of those aspects of this work that were sufficiently mature to withstand scrutiny were presented and are incorporated in the report [61].

Dr. Charles Friedman, a major exponent of the Learning Health System (LHS) concept has since led this nascent movement through to the incorporation of the Learning Health Community as "a grassroots not-for-profit organization". [62]. Its mission and vision are summed up on its home page:

> **LHC MISSION**
> *The Community's Mission is to galvanize a national grassroots movement in which multiple and diverse stakeholders work together to transform healthcare and health by collaboratively realizing the LHS Vision.*
>
> **LHC VISION**
> *The Learning Health Community ("Community") aims to mobilize and empower multiple and diverse stakeholders to collaboratively realize a national-scale (and ultimately global), person centered, continuous and rapid learning health system (LHS).*

The principles and exemplars that flesh out this vision and mission bear a close relationship to the work presented here. The fundamental principle is the creation of relevant knowledge:

> *[The LHS] will improve the health of individuals and populations. The LHS will accomplish this by generating information and knowledge from data captured and updated over time – as an ongoing and natural by-product of contributions by individuals, care delivery systems, public health programs, and clinical research – and sharing and disseminating what is learned in timely and actionable forms that directly enable individuals, clinicians, and public health entities to separately and collaboratively make informed health decisions… The proximal goal of the LHS is to efficiently and equitably serve the learning needs of all participants, as well as the overall public good.*

This is consistent with, if somewhat broader than, the goals of the Patient Centered Outcomes Research Institute.

4.2 PCORI and PCORnet

The Patient-Centered Outcomes Research Institute (PCORI) was established under the Affordable Care Act with the mission to support research with a focus on the patient, the caregiver, the patient advocate, or the citizen in general, as increasing attention is paid to health maintenance and self-management. The main case for funding this initiative, however, was the prohibitive cost of randomized control clinical trials (RCTs) and the length of time it takes for all the prescribed stages to be completed and a new product or protocol be proposed for adoption. The argument is that with sufficient numbers, observational data should prove rich enough to enable researchers to control for confounders and other possible statistical contaminants. Thus, PCORI supports comparative effectiveness research (CER) into clinical outcomes that matter to patients and their families. PCORI has funded a number of regional collaborations under two headings, Clinical Data Research Networks (CRDN) and Patient-Powered Research Networks (PPRN).

The initial goal of all CDRNs was to create an interoperable infrastructure each within their own network, and to prove it through a number of initial studies. A national network of networks, PCORnet, has been established with the 13 CDRNs and 20 PPRNs as its nodes, again as a national infrastructure for CER. One of these nodes, the Chicago-based CDRN, is introduced next.

4.3 CAPriCORN

CAPriCORN, the Chicago Area Patient Centered Outcomes Research Network, is a remarkable alliance of Chicago-land institutions representing a very diverse community—diverse both in the type of institutions involved and, importantly, in the populations they serve. Among certain of the institutions in CAPriCORN, there is substantial overlap in populations, sometimes as high as 20%. This is largely explained by the economics of healthcare for those with inadequate or no insurance. With no primary care physician (PCP) registration, the Emergency Room (ER) becomes the place where care is received.

This phenomenon presents certain challenges in the design of an integrated information system in the absence of a unique (national or local) patient identifier. Notwithstanding the adoption of a common data model, the architecture of the virtual CAPriCORN repository is essentially a federated one. Indeed, it is doubtful whether the term “repository” is justified, even with the qualifier “virtual” attached. Except where its centralized IRB has given approval to approach patients for consent, all data for sharing within CAPriCORN—and in the wider community at a

later stage—are in a HIPAA-compliant[ix] [63], de-identified format, and assembled only on a study by study basis. The Informatics Working Group has devised a distributed data architecture, a data model with appropriate standards, and a designed data flow engineered to ensure that no protected health information (PHI) is released other than under strictly controlled conditions, at the same time as maintaining the research value of the data that is released. Indeed, even de-identified data is released only for a single study at a time, not as an aggregate population. A pseudo-identifier is generated for each patient in such a way that patients' records that are distributed across different providers in the network can be matched and integrated, without data being moved outside protected home institution environments. Consent is always sought when access to PHI or directly to the patient for patient-reported outcomes is necessary. CAPriCORN was described in an early paper which is listed as subsidiary:

h. Abel N Kho, Denise M Hynes, Satyender Goel, **Anthony E Solomonides**, Ron Price, Bala Hota, Shannon A Sims, Neil Bahroos, Francisco Angulo, William E Trick, Elizabeth Tarlov, Fred D Rachman, Andrew Hamilton, Erin O Kaleba, Sameer Badlani, Samuel L Volchenboum, Jonathan C Silverstein, Jonathan N Tobin, Michael A Schwartz, David Levine, John B Wong, Richard H Kennedy, Jerry A Krishnan, David O Meltzer, John M Collins, Terry Mazany. *CAPriCORN: Chicago Area Patient-Centered Outcomes Research Network* **J Am Med Inform Assoc** 2014;21:607–611.

In the first phase of the project, which was intended as a proof of concept, NorthShore University HealthSystem, where I work, participated in studies of four conditions: Anaemia in in-patients; Obesity and Overweight; Asthma; and Recurrent *Clostridium difficile* Infection. These studies were conducted on de-identified data except for small numbers of patients who consented to be interviewed about their outcomes.

In the second, "live" phase, colleagues and I are participating in a number of national studies, including ADAPTABLE, a study of Aspirin dosing for patients with a documented history of cardiovascular disease, and two obesity-related studies, one on Bariatric Surgery, comparing different types of surgical procedure for their effects on the patient's subsequent health and weight, and another, which I co-lead in our own CDRN, on Short- and Long-Term Effects of Antibiotics on Childhood Growth: do antibiotics in the first two years of life affect the child's weight as he or she grows up? [64] Further studies of patients with high healthcare needs and costs, of patients with chronic obstructive pulmonary disease (COPD), and others are being planned.

4.4 The CAPriCORN Technical Infrastructure

The development of the technical infrastructure was necessarily phased, both to address the heterogeneity of the collaborating institutions, and also to match the evolving ideas (and sometimes asserted requirements) of the overarching national network, PCORnet. It was also necessary to phase technology and workflow specifications so as to remain in step with the

[ix] The U.S. Department of Health and Human Services ("HHS") issued the Privacy Rule to implement the requirement of the Health Insurance Portability and Accountability Act of 1996 ("HIPAA"). In effect, the Privacy Rule requires all identifying information to be removed before any data set can be shared publically.

concurrent development of a common IRB, with its associated policies and procedures, and to remain respectful of the entirely new Patient and Clinician Advisory Council (PCAC) which was poised to become the voice of stakeholders in the system.

First to be designed and agreed was a basic data model to serve as a common design for a CAPriCORN data mart at each institution. It was also axiomatic that there should be a means, ideally independent of the data mart, for a "hashing" algorithm to be applied to patient identifiers so as to generate an almost certainly unique "de-identifier" (with about 98% certainty). This would provide the means of de-duplicating patients who visited and had records at more than one institution, and a private "crosswalk" table at each institution to enable a researcher with all necessary credentials and permissions to re-identify an anonymized patient should it become necessary to contact them – e.g. for consent and follow-up. Next, the means of querying the institutional data marts was prescribed by PCORnet; it would be the ad hoc distributed query engine, PopMedNet [65], chosen because of its historic links to one of the PCORnet coordinating centres. There then followed the most intense debate, in which I played a leading part, on the correct workflow and formulation of a so-called "Master Protocol" which is to be considered a prefix to any other protocol for submission to the central IRB.

The import of the somewhat complex processes of de-identification, de-duplication and *virtual* data integration that the infrastructure performs are described in the penultimate paper in this collection,

J. Anthony Solomonides, Satyender Goel, Denise Hynes, Jonathan C. Silverstein, *et al*. ***Patient-Centered Outcomes Research in Practice: The CAPriCORN Infrastructure***. In *MEDINFO 2015: eHealth-enabled Health* **Studies in Health Technology and Informatics** Vol 216 2015 (Neil Sarkar, Andrew Georgiou, Paulo Mazzoncini de Azevedo Marques, Eds.) pp. 584-588

Chapter 5
Discussion - Engineering a Solution

Blank page

5 Engineering a solution

In this dissertation, I have considered a number of interlinked concepts, propositions and relations, and put forward a set of design theses, to support *the role of informatics* in the overall goal of *knowledge-based*, *information-driven*, *integrated*, *patient-centred*, *collaborative* healthcare and research. This rather ambitious scope may be delimited by exclusion: the work is not concerned explicitly with *genomics* or *bioinformatics*, but it does encompass certain aspects of *translational medicine* and *personalized healthcare*, which I take to be subsumed in some sense under "knowledge-based" and "information-driven". Although I do not exclude *public health informatics*, my exposure extends only to surveillance of infectious diseases, patient engagement, and the effectiveness of screening programmes. I do take *ethical, legal, social and economic issues (ELSE)* to be included, at least to the extent that I aim at an infrastructure that encompasses these issues and aims to incorporate them in technical designs in an effort to meet ethicists', lawyers', policy makers', and economists' concerns halfway. To a first approximation, the aim has been to integrate two strands of work over the last decade or more: the *informatics of medical records* on one hand and the *distributed computational infra-structures for healthcare and biomedical research* on the other.

5.1 Engineering Analogies

Two engineering analogies are sometimes made in the discussion of biomedical and healthcare informatics: the "airline analogy" (AA) and the "bridge analogy" (BA).
In AA, the main question concerns patient safety: *air travel is notoriously safe, so why can't we learn from practices in that industry to make medical and healthcare systems safer?* It is asserted that the airline industry is safe because (a) the design, manufacture and implementation of its systems (aircraft, air traffic control, etc.) follow strict engineering principles and are fully tested before deployment, and (b) in that industry's operations, there is a "no fault" system for reporting human error and near misses. These two elements of safety are evident in the work of several national and international medical informatics associations, such as the European Federation for Medical Informatics (EFMI), the International Medical Informatics Association (IMIA), and the American Medical Informatics Association (AMIA) [66], leading, for example, to work on the evaluation of electronic health record systems (EHR) [67] and to analyses of otherwise unreported systems failures, e.g. in anaesthesia pumps [68]. Thus, the AA focuses on "how", seeking and providing evidence that a system is safe:

- Requirements are established through an iterative process which defines scope, establishes validity (self-consistent, unambiguous, capture what is needed) and adequacy ("complete enough")
- Formal proof is sought that a specification conforms to requirements.
- Formal proof is sought that a system does what is has been specified to do.

In BA, the emphasis shifts from the observation that a bridge can be built on sound civil engineering principles and be certain to be "fit for purpose" – a safe construction that can bear the weight it was designed to and can last for a long time subject to some regular maintenance – to the observation that some unintended and unexpected effects of the bridge may be less than desirable – e.g. change in traffic flows that result in a congested city centre. In healthcare informatics, the BA is reflected in the way physicians' workflows are modified when electronic

systems are introduced. Even where a system is consciously introduced through a "business process re-engineering" programme, changes in workflow may be due to working around a sclerotic technology as much as to a deliberate decision to improve process. The BA focuses on "what" and "why", beginning with the question whether a bridge should be built:

- Why should a bridge be built?
- Where should the bridge be built?
- What impact will it have? (rather than merely Will it be safe?)
- Must those who opposed it refuse to use it? *May* they refuse to use it?

We may distinguish two "pure" paradigms for systems development: in the first, a need is identified and a system commissioned, i.e. development begins with "why" and "what" and moves to "how"; this is more or less the case in much industrial systems development. In the second paradigm, a potential innovation is identified through a technological advance and a service or application is envisioned based on it, i.e. "how" leads to "what" and thence to "why". This can be seen, for example, in the development of mobile communication technologies, from smartphones to the ubiquitous "apps" that run on them. In other fields, these paradigms may be described as "pull" and "push" models. In the majority of cases, of course, the reality of systems development exhibits elements of both these paradigms.

These themes have been explored and illustrated in the work presented here through a number of publications arising from specific projects as well as in a few "position" papers.

5.2 Design Methodology

Two distinct approaches have been adopted in my research, one leading from medical or healthcare issues to technological innovation and a complementary one that begins with technology and moves to healthcare applications. These approaches have been adopted singly and together in different projects. Examples include:

Medicine to Technology The EuroPGD Code project was commissioned by the European Agency for Health and Consumers and the European Society for Human Reproduction and Embryology to support data collection on preimplantation genetic diagnosis (PGD) in Europe. As the principal technical investigator, I worked directly with physicians and researchers to understand their research goals and hypotheses, designed a data collection system and guided its implementation. PGD is a method of screening for specific conditions in the process of assisted reproduction. The project concerned the subsequent developmental health of children born following *in vitro* fertilization (IVF) and PGD.

Technology to Medicine An internally funded project with Dr Kay Wilkinson at the University of the West of England, Bristol, began with the assumption that methods of artificial intelligence could be applied in the analysis of histopathology reports of patients with inflammatory bowel disease (IBD). This was an early example of anticipated value in the analysis of medical records. Using technologies from the design of decision support systems and from natural language processing, it was shown that although reported conclusions in the reports were consistent with findings, the latter were often not sufficient to warrant the conclusion. Further examination of

the reports with the pathologists who had issued them revealed patterns of reporting based on what was considered "relevant" or "obvious" and thus worthy of recording or not.

Technology and Medicine The MammoGrid project was based on the hypothesis that a grid computing approach to the sharing of mammography data in two distinct populations would enhance diagnostic collaboration between physicians and provide a diverse population base for the study of certain physiological hypotheses concerning breast cancer risk. The project developed a highly interactive approach and a common language (based on "use cases") that allowed both the expression of medical requirements to the technologists and a clear presentation of technical possibilities to the physicians.

5.3 Technology and Policy

Healthgrids

The technological strand of my work has arisen from other research in software engineering, notably in the realm of physics and engineering applications. Translating advances made in the context of the new physics experiments at CERN, I was closely identified with the development of "healthgrids", the application of grid computing principles in support of medical research and healthcare. Developed through a number of highly collaborative projects, this work led to several exemplars which eventually informed the study of the principles of this approach in the SHARE project. Earliest among the paradigmatic projects was MammoGrid in which the effectiveness of a distributed infrastructure to support diagnostic practice and epidemiology was demonstrated. This was particularly successful in epidemiological terms in that it provided a means to validate known research findings in breast cancer physiology through the study of two populations, demonstrating the effectiveness of a distributed learning system as well as the applicability of grid principles to healthcare. In the light of this and other successful European projects, the European Commission funded the SHARE project to produce a research road map for healthgrids. This was undertaken first through a study of existing systems and projects, leading to a "state of the art" report; it then followed through with several in-depth domain case studies and, in parallel, a thorough mapping of ethical, legal and socio-organizational issues; finally, the project developed a road map in two stages with a multifaceted expert review to validate the findings and principles between the two stages. The concepts of grid computing have since been incorporated with various business models into what has come to be known as cloud computing. This is commonly spoken of inaccurately as "the cloud", a singular reference that may ultimately only be justified through an abstract framework provided by the infrastructure to support different business models. An analysis of the different paths taken by the grid and cloud paradigms, with particular reference to biomedical research (including pharmaceutical industries) and to healthcare will also be used as a vehicle to explore the dimensions of biomedical and healthcare informatics.

Large-Scale Systems

There is a second interesting perspective on the first question above, (a) *Should a health system maintain a single integrated record for each patient?* In pragmatic terms, there is an extensive history of failures in healthcare systems across the world attempting to adopt or to construct an integrated health record. The present author was a public critic of the *Connecting for Health*

programme in England, whose overreach and poor development process was evident from relatively early on [69]. Yet this was what became, after a number of political twists and turns, of the visionary "Burns Report" *Information for Health* [70] whose process had been rejected as "too slow". More recently, it has been reported in the United States that the Veterans Administration and the Department of Defense have abandoned an attempt to integrate their respective EHR systems after investing several hundred millions of dollars in the attempt. This despite the widely acknowledged quality of the VA's own VistA system which has been so successful[x] that it has been marketed as a product in its own right. Comparing the critique of these two megaprojects by the National Audit Office in the UK and the Government Accountability Office in the US, shows a remarkable overlap in problems identified, criticisms and reasons for failure. The more recent failure of the launch of *HealthCare.gov*, the "Obamacare" website, is symptomatic of a different kind of complexity in the US health care system and will not concern us. However, the complexity that is recognizable in all but the most rudimentary healthcare systems forms the backdrop for an essential thread in the work described here.

5.4 Compliance and Agency

The Institute of Medicine of the National Academies in the United States has issued a number of reports in its Learning Health System Series based on discussions in its Roundtable on Value and Science-Driven Health Care. The report on infrastructure [59] expresses a degree of enthusiasm for the "Ultra Large Scale Systems" [71] approach pioneered by Carnegie Mellon University computer scientists. There are interesting similarities and differences between ULSS and the autonomous "knowledge (health)grid" that was envisaged in the SHARE reports and roadmaps. The critical similarity from the point of view of my own research is in the requirement that the local systems loosely federated in a greater scheme (a) are autonomous, (b) have the capacity to read each other in increasingly sophisticated ways, so that (c) eventually they can interoperate with minimal human intervention. This is an important principle behind my own work on automation of privacy compliance, in which human intervention has been compared with the "red flagging" of early motorcars in the UK because a regulation – intended to manage risk to pedestrians in the movement on public roads of motorized agricultural machinery – was also applied to them as "motorized vehicles". Naturally, in due course, this inappropriate regulation was replaced by a web of legislation and regulatory requirements, including, licensing of cars and drivers, annual checks on vehicle condition, varying speed limits, severe sanctions on driving while unfit, and so on. Accidents certainly occur, but on nothing like the scale that would suggest that human speed is the only appropriate speed for such vehicles[xi]. And here I draw a parallel: whatever other problems the motorcar may have brought in its wake, it has increased mobility, and mobility of data for the purposes of healthcare and healthcare-related research is one of our primary goals.

---

[x] Praised in particular by Professor Denis Protti, adviser to the NHS's *Connecting for Health* programme, at a public lecture at University College London, in 2005, celebrating ten years of *CHIME*, UCL's Centre for Health Informatics & Multiprofessional Education.

[xi] There is another side to this, of course, in the problems brought about by large numbers of motorcars and failure to adopt public modes of transport. This must also be considered (see the discussion of engineering analogies).

So the argument here rests on what a system can achieve given the right degree of autonomy and, therefore, on what the right degree of autonomy is. This belongs to a broader debate on the embodiment of policy in technology and the extent to which policy is "technology-savvy". The study of transport systems, broadly conceived, extending from the engineering of small and large vehicles to entire systems, their management and social impact, provides a useful metaphor in the study of complex information systems. This is highly effectively illustrated in the work of Joseph Sussman and his group at MIT [72]. This is in a sense the landing place for the bridge from the first strand of work on information-driven healthcare.

5.5 Patient Engagement

The work with Hanene Rahmouni was completed in a European context, exploring the varying interpretations of the European Data Protection Directive as national legislation in member states. I have been able to translate, transplant and adapt the argument to the United States context and current research trends there. The adoption of electronic records in the US is somewhat patchy, but there are significant national measures to accelerate the process. More importantly, the rapid growth in the disciplines of genomic and translational medicine, coupled with the prohibitive cost of clinical trials, has led to a number of new approaches to research. Two major trends in this respect are comparative effectiveness research (CER) and patient-centred outcomes research (PCOR). CER is essentially the study of disease development, comorbidities and the relative efficacy of treatments in real patients, and is proposed as a means of formalizing observational studies to complement clinical trials. PCOR has been promoted and heavily favoured in funding mechanisms as a means of involving the patient, at least as represented by patient organizations and patient advocates, in the formulation of research problems, hypotheses and goals that reflect real patient interests. In support of these trends in research, the Institute of Medicine of the National Academies of Science has proposed the concept of a "learning health system", a philosophy and a blueprint for a knowledge-driven healthcare system that applies best evidence-based practice and analyses its own effectiveness to improve every aspect of its operations, from diagnostic and treatment workflows to its approach to population health. My current work in this US context demonstrates the use of integrated informatics to support these goals, from concrete developments, such as new ways to record patient-physician interactions through structured clinical documentation systems (SCDS) [73] to more abstract principles, such as *information reciprocity* between patients, physicians and researchers. The latter proposes and fleshes out a system that adapts technologies from other contexts to support the maintenance of informed consent, patient feedback and education, improvement of social and organizational aspects of healthcare, and expression of unmet research needs as experienced by physicians and patients.

5.6 The Patient in the Learning Health System

The PCORI-funded study on Short- and Long-Term Effects of Antibiotics on Childhood Growth has engaged me as a technical advisor to a stakeholder advisory group consisting of parents, other carers and primary physicians. The interaction with stakeholders has provided first-hand experience of the kind and range of insights that can be brought to bear from a patient perspective. The translation of technical documents for lay stakeholders requires a careful appraisal of the principles on which a study is based and of the research questions it proposes to address. It is sometimes said that the best way to learn something is to teach it; the process of

writing, first, a set of explanatory notes on the analytic plan for the study, and then contributing to a lay pamphlet to describe and explain the project to parents who have no vested interest in the study, unlike the stakeholder group, has underscored both the value of open communication, but also the exceptional value of the committed stakeholders.

The final paper in this collection is a reflection of this experience as well as a summation, in a certain sense, of the diverse elements presented here. It is an affirmation of a commitment to engage patients' and carers', as well as primary physicians', intimate knowledge of many conditions where research has traditionally focused only on medication, on effectiveness by a narrow set of clinical measures, and the patient's wellbeing is at best an afterthought. While I believe this to be important in the provision of healthcare, I am now persuaded that it is an indispensable element in research also. Moreover, as an intelligent patient, I wish to be informed of what is going on in research that may impinge on me for all the good reasons that I may be able to impact the research process itself, that the research team may have the benefit of insights from a radically different point of view, and that the ultimate success of the proposed intervention or innovation may have as much to do with its reception by patients as with the statistical results of the study. To repeat the earlier formulation, rather than doing things *to* research subjects, or *for* philanthropists, we may do research *with* patient collaborators. The main thrust of the paper, however, is not to argue the moral case so much, as to assert that all the necessary elements are present and available.

K. Anthony Solomonides. ***The Learning Patient in the Learning Health System***. (Submitted for review to *MEDINFO 2017*)

Bibliography & References

Blank page

References

[1] *ACP-JC* The Journal Club of the American College of Physicians, supported by a section in the journal Annals of Internal Medicine. http://annals.org/aim/journal-club

[2] *PURLs* The Journal of Family Practice. PURLs: Priority Updates from the Research Literature. http://www.mdedge.com/jfponline/purls. See also University of Chicago & NorthShore Department of Family Medicine for an example of participation. http://familymedicine.bsd.uchicago.edu/Research/PurlsPriorityUpdatesFromTheResearchLiterature

[3] *UpToDate* http://www.uptodate.com/home. Also T Isaac, J Zheng, and A Jha. *Use of UpToDate and Outcomes in US Hospitals*. J. Hosp. Med., 7: 85–90. doi:10.1002/jhm.944

[4] *Isabel* diagnostic tool http://www.isabelhealthcare.com/validation/peer-reviews/impact-studies. See also EJ Henderson and GP Rubin. The utility of an online diagnostic decision support system (Isabel) in general practice: a process evaluation. J R Soc Med Sh Rep 2013;4:31. DOI 10.1177/2042533313476691; J Carlson et al. The Impact of a Diagnostic Reminder System on Student Clinical Reasoning During Simulated Case Studies. J Sim Healthcare 6:11–17, 2011.

[5] *PCORI* Patient-Centered Outcomes Research Institute. http://www.pcori.org (accessed 18/12/16)

[6] *NHS Choices* – Your health and care records http://www.nhs.uk/NHSEngland/thenhs/records/healthrecords/Pages/overview.aspx

[7] Scottish Government Records Management: NHS Code Of Practice (Scotland) Version 2.1 January 2012 http://www.gov.scot/Resource/Doc/366562/0124804.pdf

[8] *Health Information and the Law Fast Facts* Who Owns Health Information? At http://www.healthinfolaw.org/lb/download-document/6640/field_article_file; also interactive map at http://www.healthinfolaw.org/comparative-analysis/who-owns-medical-records-50-state-comparison.

[9] *Federal Register*/Vol. 82, No. 12/Thursday, January 19, 2017/Rules and Regulations Federal Policy for the Protection of Human Subjects https://www.gpo.gov/fdsys/pkg/FR-2017-01-19/pdf/2017-01058.pdf In the healthcare domain this is commonly known as 45 CFR Part 46.

[10] *HealthHeritage* http://www.healthheritage.org/using-health-heritage/

[11] *WiserCare* http://welcome.wisercare.com/how-it-works/

[12] HB Rahmouni, K Munir, M Casassa Mont, T Solomonides. *Semantic Generation of Clouds Privacy Policies*. In: Cloud Computing and Services Sciences Volume 512 of the series Communications in Computer and Information Science pp 15-30 Springer 2015. HB Rahmouni, T Solomonides, M Casassa Mont, S Shiu and M Rahmouni. *A Model-driven Privacy Compliance Decision Support for Medical Data Sharing in Europe*. Methods Inf Med 2011; 50: 326–336; HB Rahmouni, T Solomonides, M Casassa Mont and S Shiu *Privacy Compliance and Enforcement on European Healthgrids: An approach through ontology* Philosophical Transactions of the Royal Society Series A September 13, 2010 368:4057-4072.

[13] *PheKB* What is the Phenotype KnowledgeBase? https://phekb.org See also: Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016 Nov;23(6):1046-1052. doi: 10.1093/jamia/ocv202.

[14] *FARSITE* Sarah Thew, Gary Leeming, John Ainsworth, Martin Gibson, Iain Buchan. FARSITE: evaluation of an automated trial feasibility assessment and recruitment tool. Trials 2011, 12(Suppl 1):A113 http://www.trialsjournal.com/content/12/S1/A113. See also John Ainsworth, Iain Buchan. Preserving consent-for-consent with feasibility- assessment and recruitment in clinical studies: FARSITE architecture. Stud Health Technol Inform 2009, 147:137-148.

[15] J Ainsworth, I Buchan. *Combining Health Data Uses to Ignite Health System Learning*. Methods Inf Med 2015; 54: 479–487; followed by: S Denaxas; CP Friedman; A Geissbuhler; H Hemingway; D Kalra; M Kimura; KA Kuhn; HA Payne; FGB de Quiros; JC Wyatt. *Discussion of "Combining Health Data Uses to Ignite Health System Learning"* Methods Inf Med 2015; 54: 488–499.

[16] *Boot Camp Translation* Ned Norman, Chris Bennett, Shirley Cowart, Maret Felzien, Martha Flores, Rafael Flores, Connie Haynes, Mike Hernandez, Mary Petra Rodriquez, Norah Sanchez, Sergio Sanchez, Kathy Winkelman, Steve Winkelman, Linda Zittleman, and John M. Westfall. Boot Camp Translation: A Method for Building a Community of Solution. J Am Board Fam Med 2013;26:254–263

[17] John M. Westfall, Linda Zittleman, Maret Felzien, Ned Norman, Montelle Tamez, Paige Backlund-Jarquin and Don Nease. Reinventing the Wheel of Medical Evidence: How the Boot Camp Translation Process Is Making Gains. Health Aff April 2016 vol. 35 no. 4 613-618

[18] Wikipedia – *Fat Acceptance Movement*. https://en.wikipedia.org/wiki/Fat_acceptance_movement

[19] *OpenNotes* http://www.opennotes.org/what-is-opennotes-2/why-open-visit-notes/; also patients' stories at http://www.opennotes.org/the-open-patient-documentary/

[20] Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. *Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead*. Ann Intern Med. 2012:157:461-70. Linked editorial: Michael Meltsner. A Patient's View of OpenNotes. Ann Intern Med. 2012:157:523-24.

[21] *Dodd* David Creasey. *An Independent 'Health Information Bank' could solve data security issues*. (Interview with Dr. Bill Dodd) The British Journal of Healthcare Computing and Information Management, Vol 14 No. 7; October 1997.

[22] Marion J. Ball, Melinda Y Costin and Christoph Lehmann. *The Personal Health Record: Consumers Banking on their Health*. Studies in Health Technology and Informatics Vol. 134 eHealth: Combining Health Telematics, Telemedicine, Biomedical Engineering and 35 Bioinformatics to the Edge. Edited by B. Blobel, P. Pharow and M. Nerlich. IOS Press, 2008

[23] Denis J Protti. *The Health Information Bank: Revisiting Bill Dodd's Idea of 10 Years Ago*. ElectronicHealthcare, 6(4) March 2008

[24] Amnon Shabo (Shvo). *It's Time for Health Record Banking! Methods of Information in Medicine* February 2014 [doi: 10.3414/ME13-02-0048]

[25] Patricia Flatley Brennan. *Personal Health Records: Whose Right? Whose Responsibility? Whose Cost?* Presentation at AHIMA National Convention 2007 available at http://healthsystems.engr.wisc.edu/papers_presentations/PHRs_An_Overview.ppt

[26] Rodriguez, Margarita Morales; Casper, Gail; Brennan, Patricia Flatley. *Patient-centered Design: The Potential of User-centered Design in Personal Health Records* Journal of AHIMA 78, no.4 (April 2007): 44-46.

[27] Platt, R., C. Dezii, B. Evans, J. Finkelstein, D. Goldmann, S. Huang, G. Meyer, H. Pierce, V. Roger, L. Savitz, and H. Selker. 2015. *Revisiting the Common Rule and continuous improvement in health care: A learning health system perspective*. National Academy of Medicine, Washington, DC: http://nam.edu/perspectives-2015-revisiting-the- common-rule-and-continuous-improvement-in-health-care-a-learning-health-system- perspective/.

[28] Steven Galson and Gregory Simon. *Real-World Evidence to Guide the Approval and Use of New Treatments* (Perspectives | Expert Voices in Health & Health Care). National Academy of Medicine, 2016.

[29] Harold C. Sox, Roger J. Lewis. Pragmatic Trials: Practical Answers to "Real World" Questions (JAMA Guide to Statistics and Methods). JAMA September 20, 2016 Vol. 316:11.

[30] Ian Ford and John Norrie. Pragmatic Trials (The Changing Face of Clinical Trials). N Engl J Med 2016;375:454-63

[31] Nikolaos A. Patsopoulos. *A pragmatic view on pragmatic trials*. Dialogues in Clinical Neuroscience - Vol 13 No. 2 . 2011

[32] *Learning Health System - Core Values* Web page at http://www.learninghealth.org/corevalues/ (accessed 18/12/16)

[33] Daniel Schwartz and Joseph Lellouch. *Explanatory and Pragmatic Attitudes in Therapeutical Trials*. J. chron. Dis. 1967, Vol. 20, pp. 637-648.

[34] *PRECIS-2* Kirsty Loudon, Shaun Treweek, Frank Sullivan, Peter Donnan, Kevin E Thorpe, Merrick Zwarenstein. *The PRECIS-2 tool: designing trials that are fit for purpose* BMJ 2015;350:h2147 doi: 10.1136/bmj.h2147

[35] JW Song and KC Chung. Observational Studies: Cohort and Case-Control Studies. Plast Reconstr Surg. 2010 December ; 126(6): 2234–2242. doi:10.1097/PRS.0b013e3181f44abc.

[36] *ABX* PCORnet Obesity Observational Study: Short- and Long-term Effects of Antibiotics on Childhood Growth. http://www.pcori.org/research-results/2015/pcornet-obesity-observational-study-short-and-long-term-effects-antibiotics.

[37] Harold Sox. *The Patient-Centered Outcomes Research Institute Should Focus On High-Impact Problems That Can Be Solved Quickly*. Health Affairs 31, No. 10 (2012): 2176–2182 doi: 10.1377/hlthaff.2012.0171

[38] L. Sweeney. *k-anonymity: a model for protecting privacy*. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[39] Tobias A. Knoch, et al. *e-Human Grid Ecology - Understanding and Approaching the Inverse Tragedy of the Commons in the e-Grid Society*. Studies in Health Technology and Informatics Volume 147: Healthgrid Research, Innovation and Business Case (2009) pp. 269-276. doi 10.3233/978-1-60750-027-8-269.

[40] *Condor* Douglas Thain, Todd Tannenbaum, and Miron Livny. *Distributed Computing in Practice: The Condor Experience*. Concurrency and Computation: Practice & Experience – Grid Performance Volume 17 Issue 2-4, Feb 2005 pp. 323-56.

[41] *SETI* Search for Extraterrestrial Intelligence (SETI) (home page: http://setiathome.ssl.berkeley.edu accessed on 10 May 2016)

[42] Foster, I. *Internet Computing and the Emerging Grid*. Nature Web Matters, 2000 (retrieved from http://www.nature.com/nature/webmatters/grid/grid.html on 10 May 2016)

[43] Ian Foster, Carl Kesselman and Steven Tuecke. *The Anatomy of the Grid: Enabling Scalable Virtual Organizations.* International Journal of High Performance Computing Applications Volume 15 Issue 3, August 2001, pp. 200-22.

[44] Foster, I. And Kesselman, C. (eds.). *The Grid: Blueprint for a New Computing Infrastructure.* Morgan Kaufmann, 1999.

[45] S. Roberto Amendolia, Michael Brady, Richard McClatchey, Miguel Mulet-Parada, Mohammed Odeh and Tony Solomonides. *MammoGrid: Large-Scale Distributed Mammogram Analysis* in *The New Navigators: from Professionals to Patients* (Proceedings of Medical Informatics Europe 2003) Robert Baud, Marius Fieschi, Pierre Le Beux, Patrick Ruch (Eds.). Studies in Health Technology and Informatics, Vol 95 (2003) IOS Press

[46] R. Highnam and J.M. Brady. *Mammographic Image Analysis* (Computational Imaging and Vision, Volume 5) Springer (1999)

[47] ESHRE PDG Consortium Steering Committee. *ESHRE Preimplantation Genetic Diagnosis (PGD) Consortium: preliminary assessment of data from January 1997 to September 1998.*Hum Reprod (1999) 14 (12): 3138-3148. doi.org/10.1093/humrep/14.12.3138

[48] Breton V, Solomonides AE, McClatchey RH: A perspective on the Healthgrid initiative. Second International Workshop on Biomedical Computations on the Grid, at the 4th IEEE/ACM International Symposium on Cluster Computing and the Grid; Chicago 2004. (Accessed 29 August 16 at: http://arxiv.org/abs/cs/0402025)

[49] Vincent Breton, Kevin Dean, Tony Solomonides, *et al*. *The Healthgrid White Paper*. Studies in Health Technology and Informatics, Volume 112: From Grid to Healthgrid (2005) pp. 249 – 321.

[50] Andrew C. von Eschenbach and Kenneth Buetow. *Cancer Informatics Vision: caBIG™*. Cancer Informatics 2006:2 22–23

[51] Clinical Data Interchange Standards Consortium (CDISC) *Where is caBIG Going?* Posted 9 July 2012 at https://www.cdisc.org/where-cabig-going%3F

[52] Catherine J Staes, Wu Xu, Samuel D LeFevre, *et al*. *A case for using grid architecture for state public health informatics: the Utah perspective.* BMC Medical Informatics and Decision Making 2009, **9**:32 doi:10.1186/1472-6947-9-32

[53] Wim Kok, *Enlarging the European Union Achievements and Challenges.* European University Institute (2003) http://cadmus.eui.eu/bitstream/handle/1814/2515/200303KokReport_EN.pdf

[54] F Martin-Sanchez, I Iakovidis, S Nørager, *et al*. *Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care*. Journal of Biomedical Informatics 37:1(2004) pp. 30-42 http://dx.doi.org/10.1016/j.jbi.2003.09.003

[55] *SHARE* EU Project ID: 027694. Funded under: FP6-IST. Supporting and structuring HealthGrid activities & research in Europe: developing a roadmap. http://cordis.europa.eu/project/rcn/79499_en.html; see also http://www1.uwe.ac.uk/et/research/cccs/projects/share.aspx

[56] M Cannataro (Ed.) *Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare* (2 Volumes) IGI Global 2009

[57] Ian Foster's blog, 8th January 2008. *There's Grid in them thar Clouds*. http://ianfoster.typepad.com/blog/2008/01/theres-grid-in.html

[58] Lamia Youseff, Maria Butrico, and Dilma Da Silva. *Toward a Unified Ontology of Cloud Computing*. Proceedings 2008 Grid Computing Environments Workshop, IEEE 2008. doi:10.1109/GCE.2008.4738443 at http://ieeexplore.ieee.org/document/4738443/

[59] *Institute of Medicine* (now the National Academy of Medicine) Roundtable on Evidence-Based Medicine. *The Learning Healthcare System: Workshop Summary*. National Academies Press (2007) https://www.nap.edu/download/11903

[60] *Institute of Medicine* (now the National Academy of Medicine) Roundtable on Value & Science-Driven Health Care. *Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary.* National Academies Press (2011) https://www.nap.edu/download/12912

[61] Tony Solomonides. *Healthgrids, the Share Project, and Beyond*. In [59], pp. 202-210.

[62] The North Carolina Healthcare Information & Communications Alliance, Inc. (NCHICA) Workshop 27 April 2015 at http://nchica.org/wp-content/uploads/2015/02/Friedman.pdf (accessed 29 August 2016).

[63] *HIPAA* US Dept. of Health and Human Services. *Health Insurance Portability and Accountability Act* (1996), Public Law 104-191. https://www.hhs.gov/hipaa/for-professionals/index.html

[64] *ABX* PCORnet Obesity Observational Study: Short- and Long-term Effects of Antibiotics on Childhood Growth http://www.pcori.org/research-results/2015/pcornet-obesity-observational-study-short-and-long-term-effects-antibiotics

[65] *PopMedNet* http://www.popmednet.org

[66] *EFMI* https://www.efmi.org; *IMIA* http://imia-medinfo.org/wp/; *AMIA* https://www.amia.org

[67] Jan Talmon, Elske Ammenwerth, Jytte Brender, Nicolette de Keizer, Pirkko Nykänen, and Michael Rigby. *STARE-HI—Statement on reporting of evaluation studies in Health Informatics*. IJMI 78 (2009) pp 1-9.

[68] Paul T Lee, Frankie Thompson and Harold Thimbleby. *Analysis of infusion pump error logs and their significance for health care*. British Journal of Nursing (Intravenous Supplement), 21:8 (2012)

[69] Brian Randell. *A computer scientist's reactions to NPfIT*. Journal of Information Technology (2007) 22, 222–234. See also: Helen Boddy. *The NHS23 still favour an independent review*. BCS Health Informatics Now, March 2008, pp. 21-22. Available at http://www.bcs.org/upload/pdf/hinow-mar08.pdf, and *NHS23* http://editthis.info/nhs_it_info/Main_Page.

[70] NHS Executive Information for Health (the "Burns Report") http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4014469.pdf

[71] Peter H. Feiler, Kevin Sullivan (University of Virginia), Kurt C. Wallnau, Richard P. Gabriel (Sun Microsystems), John B. Goodenough, Richard C. Linger (Oak Ridge National Laboratory), Thomas A. Longstaff, Rick Kazman, Mark H. Klein, Linda M. Northrop, Douglas Schmidt (Vanderbilt University). *Ultra-Large-Scale Systems: The Software Challenge of the Future*. CMU SEI (2006). http://resources.sei.cmu.edu/asset_files/Book/2006_014_001_30542.pdf

[72] Carlos A. Osorio, Dov Dori, and Joseph Sussman. *COIM: An Object-Process Based Method for Analyzing Architectures of Complex, Interconnected, Large-Scale Socio-Technical Systems*. Systems Engineering Vol. 14, No. 4, 2011.

[73] Goldberg RP, Gafni-Kane A, Jirschele K, Silver R, Maurer D, Solomonides T, Simmons A, Silverstein J. *An automatic female pelvic medicine and reconstructive surgery registry and complications manager developed in an electronic medical record.* **Female Pelvic Med Reconstr Surg**. 2014 Nov-Dec; 20(6):302-4.

[74] Pablo Saiz *et al. AliEn Resource Broker.* Proceedings of the 10th Int. Conf. on Computing for High Energy Physics (CHEP'03) San Diego, USA. March 2003

[75] A. Redolfi *et al. Grid infrastructures for computational neuroscience: the neuGRID example,* Future Neurology, Volume 4, Number 6, 2009, pp. 703-722.

Blank page

Appendices

Blank page

APPENDIX A

Research Students

In chronological order, my four PhD students with a Healthcare or Medical Informatics theme were Kay Wilkinson, Hanene Rahmouni, Sotiris Fanou, and Mark Olive. Two names in particular, Hanene's and Mark's, appear on some papers. Relative contributions are described here.

Kay Wilkinson MB BS PhD (2006)

Kay's project was to explore the extent to which Inflammatory Bowel Disease patients' charts could be mined for knowledge, in particular for the differentiation of Crohn's Disease from Ulcerative Colitis. The project was co-supervised by Dr. Alastair Forbes, at that time a Consultant Gastroenterologist at St. Mark's Hospital, Northwick Park, now a Professor at Norwich Medical School, University of East Anglia. The only data available to us in electronic form were histopathology reports. The project examined pathologists' inter- and intra-consistency when reviewing old reports with the conclusions removed. It translated the British Society of Gastroenterology's then-current guidelines on the interpretation of histopathology reports into a decision tree and used natural language processing and abductive logic to identify information that should have been recorded in the findings for the conclusion to be justified. Interviews with pathologists confirmed the missing information with the view that "it was too obvious to be recorded". This has become a familiar theme as I have worked through many more projects: what an informatician may consider necessary for completeness and what a practising physician requires for effective healthcare provision are two very different things.

Hanene Rahmouni BSc PhD (2010)

The SHARE project had identified a number of security, privacy and confidentiality issues in the possible use of healthgrids for research and, even more so, in healthcare practice. The problem formally posed to Hanene was how to translate the European Data Protection directive into operational terms so that for the majority of straightforward cases, data sharing and transmission need not be held up by the need for a human expert to review. In fact, the European directive had been variously translated into national law in the member states, so the first problem was to determine how to represent that. Hanene adopted an ontology-based approach which worked very effectively. The problem then became how to translate the declarative Description Logic (DL) of the legislation to a deontic logic of permissions and obligations. This was accomplished through, first, factoring the sharing process into three steps: preconditions to be satisfied, permitted action, and post-conditions to be imposed on the parties involved, and second, by mapping the logic from DL to rules in the eXtensible Access Control Markup Language (XACML).

Sotiris Fanou BA MSc PhD (2012)

Sotiris's PhD was based in the School of Health and Social Care, with my co-supervision of technical and informatics aspects. The project addressed the question: can individuals with limited intellectual abilities ("people with learning difficulties" (PLD) in their own preferred terms) be helped to create a simple wiki to support their work? In this specific case, these PLD had been singled out for their reasoning and communication abilities to serve as basic health trainers for other PLD who lived in their communities. Thus, the project was essentially to extend the principles of end-user computing and co-creation of useful artefacts to the case where the user/co-creators had limited intellectual abilities. Once they had designed the wiki, the health trainers were to use it to share materials to support each other in their work. The project was a limited success, with only a relatively few PLD health trainers contributing either to the functionality of the wiki or to materials on it.

Mark Olive BA MPhil (2017)

Mark's project also stemmed from SHARE, and in a sense, from the earlier project MammoGrid. Here the question at issue was the use, or usability, of evidence from practice: the project was dubbed "practice-based evidence for evidence-based practice" (PBE4EBP). In the MammoGrid project, mammograms were digitized and standardized using a proprietary method known as Standard Mammogram Form™ which (at least in theory) enabled radiologists to compare mammograms with each other as if they had been taken on the same machine using the same settings. An additional facility in the commercial product was "Find One Like It", i.e. a search of the image database and of the associated patient data to produce the best matches for a patient's series of images and history. This would to support case-based prognosis, based on everything that was known about the patient at hand. The critical point here was that the evidence would not have been validated through a rigorously controlled study, but would in effect be an open-ended observational trial, against a background of changing conditions and practices in radiology. Seeking a more controllable problem to explore the possibility of evidence from practice, I suggested that Mark should study variance reporting in integrated care pathways. Following relapse of a serious illness, Mark has continued to a more limited project along these lines for the degree of MPhil.

Other research students

I also served as Director of Studies for **Computer Science** students:

Piotr Stanczyk – Image assessment using multi-fractal image processing techniques. Collaboration with *British Steel*.

Steve Jenkins – Questionnaire and survey editing software; commercial product *SnapSurveys*.

Peter Hale – End-user computing in engineering design; collaboration with *British Aerospace*.

APPENDIX B

Principal Publications

These papers are part of the formal submission for the degree of Doctor of Philosophy (DPhil).

A. F Estrella, C del Frate, T Hauer, R McClatchey, M Odeh, D Rogulin, S R Amendolia, D Schottlander, **T Solomonides**, R Warren. *Resolving Clinicians' Queries Across a Grids Infrastructure* **Methods of Information in Medicine** Vol **44** No 2. 2005 pp 149-153. ISSN 0026-1270 Schattauer publishers.

B. R Warren, **T Solomonides**, C del Frate, I Warsi, J Ding, M Odeh, R McClatchey, C Tromans, M Brady, R Highnam, M. Cordell, F Estrella & R Amendolia. *MammoGrid – A Prototype Distributed Mammographic Database for Europe* **Clinical Radiology** Vol **62** No 11 pp 1044-1051. DOI 10.1016/j.crad.2006.09.032 November 2007, Elsevier publishers.

L. Estrella F, Hauer T, McClatchey R, Odeh M, Rogulin D, **Solomonides T**. *Experiences of engineering Grid-based medical software*. **Int J Med Inform.** 2007 Aug; **76(8)**:621-32.

M. Olive, M., Lashwood, A. and **Solomonides, T**. (2011). *A retrospective study of paediatric health and development following pre-implantation genetic diagnosis and screening*. In: Olive, M. and Solomonides, T. (ed.) **IEEE Proceedings of the 2011 24th International Symposium on Computer-Based Medical Systems** (CBMS 2011), p. 32-38.

N. V. Breton, K. Dean, **T. Solomonides**, *The HealthGrid White Paper*, in *From Grid to Healthgrid*. **Studies in Health Technology and Informatics**, vol. 112, ISBN 1-58603-510-X, ISSN 0926-9630 IOS Press.

O. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez *SHARE road map for HealthGrids: Methodology* **International Journal of Medical Informatics**, 78S (2009) S3–S12

P. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez. *SHARE, from Vision to Road Map: Technical Steps.* **Studies in Health Technology and Informatics** Volume 129: ***Building Sustainable Health Systems*** -

Proceedings of the 12th World Congress on Health Informatics – MedInfo 2007. IOS Press 2007

Q. Mark Olive, Hanene Boussi Rahmouni and **Tony Solomonides** (UWE, Bristol, UK); Vincent Breton, Nicolas Jacq and Yannick Legré (IN2P3, CNRS, Clermont-Ferrand, France & HealthGrid, EU); Ignacio Blanquer and Vicente Hernandez (Universidad Politécnica de Valencia, Spain); Isabelle Andoulsi and Jean Herveg (Universitaires Notre-Dame de la Paix, Belgium); Celine Van Doosselaere and Petra Wilson (European Health Management Association, EU); Alexander Dobrev, Karl Stroetmann and Veli Stroetmann (Empirica GmbH, Germany). *SHARE: A European Healthgrid Roadmap* in **Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare** (Mario Cannataro, Ed). Chapter 1, pp. 1–27. IGI-Global, Hershey, PA. 2009.

R. A E Solomonides. *Compliance and Creativity in Grid Computing.* 16th World Congress on Medical Law, Bioethics Track. Toulouse 2006.

S. Anthony Solomonides, Satyender Goel, Denise Hynes, Jonathan C. Silverstein, *et al*. ***Patient-Centered Outcomes Research in Practice: The CAPriCORN Infrastructure.*** In *MEDINFO 2015: eHealth-enabled Health* **Studies in Health Technology and Informatics** Vol 216 2015 (Neil Sarkar, Andrew Georgiou, Paulo Mazzoncini de Azevedo Marques, Eds.) pp. 584-588

T. Anthony Solomonides. ***The Learning Patient in the Learning Health System.*** (Submitted for review to *MEDINFO 2017*)

APPENDIX C

Subsidiary Publications

a. S. Roberto Amendolia, Michael Brady, Richard McClatchey, Miguel Mulet-Parada, Mohammed Odeh and **Tony Solomonides**. *MammoGrid: Large-Scale Distributed Mammogram Analysis* in *The New Navigators: from Professionals to Patients* (Proceedings of Medical Informatics Europe 2003) Robert Baud, Marius Fieschi, Pierre Le Beux, Patrick Ruch (Eds.). **Studies in Health Technology and Informatics**, Vol 95 (2003) IOS Press

This was a high-level description of the MammoGrid project as proposed for funding. I made a substantial contribution both in writing the proposal and in the writing of the paper. I was chosen to present the paper at the MIE2003 conference.

b. R Warren, D Thompson, C del Frate, R Highnam, C Tromans, I Warsi, J Ding, F Estrella, **T Solomonides**, M Odeh, R McClatchey, M. Bazzocchi, S R Amendolia & M Brady. *A Comparison of Some Anthropometric Parameters Between an Italian and a UK Population : "Proof of Principle" of a European Project using MammoGrid.* **Clinical Radiology** Vol **62** No 11 pp 1052-1060. DOI 10.1016/j.crad.2007.04.002 November 2007, Elsevier publishers.

This paper is linked to the main paper B, and appears immediately after B in the journal. This represents the main scientific contribution of the project. The journal would not publish it until a suitable description of the technology was in the public domain and could be referenced. Although the journal was initially reluctant to publish a technical paper, it eventually accepted paper B above to be published alongside this one.

c. Tamás Hauer, Dmitry Rogulin, Sonja Zillner, Andrew Branson, Jetendr Shamdasani, Alexey Tsymbal, Martin Huber, **Tony Solomonides**, Richard McClatchey. *An Architecture for Semantic Navigation and Reasoning with Patient Data - Experiences of the Health-e-Child Project.* in The Semantic Web – Proceedings of the Seventh International Semantic Web Conference (ISWC 2008). **Lecture Notes in Computer Science** Vol. 5318; pp. 737-750. Springer 2008.

Dr. Hauer wrote this paper with my collaboration. I particularly contributed to the precise description of user requests for information, such as formal descriptions of the results of cascades of refinements to a query.

d. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez, Isabelle Andoulsi, Jean Herveg, Petra Wilson. *SHARE Roadmap 1: Towards a debate.* **Studies in Health Technology and Informatics** Volume 126: ***From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences*** - Proceedings of HealthGrid 2007. IOS Press 2007

e. Mark Olive, Hanene Rahmouni, **Tony Solomonides**, Vincent Breton, Yannick Legré, Ignacio Blanquer, Vicente Hernandez, Isabelle Andoulsi, Jean Herveg, Petra Wilson. *SHARE Roadmap 1: Towards a debate.* **Studies in Health Technology and Informatics** Volume 126: ***From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences*** - Proceedings of HealthGrid 2007 pp.164–173. IOS Press 2007

The SHARE road map in publication H. above was developed in two stages. The preliminary SHARE Roadmap 1 was published at the end of the first year of the project and summarized for presentation as a peer reviewed paper at HealthGrid 2007. (It was also presented without a corresponding publication at a number of grid computing and healthcare management events for public criticism.) Order of authorship is again by institution, with the most senior member of the institutional writing team last within its group. I was the principal author and presenter of this paper.

f. Penny Duquenoy, Carlisle George, **Anthony Solomonides**. *Considering something 'ELSE': Ethical, legal and socio-economic factors in medical imaging and medical informatics*. **Computer Methods and Programs in Biomedicine: Medical Imaging and Medical Informatics (MIMI)** Vol 92:3, 2008, pp. 227–237.

This paper was the result of a continuing collaboration on the theme of ethics in healthcare and biomedical computing. Work was at various stages presented at workshops at Middlesex University and finally at the Second International Conference on Medical Imaging and Medical Informatics, Beijing, 2007. An expanded paper was selected for inclusion and was subsequently re-reviewed and published.

g. **Tony Solomonides** *Healthgrids, the SHARE project, medical data and agents: retrospect and prospect* in **Lecture Notes in Artificial Intelligence (LNCS 7057)** ***Principles and Practice of Multi-Agent Systems*** Revised Selected Papers from the 13th international conference on Principles and Practice of Multi-Agent Systems (PRIMA 2010) pp. 523-534 Springer-Verlag Berlin, Heidelberg 2012

This was a peer-reviewed paper based on an invited presentation at PRIMA 2010. Its principal interest is in addressing the possible use of agent technologies as a means of automating certain aspects of data distribution and management.

h. Abel N Kho, Denise M Hynes, Satyender Goel, **Anthony E Solomonides**, Ron Price, Bala Hota, Shannon A Sims, Neil Bahroos, Francisco Angulo, William E Trick, Elizabeth Tarlov, Fred D Rachman, Andrew Hamilton, Erin O Kaleba, Sameer Badlani, Samuel L Volchenboum, Jonathan C Silverstein, Jonathan N Tobin, Michael A Schwartz, David Levine, John B Wong, Richard H Kennedy, Jerry A Krishnan, David O Meltzer, John M Collins, Terry Mazany. *CAPriCORN: Chicago Area Patient-Centered Outcomes Research Network* **J Am Med Inform Assoc** 2014;21:607–611.

This paper was published immediately after the award of the grant by PCORI to the CAPriCORN consortium. It is essentially a description of the project plan and the partial progress that had been achieved already in anticipation of the grant. Abel Kho drafted the paper based on the proposal document, to which Solomonides had contributed extensively, both in preliminary analysis of data in different systems, in designing a common data model, and in writing several sections. Solomonides provided review and minor edits in the final version prior to submission.

i. Goldberg RP, Gafni-Kane A, Jirschele K, Silver R, Maurer D, **Solomonides T**, Simmons A, Silverstein J. *An automatic female pelvic medicine and reconstructive surgery registry and complications manager developed in an electronic medical record.* **Female Pelvic Med Reconstr Surg**. 2014 Nov-Dec; 20(6):302-4.

This paper is based on fundamental work on structured clinical documentation carried out by Silverstein and Solomonides in Research, implemented by Simmons and Maurer in Health IT, to serve Drs. Goldberg, Gafni-Kane, Jirschele and Silver in the Department of Obstetrics and Gynecology. The paper was written by Dr. Gafni-Kane with editorial changes by Solomonides and Silverstein.

Blank page

ANNEX
The Principal Publications

Blank page

Resolving Clinicians Queries Across a Grids Infrastructure

Florida Estrella[1], Chiara del Frate[2], Tamas Hauer[1], Richard McClatchey[1], Mohammed Odeh[1], Dmitry Rogulin[1], Salvator Roberto Amendolia[3], David Schottlander[4], Tony Solomonides[1], Ruth Warren[5]

[1]CCCS Research Centre, University of the West of England, Frenchay, Bristol BS16 1QY, UK
[2]Istituto di Radiologia, Università di Udine, Italy
[3]ETT Division, CERN, 1211 Geneva 23, Switzerland
[4]Mirada Solutions Limited, Mill Street, Oxford, OX2 0JX, UK
[5]Breast Care Unit, Addensbrooke Hospital, Cambridge, UK

Corresponding author: Professor Richard H McClatchey, CCCS Director, University of the West of England, Coldharbour Lane, Frenchay, Bristol BS16 1QY, UK
Email: richard.mcclatchey@cern.ch

Abstract: The past decade has witnessed order of magnitude increases in computing power, data storage capacity and network speed, giving birth to applications which may handle large data volumes of increased complexity, distributed over the internet. Medical image analysis is one of the areas for which this unique opportunity likely brings revolutionary advances both for the scientist's research study and the clinician's everyday work. Grids [1] computing promises to resolve many of the difficulties in facilitating medical image analysis to allow radiologists to collaborate without having to co-locate. The EU-funded MammoGrid project [2] aims to investigate the feasibility of developing a Grid-enabled European database of mammograms and provide an information infrastructure which federates multiple mammogram databases. This will enable clinicians to develop new common, collaborative and co-operative approaches to the analysis of mammographic data. This paper focuses on one of the key requirements for large-scale distributed mammogram analysis: resolving queries across a grid-connected federation of images.

Keywords: distributed database, queries, meta-data, mammography, medical image analysis, epidemiological studies

---

1. Introduction

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, SPEC etc. This trend will increase as high bandwidth PACS systems are installed in large numbers of hospitals (currently, primarily in large teaching hospitals).

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 (3-D images, time sequences, multiple imaging protocols). In most cases, no single imaging modality suffices, since clinically significant signs are subtle and because there are many parameters that affect the appearance of an image, like:

- Patient age, diet, lifestyle, clinical history, …
- Image acquisition parameters
- Anatomical and physiological variations.

Breast cancer as a medical condition, and mammograms as images, are extremely complex with many dimensions of variability across the population. Similarly, the way diagnostic systems are used and maintained by clinicians varies between imaging centres and breast screening programmes, and in consequence so does the appearance of the mammograms generated. It is necessary to understand this variability to be able to study the epidemiology of breast cancer and enhance the usefulness of mammography breast screening by integrating Computer Aided Diagnostic tools [3] and quality control [4], [5] in the process. A geographically distributed database that reflects the spread of pathologies across the European population is an essential tool for the epidemiologist and the understanding of the variation in image acquisition protocols is invaluable to the end-user who runs a screening programme.

In order to make the most of such a database it is necessary to have the right tools. This requires an infrastructure to make the large volume of data available to all the centres in an acceptable time, a capable data-mining engine that enables queries based on patient details and text annotations, standardization software to enable the comparison of images from different patients and centres, image analysis algorithms that provide quantitative information, which is otherwise unavailable from visual inspection alone, and detection systems that help in visual diagnosis.

Usually, related personal and clinical information is important (age, gender, selection criteria, disease status). 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 exemplars in response to any given query (e.g. "show me all women in their 50s that developed a tumor within 5 years of starting HRT") that is statistically significant. Overcoming this problem implies constructing a huge, multi-centre – federated – database, while overcoming statistical biases such as lifestyle and diet leads to a database that transcends national boundaries. For ***any*** medical condition, there are potential gains from a pan-national database – so long as that (federated) database is as usable as if it were installed in a single site.

2. MammoGrid User Requirements

The main output of the MammoGrid project, a Grid-enabled software platform (called the MammoGrid Information Infrastructure) which federates multiple mammogram databases, will enable clinicians to develop new common, collaborative approaches to the analysis of mammograms. This will be achieved through the use of Grid-compliant services for managing

massively distributed files of mammograms, for handling the distributed execution of mammogram analysis software, for the development of Grid-aware algorithms and for the sharing of resources between multiple collaborating medical centres. All this is delivered via a novel software and hardware information infrastructure that guarantees the integrity and security of the medical data.

Actor:
Mammogram Analyst
Pre-Conditions
User-Authentication
Non-functional requirements:
CADe Software Interface Requirements
Flow of Events:
as per selection of the Mammogram Analyst to link to the appropriate extension point below
Extension Points:
(1) View Mammogram and Patient Details
(2) Annotate Mammograms and Patient Details
(2.1) Diagnose Study
(2.2) Diagnose Series
(2.3) Annotate Image
(2.4) Request CADe
(2.5) Link Annotations
(2.6) Request CADe in Mammogram Region
(3) Execute Radiological Queries
(3.1) Formulate Radiological Query
(3.2) Refine Radiological Query
Alternative Flows:
(1) Unsuccessful User Authentication
(2) CADe Interface Error
(3) Invalid Query Selections
Post-Conditions
(1) Mammogram Image Annotated
(2) Patient Details Changed
(3) Results of Query Execution (Grid)
Use-Case View:

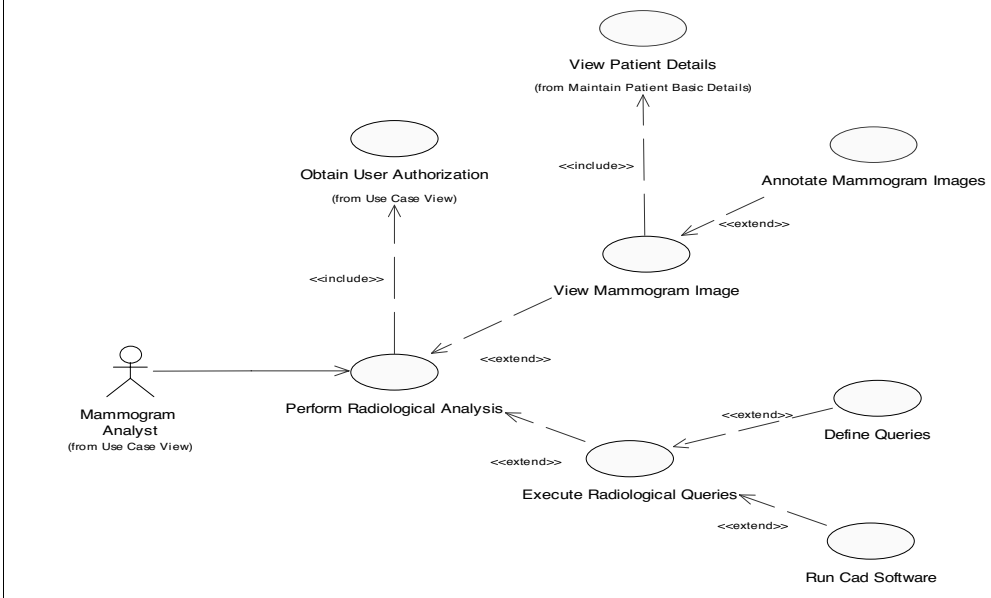


Figure 1: The Mammogram Analyst Use Case diagram

The MammoGrid project is being driven by the requirements of its user community (represented by Udine (Italy) and Cambridge (UK) hospitals along with medical imaging expertise from Oxford) that have been elicited and specified in detail [6]. Rational's Unified Process model (RUP) [7] has been used in the requirements specification for MammoGrid and in particular key requirements engineering activities. The process has identified major use-case scenarios in the use of a distributed database of mammograms deployed across a pan-European Grid and that later can be used to prove the MammoGrid prototype.

The resulting MammoGrid User Requirements Specification (URS) details two essential objectives that must be supported and tested in the MammoGrid project:

- Support of clinical research studies through access to and execution of algorithms on physically large, geographically distributed and potentially heterogeneous sets of (files of) mammographic images, just as if these images were locally resident.
- Controlled and assured access for educational/commercial companies to distributed mammograms for testing novel medical imaging diagnostic technologies in scientifically acceptable clinical trials that fulfill the criteria of evidence-based medical research.

The choice of a use-case driven approach to requirements gathering seemed most practical as the gap between the developers' background knowledge of the domain and that of the users needed to be bridged in order to tackle the problem of architectural and interaction design. The requirements elicitation process was carried out in consultation with the user community at hospitals in Udine, Cambridge and Torino. In these discussions nine core use cases with corresponding actors have been identified. The use case concerned with mammographic analysis describes the tasks that the 'Mammogram Analyst' actor, normally a radiologist or perhaps an epidemiologist, may undertake to annotate and/or view mammograms and patient details, to execute radiological queries, including use of computer aided detection (CADe) software, and to execute epidemiological queries. This use-case provides the frame for the queries which is the main subject of the further sections of this paper. A brief extract of this use case is shown in figure 1.

3. Resolving Clinicians' Queries

In the MammoGrid proof-of-concept demonstrator, real clinician queries will be handled and resolved against data resident across a Grids infrastructure. User Requirements have been gathered that will enable queries to be executed and data retrieved for the analysis of mammograms. In particular the MammoGrid project will test the access to sets of mammogram images for the purposes of breast density assessment and for the testing of CADe studies of mammograms.

Queries can be categorized into simple and complex queries. Simple queries use predicates that refer to simple attributes of meta-data saved alongside the mammographic images. One example of a simple query might be to 'find all mammograms for women aged between 50 and 55' or 'find all mammograms for all women over 50 undergoing HRT treatment'. Provided that age and HRT related data is stored for (at least a subset of) patients in the patient meta-data then it is relatively simple to select the candidate images from the complete set of images either in one location of across multiple locations. It is also possible to collect data concerning availability of requested items so as to inform the design of future protocols, thus engineering a built-in enhancement process.

There are, however, queries which refer to data that has not been stored as simple attributes in the meta-data but rather require derived data to be interrogated or an algorithm to be executed. Examples of these might be queries that refer to the semi-structured data stored with the images through annotation or clinician diagnosis or that is returned by, for example, the execution of the CADe image algorithms.

3.1. Typical Complex MammoGrid Queries

This section describes three example use-cases illustrating the nature of complex queries which the MammoGrid infrastructure should handle.

3.1.1. Use Case 1: Patient's first visit

Consider a patient on her first visit to the mammography center (following referral by GP, worried about a symptom). The typical workflow of the visit looks like this:

- Mammograms (2 ´ MLO and 2 ´ CC[1]) taken
- Radiologist reads them and annotates[2] left MLO (LMLO) and left CC (LCC)
- Radiologist requests CADe for LMLO and LCC images.
- *Query:* Radiologist requests 'find similar cases'. Example criteria might include women:
  - of same age ± 3
  - with same number of children (0), (1-2), (3-4), (5+)
  - with same age ranges of children (equivalently, age at first and last pregnancy)
  - with images that the algorithm "find one like it" matches well either in MLO or CC
- Radiologist reviews demographics and personal data and determines best four cases to request images.
- Radiologist reviews comparable images with histories and analyses:
  - consider the best match
  - take images from first diagnosis to current state
  - review growth of lesion (ideally identifies the lesion across images)

3.1.2. Use Case 2: Epidemiology Study

Consider an epidemiologist who is conducting a study on contralateral breast cancer. The typical queries she is interested in running may include:

- Find all patients in the distributed database who have developed cancer in the other breast after successful therapy (specific or otherwise) on the first cancer.
- Consider mammographic features from the time of first diagnosis and any correlation to occurrence of contralateral cancer.
- Consider measures of asymmetry and their correlation to contralateral cancer.

3.1.3. Use Case 3: Quality Control of Radiology Diagnosis

Consider the use case of comparative study of radiologists' annotations. The typical queries which can be used to survey radiologists' diagnostic processes include the following example queries:

- For a period of six months, allocate each patient who attends for screening at random to two out of three radiologists so that all three possible pairs get roughly equal numbers.
- For each patient, ask both radiologists to examine the mammograms and to make any necessary annotations.
- Submit all annotations for CADe and measure differences between radiologists' annotations and CADe (could be area if masses, counts if microcalcifications) and between the two radiologists in each case.

---

[1] MLO – Medio-Lateral Oblique, taken at 45° from shoulder to opposite hip; CC – Cranio-Caudal, taken vertically down from above.

[2] Annotation – a region is marked out as suspect or for further analysis.

- Consider correlation to experience, the length of the viewing session and the serial order of the given image in that session, and the radiologist's perception whether this was the first or second reading.

3.2. The Role of Meta-Data

During the final phase of implementation and testing, lasting until the completion of the project, the meta-data structures required to resolve the clinicians' queries will be delivered using the meta-modelling concepts of the CRISTAL project [8]. This will involve customizing a set of structures that will describe mammograms, their related medical annotations and the queries that can be issued against these data. The meta-data structures will be stored in a database at each node in the MammoGrid (e.g. at each hospital or medical centre) and will provide information on the content and usage of (sets of) mammograms.

The query handling tool will locally capture the elements of a clinician's query and will issue a query, using appropriate Grids software, against the meta-data structures held in the distributed hospitals. At each location the queries will be resolved against the meta-data and the constituent sub-queries will be remotely executed against the mammogram databases. The selected set of matching mammograms will then be either analyzed remotely or will be replicated back to the centre at which the clinician issued the query for subsequent local analysis, depending on the philosophy adopted in the underlying Grids software. All data objects will reside in standard commercial databases, which will also hold descriptions of the data items.

3.3. The Query Handler

The user will submit queries that are serviced locally and farmed out to available resources when data from the network is required. In resolving queries the system will consult the knowledge it has acquired from previous queries. Data will be immediately returned to the user and the knowledge base updated. New data is processed only when necessary. With this approach the computation required by a domain-specific application is analyzed and farmed out to appropriate data sources rather than moving or replicating potentially vast amounts of data and processing.

The querying software largely constitutes:

Query Manager components
- Query Translator,
- Query Analyser,
- Local Query Handler,
- Remote Query Handler and
- Result Handler

Data Sources
- User's Terms and Mammogrid specific meta-data,
- Local database and
- Stored query database

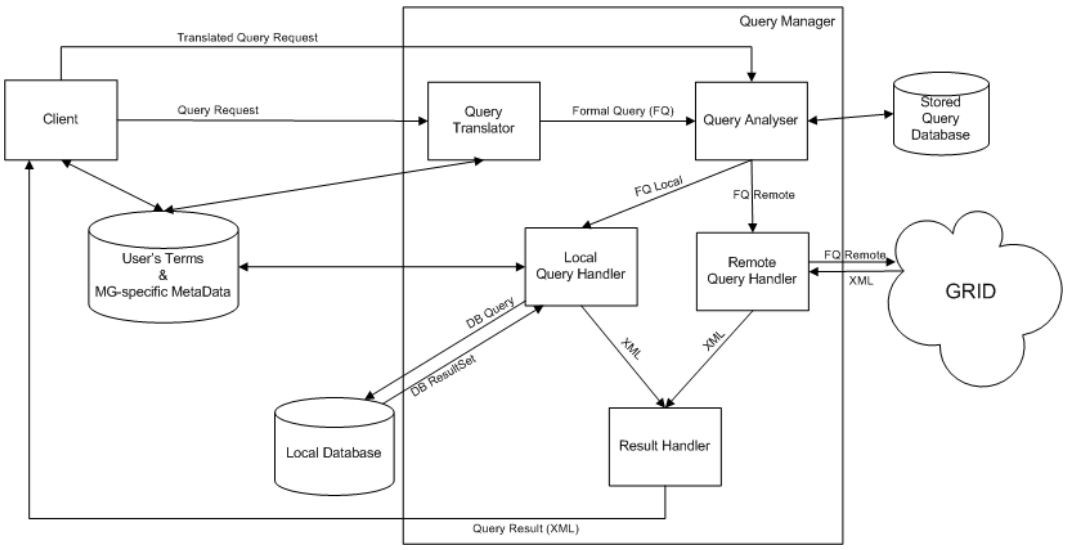


Figure 2: Query Handling in Mammogrid

Figure 2 illustrates the query handling and execution in Mammogrid. The sequence of events is as follows:

(1) Clients (e.g. end-users, applications) define their mammogram analysis in terms of queries they wish to be resolved across the collection of data repositories (either locally- or remotely-held data). This uses descriptive information (User's Terms and MG-specific metadata) about the query domain (both graphical specifications and user-specific terms) to translate the user query into a 'data request' using standard terms.

(2) Query Translator takes the user request and translates to a MG-defined formal query representation.

(3) Queries are executed at the location where the relevant data resides. That is, the sub-queries are moved to the data, rather than large quantities of data being moved to the clinician, which is prohibitively expensive given the quantities of data. The Query Analyser takes a formal query representation and de-composes into (a) formal query for local processing and (b) formal query for remote processing. It then forwards these de-composed queries to the Local Query Handler and the Remote Query Handler for the resolution of the request.

(4) The Local Query Handler generates query language statements (e.g. SQL) in the query language of the associated Local DB (e.g. MySQL). The result set is converted to XML and routed to the Result Handler.

(5) The Remote Query Handler is a portal for propagating a queries and results between sites. This handler forwards the formal query for remote processing (3b above) to the Query Analyser of the remote site. The remote query result set is converted to XML and routed to the Result Handler.

(6) The Result Handler is responsible for collecting query results – both local and remote. The query handlers return XML results, and these are "joined" to create the overall result to be sent back to the requestor – either the client of the Remote Query Handler.

Figure 3 shows the propagation of queries between sites.

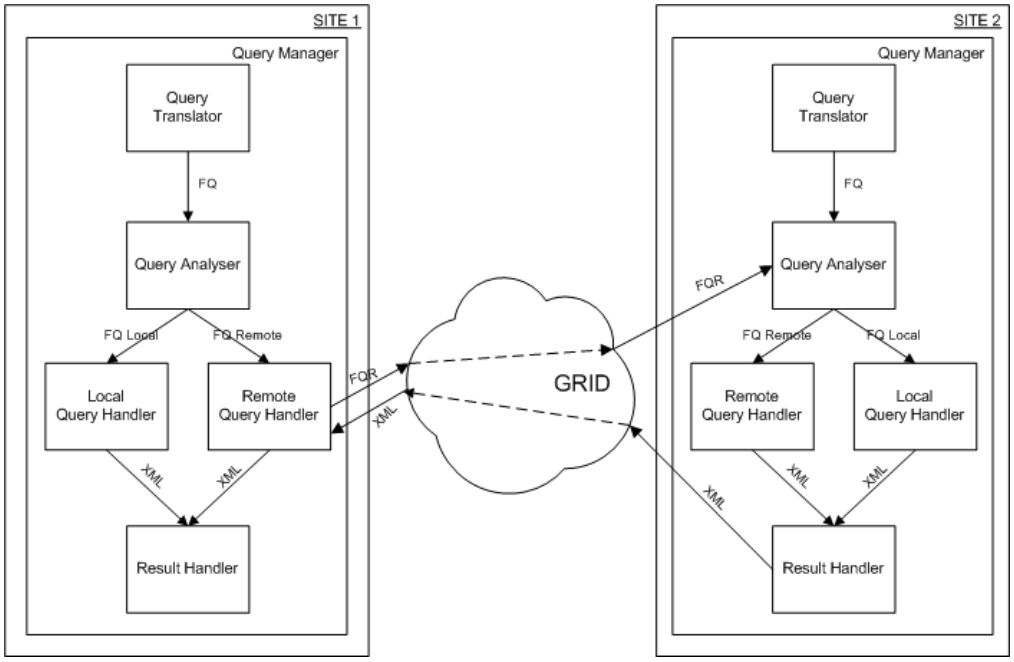


Figure 3: Propagation of Queries between Sites

4. Conclusions

The application of computer science in medicine is relatively young. Proliferation of information technology in medical sciences will undoubtedly continue, addressing clinical demands and providing increasing functionality. The MammoGrid project aims to advance deep inside this territory and explore the requirements of evidence-based, computation-aided radiology, as specified by medical scientists and practicing clinicians. This paper has emphasized two pillars which are likely to prove essential to the success of such a project: the importance of extensive requirements analysis and a design which caters for the complexity of the data.

The very nature of a project like MammoGrid implies that it is inconceivable to define an exhaustive list or even complete classification of all possible queries which the radiologists may need to run against the distributed database. Inevitably, when the user community starts using such a system, the requirements will undergo adjustments and extension. This paper has illustrated the kind of complexity of the expected queries, based on initial consultation of radiologists. It is suggested that the design with extensive use of meta-data, as in the by now well-tested CRISTAL system, is both capable of handling such complex queries in an efficient way and flexible enough to

adapt to changing requirements. A design which handles queries using a reflexive data model has been presented as the proposed query model for the MammoGrid infrastructure.

In its first year, the MammoGrid project has faced interesting challenges originating from the interplay between medical and computer sciences and has witnessed the excitement of the user community whose expectations from the a new paradigm are understandably high. As the MammoGrid project moves into the implementation and testing phase, further challenges are anticipated which will test these ideas to the full. We hope to return to this subject in future publications.

5. References

[1] I. Foster & C. Kesselman., "The Grid: Blueprint for a New Computing Infrastructure". Morgan Kaufmann publishers, 1998. ISBN 1558604758

[2] The Information Societies Technology project: "MammoGrid - A European federated mammogram database implemented on a GRID infrastructure". EU Contract *IST-2001-37614*

[3] C.J. Viborny, M.L. Giger, R.M. Nishikawa, "Computer aided detection and diagnosis of breast cancer", *Radiol. Clin. N. Am.* **38(4)**, 725-740, 2000.

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

[5] S. Ciatto, M. Rosselli Del Turco, P. Burke, C. Visioli, E. Paci and M. Zappa "Comparison of standard and double reading and computer-aided detection (CAD) of interval cancers at prior negative screening mammograms: a blind review", British Journal of Cancer, 89, 1645-1649 (2003)

[6] R. McClatchey et al. MammoGrid User Requirements Document V1.0 "Requirements for Large-Scale Distributed Medical Image Analysis", available on request from authors and see: http://mammogrid.vitamib.com/

[7] The Rational Unified Process Model. See http://www.rational.com For particular details consult:

http://www.rational.com/products/rup/whitepapers.jsp

[8] CRISTAL – Common Repositories and Information Systems to Track Assembly Lifecycles. F. Estrella, "Objects, Patterns and Descriptions in Data Management", PhD Thesis, University of the West of England, Bristol, England, December 2000.

Published in:
Clinical Radiology (2007) 62, 1044-1051
Received 12 May 2006; received in revised form 18 September 2006; accepted 29 September 2006

MammoGrid—a prototype distributed mammographic database for Europe

R. Warren[a], A.E. Solomonides[b], C. del Frate[c], I. Warsi[a], J. Ding[a], M. Odeh[b], R. McClatchey[b], C. Tromans[d], M. Brady[d], R. Highnam[e], M. Cordell[e], F. Estrella[b], M. Bazzocchi[c], S.R. Amendolia[f]

[a] Department of Radiology, University of Cambridge, Addenbrooke's Hospital, Cambridge, [b] Centre for Complex Cooperative Systems, University of West of England, Frenchay, Bristol, UK, [c] Insitituto di Radiologia, Universita di Udine, Italy, [d] Department of Engineering Science, Oxford University, Oxford, [e] Siemens Molecular Imaging, Oxford, UK, [f] ETT Division, CERN, 1211 Geneva 23, Switzerland

This paper describes the prototype for a Europe-wide distributed database of mammograms entitled MammoGrid, which was developed as part of an EU-funded project. The MammoGrid database appears to the user to be a single database, but the mammograms that comprise it are in fact retained and curated in the centres that generated them. Linked to each image is a potentially large and expandable set of patient information, known as metadata. Transmission of mammograms and metadata is secure, and a data acquisition system has been developed to upload and download mammograms from the distributed database, and then annotate them, rewriting the annotations to the database. The user can be anywhere in the world, but access rights can be applied. The paper aims to raise awareness among radiologists of the potential of emerging ''grid'' technology (''the second-generation Internet'').

Introduction

The increasing use of electronic formats for radiological images, including mammography (the particular focus of the authors), together with the fast, secure transmission of images and patient data, potentially enables many hospitals and imaging centres throughout Europe to be linked together to form a single ''virtual organization''. The technological possibilities are co-evolving with an appreciation of potential uses. Huge ''federated'' databases of mammograms, which appear to the user to be a single database, but are in fact retained and curated in the centres that generated them, has been tested as a useful model. Linked to each image would be a potentially large and expandable set of relevant information, known as metadata. This might comprise: patient age, exogenous hormone exposure, family and clinical history, even genetic information; information known to correspond to risk factors (e.g., diet, parity, breast density); as well as image acquisition parameters, including breast compression and exposure data, which affect image appearance. Levels of access to the images and metadata in the database would vary according to the certificated rights of the user: radiologists might have access to all of it, whereas epidemiologists and researchers would have more limited access, protecting patient privacy, in accordance with European legislation. Such databases might consist of millions of mammograms providing a common resource of images and metadata that might be used for a variety of purposes, such as education, research, and access to expert second reading.

Currently, such virtual organizations and huge federated databases do not exist. However, the enabling technology to realize them does. It is variously known as the ''second-generation Internet'', or the ''grid''.* The grid may be understood with reference to the Internet. The Internet provides access to a massive number of files, each of which

* The word ''grid'' as used in this paper should not be confused with the anti-scatter grid familiar to all mammographers. The two uses of the same word is unfortunate, but there is no relation at all between the two meanings.

has a unique resource locator (URL), or web address used to locate that file. There are almost no restrictions on what the content may be (or the accuracy) of files posted on the Internet. Also, although methods for secure financial transactions are now available, the Internet is notoriously insecure—most computer viruses are distributed via it. Despite these limitations, use of the Internet has grown explosively over recent years. At its technical heart, the Internet comprises a remarkably small set of software standards and protocols that have been universally adopted.[†] The grid is, similarly, an emerging set of software standards that builds upon those developed for the Internet, to provide for consultation of files and for distributed computing. The name derives from an analogy with the electrical power grid. The grid aims to provide computing power in much the same way. Like the Internet, it is not necessary to know how the grid works in order to use it. It is sufficient for this paper to accept the following assertions: (1) the grid is an emerging set of standards, not yet completely agreed; and (2) it has the potential to provide the technological infrastructure to realize the virtual organization and federated database concepts.

The EU funded project ''MammoGrid'' set out to explore the following conjecture: grid technology and standards have evolved to the point where a prototype, federated database of mammograms might be constructed, based on centres in three European countries (UK, Italy, Switzerland; Fig. 1). More specifically, as part of the MammoGrid project, the following parameters were explored:

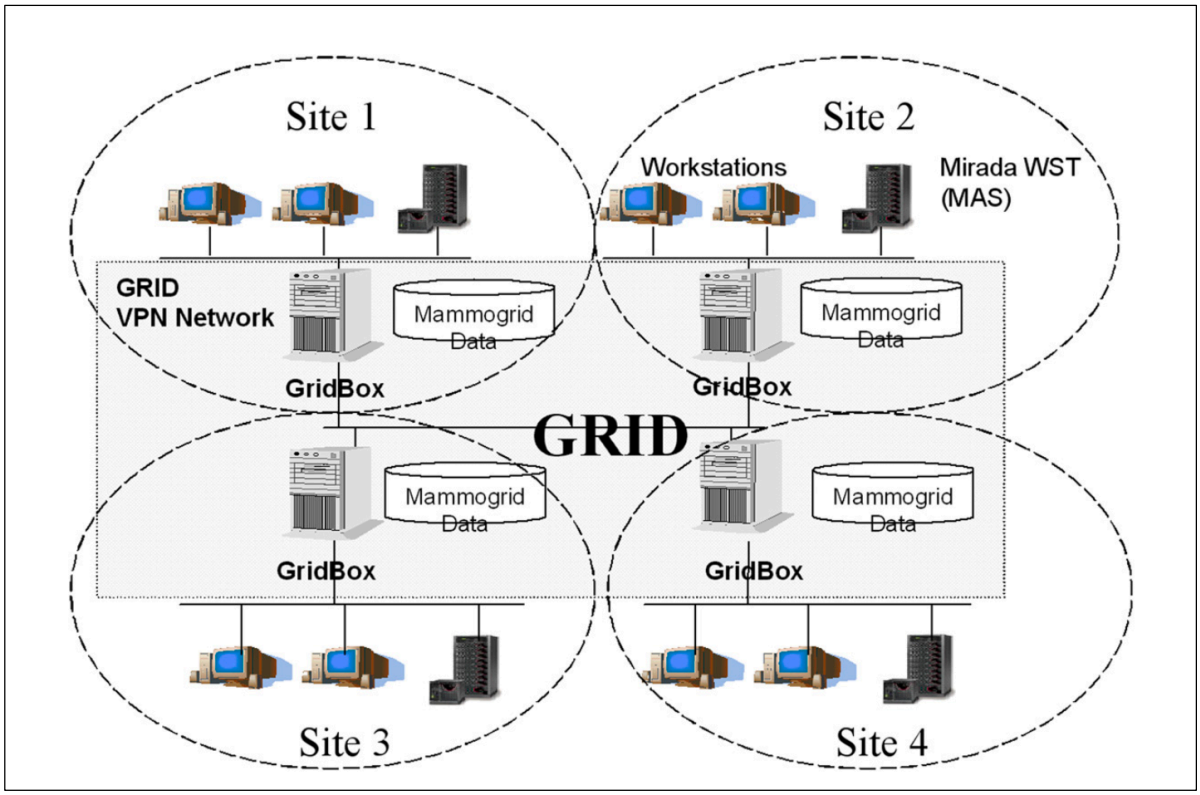


Figure 1 Schematic of the MammoGrid project. Four sites are shown contributing mammograms to the distributed database, at Udine, CERN (Geneva), and Cambridge. An additional node is shown at Siemens Molecular Imaging (formerly CTI Mirada) in Oxford, where the MammoGrid data acquisition system was developed. One node, not shown on the diagram, is located in Oxford University.

[†] The software standards are: html - hypertext markup language; http - hypertext transfer protocol; and tcp/ip: transmission control protocol & internet protocol. The familiar www is a particular network application that uses tcp/ip, and http enables resources to be communicated over the Internet.

Image standardization

As noted above, the appearance of a mammogram is greatly affected by differences in image-acquisition processes (machine type, filter, exposure time, etc.) As part of the MammoGrid project, the possibility of standardizing images using the Standard Mammogram Form (SMF) representation, developed by Highnam and Brady, [1] to support the applications of research study was explored. The Generate-SMF program used in this project is installed on the ''grid box'' employed between the partner organizations of the MammoGrid project. [2,3] The SMF program supports two areas of research by the clinical partners in the collaboration: (1) Breast density as a risk factor. Mammographic density is recognized as a major risk factor for breast cancer. [4] The SMF program provides a measure of the amount and proportion of dense tissue, so that partners in Cambridge and Udine could compare breast density measurements provided by SMF with the standard method of visual assessment [5] and an automated, two-dimensional (2D), interactive computer program. [6] (2) Computer-aided detection of microcalcifications and masses. Before MammoGrid, the partners from Sassari and Pisa had developed a system named ''CALMA'' (computer-aided library for mammography) for the detection of lesions and microcalcifications with good sensitivity and specificity, as previously reported. [7,8] The database of mammograms generated during the MammoGrid project at Cambridge and Udine was used to reassess the sensitivity and specificity of CALMA, and to examine whether its performance would be improved using the standardized images generated by SMF.

The present study tests the use of the database for research in three areas: (1) epidemiological studies of breast cancer risk using mammographic density estimated by a computer method; (2) to train pattern recognition algorithms, such as those used for computer-aided detection (CAD). The variability of breast anatomy and imaging conditions (machine type, exposure time, tube voltage, filter, etc.) imply the need for a huge number of training cases for pattern-recognition algorithms. [9] For data mining applications (''FindOneLikeIt'') in which a radiologist queries the database to find images (or a region of interest in an image) that resemble the current case and for which the diagnosis is biopsy proven.

Material and methods

The project has approval from the Cambridge and Norfolk research ethical committees (LREC; two separate submissions). UK women participating in one subproject where heights and weights were obtained gave written informed consent to participate. Italian women gave consent by agreeing to informed participation, which involved submitting to height and weight measurement and allowing their data to be used for research. The MammoGrid project has approval from the Cambridge and Norwich LRECs (2002) for storage of anonymized images with associated metadata in an encrypted format for viewing between the project partners, and for cancers and controls from screening they did not require informed consent to be obtained in view of the anonymous nature of the project and the soundness of the encryption process.

The MammoGrid project has several technical aspects: (1) image standardization using SMF; (2) the development of a workstation on which images can be acquired, annotated, and uploaded to the grid; (3) the distribution of data, images, and clinician queries across grid-based databases. For the UK, ethical approval was obtained for use of anonymized data, but patient consent was not requested. The Italian participants did not require any ethical application. Ethical, legal, confidentiality, and security constraints apply differently in the partner countries of origin, and must be respected.

The aim has been to use, wherever possible, freely available software to make queries across a widely distributed, federated database of mammographic images, and to perform epidemiological studies and CAD on the sets of returned images. For example, in the MammoGrid project, radiologists may annotate (i.e., mark out) different regions of a mammogram, which are then subjected to different CAD algorithms (including CALMA) and compared with stored mammograms in the database. As any one of these stages may be executed independently or take some time to be completed, the process must be controlled in a way that recognizes the current state of the computation and ensures that results are meaningfully assembled from the various partial outcomes. To provide for these possibilities, MammoGrid has built on the results of previous European grid projects, such as EGEE. [10-12] The details of EGEE and how it was used in MammoGrid have been described elsewhere. [13]

To be clinically useful, it is evident that the MammoGrid system has had to attain high levels of data integrity, quality, and consistency: these requirements have been met (by developing standard services, standard data formats, and

strict automated quality checks.) Once approval has been obtained from the relevant ethical committee on behalf of patients whose anonymized data will be held in the database, the system design enables certification and authorization services to guarantee anonymity and security. These critically important technical issues are described elsewhere. [13]

Image standardization: SMF

The SMF technology is a fully automated, objective measurement tool to estimate the volume of glandular tissue in the breast from a mammogram. [1,14] The SMF algorithm explicitly considers breast compression, exposure and tube voltage, and computes two volumetric measures of breast density, (1) the absolute volume (cm^3) of the breast that is dense (SMF volume) and (2) the percentage of the volume of the breast that is dense (SMF%). SMF is different from other volumetric research methods in that it incorporates a full physics model rather than using step-wedges in each image. [9,15] It has been described in two recent studies. [16,17] It is also used in the MammoGrid project to standardize the images on the grid box before using the CALMA CAD program to test whether improved detection follows.

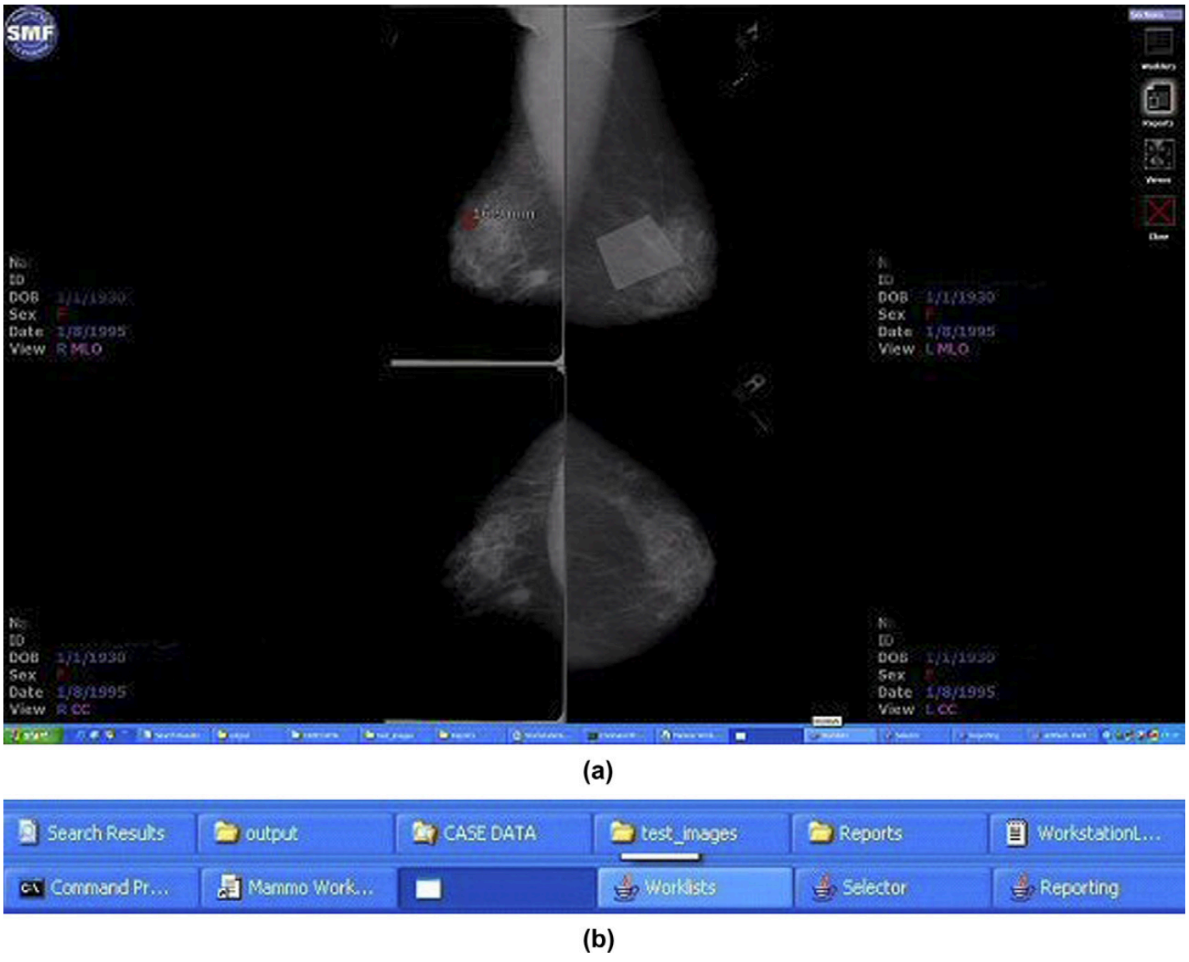


Figure 2 The MammoGrid data acquisition system, developed at Siemens Molecular Imaging (formerly CTI Mirada), in Oxford. A typical screenshot is shown (a) and an enlarged version of the user-interaction buttons shown in (b). The aim of this part of the project was not to develop a state-of-the-art visualization workstation, rather to develop one that could interact with the grid, to enable uploading/downloading of mammograms and annotations.

MammoGrid data acquisition workstation

A screenshot of the MammoGrid data acquisition workstation is shown in Fig. 2. The MammoGrid workflow currently begins with films, which must be digitized. The digitized film, with corresponding metadata, then needs to be uploaded, on the federated database so that it can subsequently be downloaded for processing. Subsequent processing may involve a radiologist supplying annotations, and there may be several such annotations, either done serially, or independently. Alternatively, processing may involve computer algorithms, either GenerateSMF to compute density, or to detect microcalcifications and/or masses. Again, such processes may be applied serially or independently. The results of all such processing need to be automatically uploaded to the database, with appropriate guarantees of security, confidentiality, and with an appropriate audit trail. Achieving all of these aims has resulted in the MammoGrid data acquisition workstation.

In the case that a radiologist examines a mammogram, in order to supply or revise an annotation, the case needs to be displayed. The development of a high-quality retrieval and mammogram display system was not a major concern of the MammoGrid project, as such display systems are available already from a range of manufacturers. [18] A rudimentary system was developed to the extent that the distributed database concept could be evaluated and the clinical studies accomplished.

User requirements

As in any major software project, the MammoGrid partners were confronted by a dilemma: because all the end-uses of the database were unknown the system was not fully developed, thus enabling alterations as new challenges and requirements were presented.

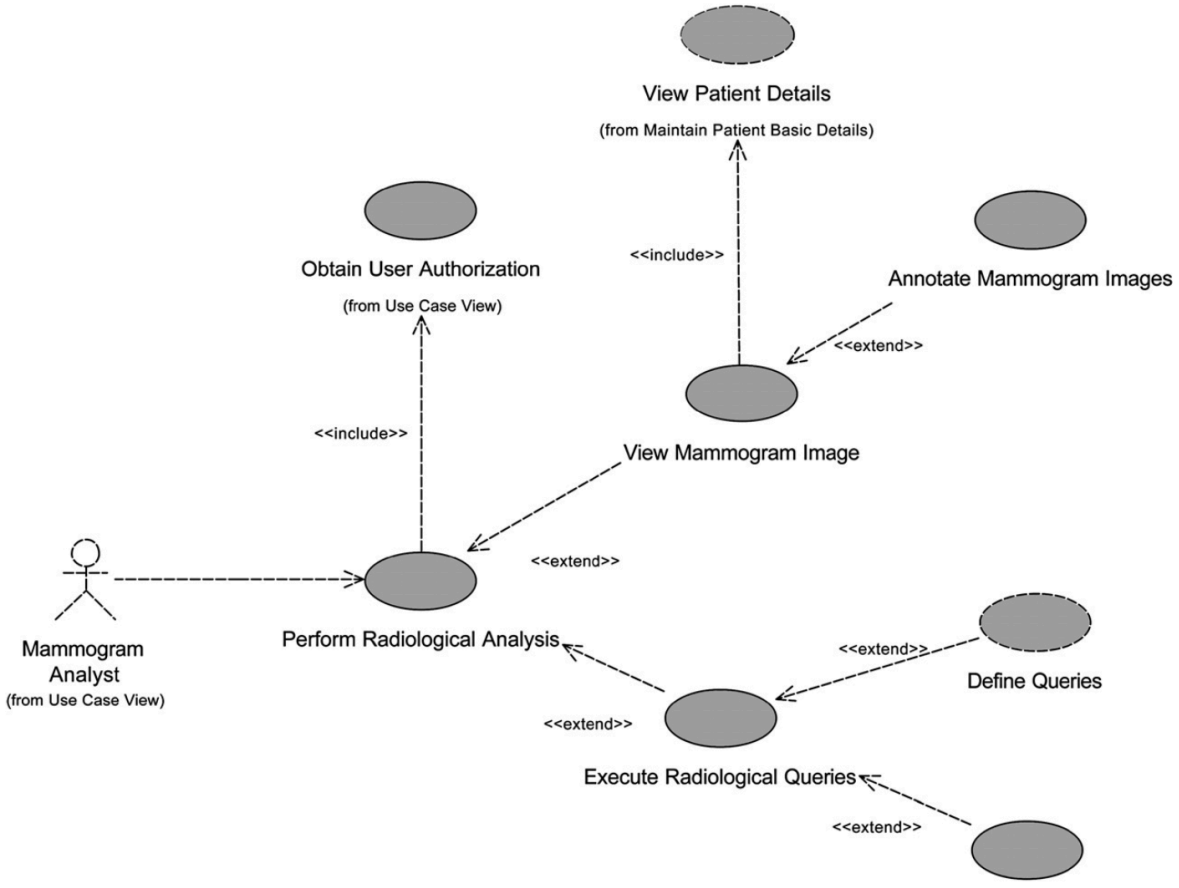


Figure 3 Typical use-case hierarchy and associated diagram. (See “User requirements”.)

The methodology based on ‘‘use-cases’’, [skeletal scenarios used to describe standard interactions between actors (roles enacted by users) and a given system] were used to determine the limits of that system (the system scope). This approach was adopted in MammoGrid to model and document the workflow of a radiology department. [19] This publication concludes that information infrastructures to support radiology not only must address issues related to the integration of clinical data from heterogeneous databases, but must facilitate access and filtering of patient data. An example of a use-case from the MammoGrid project, namely ‘‘perform radiological analysis’’ is shown in Fig. 3 (reproduced from [20]).

DICOM

The MammoGrid project conforms to the DICOM standard [21] in two respects. First, the digitized images are imported and stored in the DICOM storage format, so that the full set of image and patient-related metadata is readily available with the images, and that information exchange with other medical devices understanding the DICOM storage format is seamless. To ensure the compatibility with DICOM conformant clients, it is also necessary to exchange DICOM datasets via a communication protocol also defined by the standard.

CRISTAL and metadata

CRISTAL [11] is a distributed scientific database system used in high-energy physics experiments at CERN, The European Laboratory for Particle Physics. The CRISTAL project has studied the use of a description-driven approach using metadata-modelling techniques to manage the evolving data needs of a large community of scientists. This approach has been shown to provide many powerful features such as scalability, system evolution, interoperability, and reusability, aspects that are essential for future-proofing medical information systems. The MammoGrid project was based largely on the conjecture that CRISTAL would be ideally suited to being grid-enabled, in order to serve as the basis of a MammoGrid query handler.

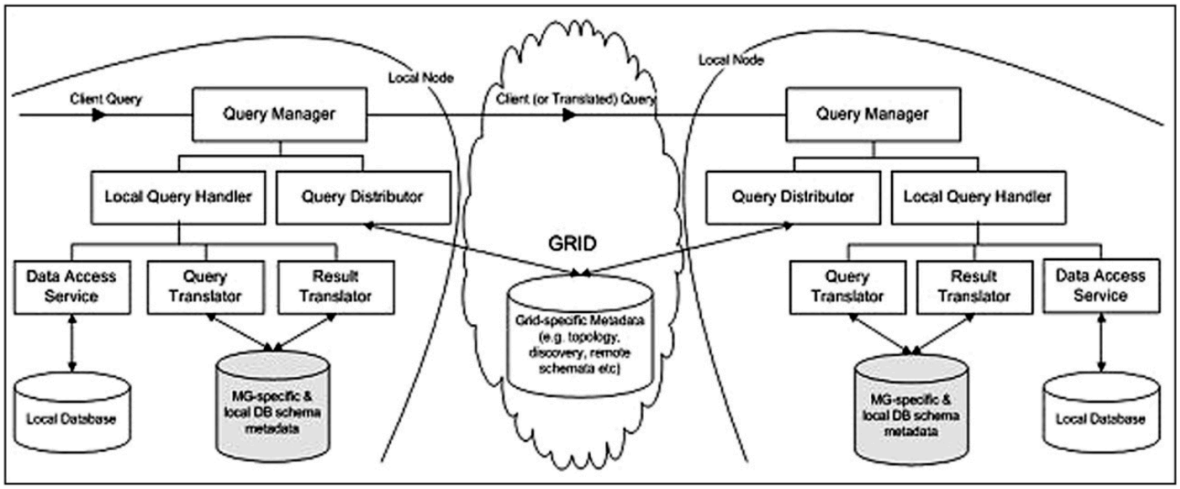


Figure 4 Handling a query in the MammoGrid distributed database system. (See CRISTAL and metadata.)

The handling of a grid database query is illustrated in Fig. 4. [12] A query-handling tool locally captures the elements of a clinician’s query and issues a formal query, using appropriate grid software, against the metadata structures and users’ structured terms, across multiple data centres in the distributed hospitals. At each location, the queries are resolved against relevant metadata and constituent sub-queries are remotely executed on the local mammogram databases. Selected sets of matching mammograms may then be either analysed remotely or replicated back to the centre at which the clinician issued the query for subsequent local analysis, depending on the philosophy adopted in the underlying grid software, but maintaining the impression that a single database has been accessed, as if resident at the local machine.

Joining centres together: ''grid boxes''

At the launch of the MammoGrid project, a software system named ''AliEn'' had been developed at CERN to provide a ''virtual file catalogue'' that enables transparent access to distributed datasets. The AliEn software was installed and configured on a set of novel secure hardware units, nicknamed grid boxes. The idea has been that each MammoGrid centre would have a single point of entry onto the MammoGrid (Fig. 1). Each grid box is responsible for storing new patient images and studies, updating the file catalogue, and propagating any changes in such a way that the view of database at the workstations is always up to date at every site.

Results

At the final EC review, the ability of the system to achieve complex queries across the grid to the satisfaction of technical reviewers, appointed by the EC, was demonstrated and confirmed as a functioning system. The experimental findings have been submitted separately to journals relevant to the clinical hypotheses; the findings of the mammography data have been published in this issue of Clinical Radiology. [22]

Discussion

MammoGrid is one of a number of European HealthGrid projects. [23,24] The technical approaches vary, but e-Diamond in the UK and MammoGrid in Europe have adopted the grid as their platform of choice proof of principle of the concept. As breast screening in the UK and in Italy has been based on film, mammograms have had to be digitized for use in both e-Diamond and MammoGrid. The UK ''e-Science'' programme has funded the e-Diamond project, [25,26] but whereas MammoGrid is based on open-source software, e-Diamond is based on (IBM) proprietary technology and concentrates on two complementary applications, namely teaching [26] and ''FindOneLikeIt''. Also, in the United States, the National Digital Mammography Archive [27] (NDMA) has adopted a radically different approach, using a large centralized archive of direct digital mammograms.

The central feature of the MammoGrid project is a geographically distributed, grid-based database of standardized images and associated patient data. The novelty of the MammoGrid approach lies in the application of grid technology and in the provision of data and tools, which enable radiologists to compare new mammograms with existing ones on the grid database, allowing them to make comparative diagnoses.

This project demonstrates the potential of the database, populated with provenance-controlled, reliable data from across Europe, and provides the prospect of statistically robust epidemiology, allowing analysis of lifestyle factors, including, for example, diet, exercise, and exogenous hormone use. A grid-based system such as MammoGrid would also be suitable for storing genetic or pathological image information. The project has attracted attention in the computer community as a paradigmatic exemplar of the application of grid technology. Although the problems required to install such systems have not yet been overcome, the project has established an approach and a prototype platform, sharing medical data, including images, across a grid. In loose collaboration with a number of other European medical grid projects, it is addressing the issues of informed consent and ethical approval, data protection, compliance with institutional, national and European regulations, and security. [12,28]

References

1. Highnam R, Brady M. *Mammographic image analysis*. 1st ed. Dordrecht: Kluwer Academic Publishers; 1999.
2. Estrella F, McClatchey R, Rogulin D, *et al. A service-based approach for managing mammography data*. Stud Health Technol Inform 2004;107:235e9.
3. Amendolia SR, Brady M, McClatchey R, *et al. MammoGrid: large-scale distributed mammogram analysis*. Stud Health Technol Inform 2003;95:194e9.
4. Boyd NF, Rommens JM, Vogt K, *et al. Mammographic breast density as an intermediate phenotype for breast cancer*. Lancet Oncol 2005;6:798e808.
5. Warner E, Lockwood G, Tritchler D, *et al. The risk of breast cancer associated with mammographic parenchymal patterns: a meta-analysis of the published literature to examine the effect of method of classification*. Cancer Detect Prev 1992;16:67e72.
6. Byng JW, Yaffe MJ, Lockwood GA, *et al. Automated analysis of mammographic densities and breast carcinoma risk*. Cancer 1997;80:66e74.

7. Bazzocchi M, Facecchia I, Zuiani C, *et al. Application of a computer-aided detection (CAD) system to digitalized mammograms for identifying microcalcifications*. Radiol Med (Torino) 2001;101:334e40.
8. Bottigli U, Delogu P, Fantacci ME, *et al. Search of Microcalcification clusters with the CALMA CAD station*. Int Soc Opt Eng (SPIE) 2002;4684:1301e10.
9. Pawluczyk O, Augustine B, Yaffe M, *et al. A volumetric method for estimation of breast density on digitized screen-film mammograms*. Med Phys 2003;30:352e64.
10. Robert Jones, *et al. The Information Societies Technology Project*. In: EU-EGEE, EU Contract IST-2003e508833, http://www.eu-egee.org/. Page accessed 22nd [May 2007].
11. F. Estrella, Z. Kovacs, J-M. Le Goff & R. McClatchey. *Meta-Data Objects as the Basis for System Evolution* Lecture Notes in Computer Science; vol. 2118, Advances in Web-Age Information Management, pp. 390e399 ISBN: 3-540-42298-6 Springer-Verlag, 2001 (Presented at the Second International Conference, WAIM 2001 Xi'an, China, July 9-11, 2001)
12. Estrella F, Hauer T, McClatchey R, *et al. A grid information infrastructure for medical image analysis*. In: Gibaud BDM, editor. Proceedings of the International Workshop on Distributed Databases and Processing in Medical Image Computing; 2004; Rennes, France 2004.
13. Amendolia SR, Estrella F, Del Frate C, *et al. Deployment of a grid-based medical imaging application*. Stud Health Technol Inform 2005;112:59e69.
14. Marias K, Behrenbruch C, Highnam R, *et al. A mammographic image analysis method to detect and measure changes in breast density*. Eur J Radiol 2004;52: 276e82.
15. Berks M, Diffey J, Hufton A, *et al. Feasibility and acceptability of stepwedge based measurement*. In: Astley S, editor. International Workshop on Digital Mammography. Berlin: Springer Verlag; 2006. p. 355e61.
16. Jeffreys M, Warren R, Highnam R, *et al. Initial experiences of using an automated volumetric measure of breast density: the standard mammogram form*. Br J Radiol 2006;79: 378e
17. Highnam R, Pan X, Warren R, *et al. Breast composition measurements using retrospective standard mammogram form (SMF).* Phys Med Biol 2006;51:2693e713.
18. Evertsz CJG, Bodicker A, Bohnenkamp S, *et al. Soft-copy reading environment for screening mammography-Screen*. In: Yaffe MJ, editor. Digital Mammography. The 5th International Workshop on Digital Mammography, IWDM 2000, Toronto, Canada. Madison WI, USA: Medical Physics Publishing; 2000. p. 566e72.
19. Bui AA, Taira RK, Dionisio JD, *et al. Evidence-based radiology: requirements for electronic access*. Acad Radiol 2002; 9:662e9.
20. Odeh M, Hauer T, McClatchey R, et al. *Use-case driven approach in requirements engineering: the MammoGrid Project*. In: Hamza MH, editor. Proceedings of the 7th IASTED Int. Conference on Software Engineering & Applications; 2003; Marina del Rey, CA. ACTA Press; 2003. p. 562e7. ACTA Press Calgary, Canada.
21. *The DICOM standard*. In: Digital imaging and communications in medicine. http://medical.nema.org/. Accessed May 2005, authors National Electrical Manufactures Association.
22. Warren R, Thompson D, del Frate C, *et al. A comparison of some anthropometric parameters between an Italian and a UK population: ''proof of principle'' of a European project using MammoGrid*. Clin Radiol 2007;62: 1052e60.
23. Amendolia SR, Estrella F, del Frate C, *et al. Deployment of Grid-based medical imaging application*. Stud Health Technol Inform 2005;112:59e69.
24. Nørager S, Paindaveine YT. *The HealthGrid terms of reference*. Brussels: EU; 2002.
25. Brady M, Gavaghan M, Simpson A, et al. *eDiamond: a grid-enabled federated database of annotated mammograms*. In: Berman F, editor. Grid computing: making the global infrastructure a reality. Chichester, UK: Wiley; 2003. p. 923e43.
26. Brady M, Gilbert F, Lloyd S, et al. *eDiamond: the UK's digital mammography national database*. In: Pisano E, editor. International Workshop on Digital Mammography. NC, USA: Chapel Hill; 2004.
27. *The National Digital Mammography Archive*. http://www. ndma.us/pdf/NDMA_brochure_2006.pdf (accessed on 23rd May 2007).
28. T Solomonides, R McClatchey, V Breton, Y Legre & S Norager (Editors). *From Grid to HealthGrid* Studies in Health Technology & Informatics vol. 112, ISBN: 1-58 603-510-X, ISSN: 0926-9630 IOS Press. (Proceedings the 3rd HealthGrid International Conference (HealthGrid 2005). Oxford, UK. April 2005.)

Experiences of engineering Grid-based medical software

F. Estrella, T. Hauer, R. McClatchey, M. Odeh, D. Rogulin, T. Solomonides

Centre for Complex Cooperative Systems, CEMS Faculty, University of the West of England, Coldharbour Lane, Frenchay, Bristol BS16 1QY, United Kingdom

{florida.estrella | tamas.hauer | dmitry.rogulin}@cern.ch;
{richard.mcclatchey | mohammed.odeh | tony.solomonides}@uwe.ac.uk

Abstract

Objectives: Grid-based technologies are emerging as potential solutions for managing and collaborating distributed resources in the biomedical domain. Few examples exist, however, of successful implementations of Grid-enabled medical systems and even fewer have been deployed for evaluation in practice. The objective of this paper is to evaluate the use in clinical practice of a Grid-based imaging prototype and to establish directions for engineering future medical Grid developments and their subsequent deployment.

Method: The MammoGrid project has deployed a prototype system for clinicians using the Grid as its information infrastructure. To assist in the specification of the system requirements (and for the first time in healthgrid applications), use-case modelling has been carried out in close collaboration with clinicians and radiologists who had no prior experience of this modelling technique. A critical qualitative and, where possible, quantitative analysis of the MammoGrid prototype is presented leading to a set of recommendations from the delivery of the first deployed Grid-based medical imaging application.

Results: We report critically on the application of software engineering techniques in the specification and implementation of the MammoGrid project and show that use-case modelling is a suitable vehicle for representing medical requirements and for communicating effectively with the clinical community. This paper also discusses the practical advantages and limitations of applying the Grid to real-life clinical applications and presents the consequent lessons learned.

Conclusions: The work presented in this paper demonstrates that given suitable commitment from collaborating radiologists it is practical to deploy in practice medical imaging analysis applications using the Grid but that standardization in and stability of the Grid software is a necessary pre-requisite for successful healthgrids. The MammoGrid prototype has therefore paved the way for further advanced Grid-based deployments in the medical and biomedical domains.

1. Introduction

The past decade has witnessed order of magnitude increases in computing power and data storage capacity, giving birth to new applications which can handle large data volumes of increased complexity. Similar increases in network speed and availability pave the way for applications distributed over the web, carrying the potential for better resource utilization and on-demand resource sharing. Medical informatics is one of the areas in which these technologically revolutionary advances could bring significant benefit both for scientists' research study and clinicians' everyday work. Recently there has been much excitement in the parallel systems community as well as that of distributed database

applications in the emergence of 'The Grid' as a promising platform for scientific and medical collaborative computing.

The Grid is a new paradigm for distributed computing defined as the "flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources" [1]. Geographically separated yet working together to solve a problem, groups of people can harness the collection of resources provided by the participants and use the shared environment of the Grid within the boundaries of so-called Virtual Organizations (VO).

In essence the Grid:

- provides a virtual platform for large-scale, resource-intensive, and distributed applications;
- offers a connectivity environment allowing management and coordination of diverse and dispersed resources;
- enables access to increased storage capacity and computing power;
- provides mechanisms for sharing and transferring large amounts of data as well as aggregating distributed resources for running computationally expensive procedures;
- utilizes a common infrastructure based on open standards thus providing a platform for interoperability and interfacing between different Grid-based applications from the particular domain. Grid computing holds the promise of harnessing extensive computing resources located at geographically dispersed locations which can be used by a dynamically configured group of collaborating institutions; consequently, it defines a suitable platform on which to base distributed medical informatics applications. In particular the Grid can address some of the following issues relevant to medical domains.

1.1. Data distribution

The Grid provides a connectivity environment for medical data distributed over different sites. It solves the location transparency issue by providing mechanisms which permit seamless access to and the management of distributed data. These mechanisms include services which deal with virtualization of distributed data regardless of their location.

1.2. Heterogeneity

The Grid addresses the issue of heterogeneity by developing common interfaces for access and integration of diverse data sources. Such generic interfaces for consistent access to existing, autonomously managed databases that are independent of underlying data models are defined by the Global Grid Forum Database Access and Integration Services (GGF-DAIS) working group. These interfaces can be used to represent an abstract view of data sources which can permit homogeneous access to heterogeneous medical data sets.

1.3. Data processing and analysis

The Grid offers a platform for transparent resource management in medical analysis. This allows the virtualization and sharing of all resources (e.g. computing resources, data storage, etc.) connected to the grid. For handling computationally intensive procedures (e.g. CADe), the platform provides automatic resource allocation and scheduling and algorithm execution, depending on the availability, capacity and location of resources.

1.4. Security and confidentiality

Enabling secure data exchange between hospitals distributed across networks is one of the major concerns of medical applications. Grid addresses security issues by providing a common infrastructure for secure access and communication between grid-connected sites. This infrastructure includes authentication and authorization mechanisms, among other things, supporting security across organizational boundaries.

1.5. Standardization and compliance

Grid technologies are increasingly being based on a common set of open standards (such as XML, SOAP, WSDL, HTTP, etc.) and this is promising for future medical image analysis standards.

In other words, Grid computing has the capacity to resolve many of the exceptional difficulties encountered in medical informatics by allowing medical doctors and researchers to collaborate without having to co-locate, thereby providing transparent access to data and computing resources. To date, however, there have been few projects that have attempted to deliver Grid computing to clinicians and there is no practice-based evidence or guidelines as to where and how the Grid can benefit clinicians. Furthermore there are few articles in the established medical informatics journals that have covered aspects of Grid computing [3,4].

In this paper, we report on how a Grid architecture has been used to provide both computing power and a distribution platform to a community of radiologists spanning multiple medical institutions, even across borders, and in so doing to investigate the particular issues surrounding the implementation and use of Grid technologies in a clinical environment. MammoGrid exploited existing and emerging technologies to build a large-scale database of mammograms and associated metadata[1] that can be used to investigate healthcare applications and to explore the potential of the Grid to support effective co-working between healthcare professionals [5,6]. The use of digitized radiological images (mammograms) enabled linkage of distant centres for the first time in a “radiological virtual organization”. The MammoGrid project aimed to demonstrate that through such a virtual organization a Grid infrastructure can support collaborative medical image analysis, and to enable radiologists to share standardized mammograms, compare diagnoses (with and without computer-aided detection), and perform epidemiological research studies across national boundaries. This paper highlights the specific constraints apparent in Grid-based distributed radiology and outlines how established software engineering techniques can be married with emerging Grid technologies to provide a first Grid-based mammogram analysis system.

The paper is structured as follows. First, we present the rationale behind the MammoGrid project and we restate its aims and objectives to provide justification for the approach followed in its development. In the next section, we investigate how this approach was constrained by two major factors: the nature of the clinical domain and the rapidly changing Grid research environment. Then, we describe (in Section 3) how we approached the MammoGrid research and development process to address these constraints. Emphasis on the requirements engineering phase of the project and the construction of a set of clinical use-cases for the capture of the radiologists’ system needs are presented in Section 4. Next, in Section 5, we identify the outcomes of the design and development phases of the MammoGrid project and outline its prototyping strategy. Then in Section 6 the service-oriented architecture of the delivered prototype is presented and discussed along with its suitability through a set of clinical tests. This is then followed by Section 7, where the main lessons from undertaking a software engineering

[1] ‘Metadata’ is used inclusively to encompass associated data (such as patient information), summary data (such as breast density) and metadata proper, such as information about the logical or physical distribution of the data.

approach in the MammoGrid project are summarized. Finally, we draw conclusions (in Section 8) on the software engineering process and give guidelines for future development and deployment of Grid-based systems in the medical domain.

2. Background

The Fifth Framework EU-funded MammoGrid project aimed to apply the Grid concept to mammography, including services for the standardization of mammograms, computer-aided detection (CADe) of salient features, especially masses and 'microcalcifications', quality control of imaging, and epidemiological research including broader aspects of patient data. In doing so, it attempted to create a paradigm for practical, Grid-based healthcare-oriented projects, particularly those which rely on imaging. There are, however, a number of factors that make patient management based on medical images particularly challenging. Often very large quantities of data, with complex structures, are involved (such as 3D images, time sequences, multiple imaging protocols, etc.). Also, clinicians rarely analyse single images in isolation but rather in the context of metadata. Metadata that may be required are clinically relevant factors such as patient age, exogenous hormone exposure, family and clinical history; for the population, natural anatomical and physiological variations; and for the technology, image acquisition parameters, including breast compression and exposure data. Thus any database of images developed at a single site may not contain enough exemplars in response to any given query to be statistically significant. Overcoming this problem implies constructing a very large and federated database, which can transcend national boundaries. However this will necessitate specialist image processing algorithms – for example, computationally heavy tasks operating on large files of images – which in turn place significant requirements on storage space, CPU power and/or network bandwidth on all participating hospitals, unless appropriate sharing of computing resources is arranged. Realising such a geographically distributed (pan-European) database therefore necessitates a Grid infrastructure, and the construction of a prototype model which would push emerging Grid technology to its limits.

The MammoGrid project was carried out between mid 2002 and the end of 2005 and involved hospitals and medical imaging experts and academics in the UK, Italy and Switzerland with experience of implementing Grid-based database solutions. A key deliverable of the project was a prototype software infrastructure based on an open-source Grid 'middleware' (i.e. software that enables an underlying Grid infrastructure to host domain applications) and a service-oriented database management system that is capable of managing federated mammogram databases distributed across Europe. The proposed solution was a medical information infrastructure delivered on a service-based, grid-aware framework, encompassing geographical regions of varying clinical protocols and diagnostic procedures, as well as lifestyles and dietary patterns. The prototype will allow, among other things, mammogram data mining, diverse and complex epidemiological studies, statistical and (CADe) analyses, and the deployment of versions of the image standardization software. It was the intention of MammoGrid to get rapid feedback from a real clinical community about the use of such a simple Grid platform to inform the next generation of Grid projects in healthcare.

The clinical workpackages encompassed in MammoGrid prototypes address three selected clinical problems:

• *Quality control*: the effect of image variability, due to differences in acquisition parameters and processing algorithms, on clinical mammography;

• *Epidemiological studies*: the effects of population variability, regional differences such as diet or body habitus and the relationship to mammographic density (a biomarker of breast cancer) which may be affected by such factors;

• Support for radiologists, in the form of tele-collaboration, second opinion, training and quality control of images.

Other initiatives against which MammoGrid may be compared include: the eDiamond project in the UK, and the NDMA project in the US. The MammoGrid approach shares many similarities with these projects, but in the case of the NDMA project (one of whose principal aims is to encourage the adoption of digital mammography in the USA) its database is implemented in IBMs DB2 on a single server. The MammoGrid project federates multiple (potentially heterogeneous) databases as its data store(s). The Italian INFN project GP-CALMA (Grid Project CALMA) has focused on a Grid implementation of tumour detection algorithms to provide clinicians with a working mammogram examination tool. MammoGrid uses aspects of the CALMA project in its computer-aided detection of microcalcifications.

More recent Grid-based research includes the BIRN [10] project in the US, which is enabling large-scale collaborations in biomedical science by utilizing the capabilities of emerging Grid technologies. BIRN provides federated medical data, which enables a software 'fabric' for seamless and secure federation of data across the network and facilitates the collaborative use of domain tools and flexible processing/analysis frameworks for the study of Alzheimer's disease. The INFOGENMED initiative [11] has given the lead to projects in moving from genomic information to individualized healthcare using data distributed across Europe. Finally the CDSS [12] project is a system which uses knowledge extracted from clinical practice to provide a classification of patients' illnesses, implemented on a Grid platform.

From the outset, the MammoGrid project posted its objectives in terms of the promised radiological and epidemiological applications, but not in terms of new Grid technology. Its technology attitude has largely been one of re-use, not invention or development; only where required functionality was missing was there a need to implement new Grid services. An information infrastructure to integrate multiple mammogram databases is clearly needed to enable clinicians to develop new common, collaborative and co-operative approaches to the analysis of mammographic images as is evident by the clinical evaluation that took place towards the end of the MammoGrid project.

3. The development environment

The development and deployment of the MammoGrid prototypes was carried out between 2002 and 2005 by a group of software engineering researchers from the University of the West of England (UWE, UK) and from European Centre for Particle Physics (CERN, Switzerland), groups of medical imaging experts from Mirada Solutions (UK) and the Universities of Oxford, Pisa and Sassari, and research radiologists from the University hospitals of Udine (Italy) and Addenbrookes, Cambridge (UK). The approach followed was very pragmatic in nature rather than one which rigorously followed a traditional software engineering process and was loosely based around an evolutionary prototyping philosophy; nevertheless, in its early stages, the project conducted its requirements capture by utilising aspects of the Rational Unified Process Model (RUP) [13] and by establishing a set of use-cases which subsequently were used as part of the clinical evaluation of the project's outcomes.

Due to the nature of the development environment there were a number of constraints on the software engineering process. *First*, the development was research-oriented both in terms of the maturity of the still-emerging Grids middleware and the novelty of the MammoGrid services for the medical community. This constraint led to several challenges: the need to raise the clinicians' awareness of Grid technologies and the use of requirements modelling techniques while managing their expectations of what these technologies might deliver; the necessity to cater for frequent releases of new underlying software technologies while at the same time adhering to existing medical standards and protocols, and the need to re-train researchers 'on-the-job' as new middleware became available.

Second, the MammoGrid project was carried out under both tight manpower and time restrictions. This necessitated careful project management of the collaboration between busy clinicians, software engineers, developers and computer science doctoral students. As a consequence, this required significant participation from the user community (radiologists, radiographers) especially during the requirements engineering phase of the project along with frequent feedback and validation of project findings (mainly through systems 'walkthroughs' or sometimes 'stomp-overs') with the software engineers and researchers. *Third*, the project required research and development effort spread over eight institutes located in the UK, Italy and Switzerland, a substantial challenge to project management. It was therefore necessary to delineate clear task boundaries and to establish inter-task dependencies, so that explicit responsibilities for the production of deliverables were established and that, as a group, the project respected those responsibilities and adhered to delivery milestones. During the early stages of the project, strong bonds of trust and mutual respect were built between project members. Also, agreed project management structures and deliverables/milestones were put in place. Project members remained committed to the common goals despite shortages of resources and despite differences in priorities between research and commercial partners.

Finally, during the later stages of the project the deployment and testing of the system were subject to delivery constraints of software from commercial partners and subject to the maturity and stability of the underlying Grid technologies as it evolved during the project. This inevitably led to delays in the successful integration of the final MammoGrid prototype and the testing and validation of the clinical use-cases.

Constraints on the requirements elicitation, specification and validation phases included the limited time available with domain experts, the geographic distance between the various stakeholders and hence the episodic nature of meetings. In the course of a visit of several days, domain experts could make themselves available for relatively frequent but rather brief meetings. Domain experts had no previous exposure to the kind of model used in software development. Software engineers on the project had some appreciation but very little experience to the particular problems of mammography and breast cancer screening prior to this exercise. Moreover, the requirements team had to span the space between radiologists, Grid experts and medical image processing specialists, whether those working on the specification of the local workstation or on the CADe software.

4. The requirements engineering phase

The MammoGrid project was driven by the requirements of its user community – Udine and Cambridge hospitals – along with medical imaging expertise from Oxford. The ultimate objective of the requirements engineering phase was to obtain an agreed, validated and essentially stable requirements specification document for the project. Two core objectives for the project followed immediately from its scope and definition:

- The support of clinical research studies through access to and execution of algorithms on physically large, geographically distributed and potentially heterogeneous sets of (files of) mammographic images, just as if these images were locally resident;
- The controlled and assured access of educational and commercial companies to distributed mammograms for testing novel medical imaging diagnostic technologies in scientifically acceptable clinical trials that fulfil the criteria of evidence-based medical research.

To facilitate requirements specification, a number of meetings took place between software engineers and radiologists at Udine and Cambridge to elicit, and then analyse and specify the functional requirements of the end-user radiologists and radiographers (radiology technicians) in addition to product-related non-functional requirements. Use-case and conceptual data (object-oriented) models

were incrementally and iteratively developed and validated as the main requirements models followed by dynamic interaction and state transition diagrams. Parallel to the requirements elicitation activities, the hardware and software requirements were established. Meanwhile, the logical view of the application architecture was developed following iterations on activities in the requirements and design workflows of RUP. The requirements were specified by a group of UWE software engineers working with domain experts from Mirada Solutions and the participating hospitals.

These activities resulted in identifying major use-case scenarios in the use of a distributed database of mammograms deployed across a pan-European grid that were validated with the domain specialists and then later used to prove the project prototype(s) in clinical evaluation. Problem domain entities (classes) were identified and described in addition to documenting relationships between such entities, resulting in a stable and validated conceptual class model which has since evolved as the logical class model. The system level use-case model of MammoGrid is shown in Fig. 1. A detailed example, the essential function of mammogram analysis carried out by a "mammogram analyst" (one of key system actors, normally enacted by a clinician) is briefly modelled and presented in Fig. 2 using the core use-case "Perform Radiological Analysis". Further details of the use-case model can be found in [14].

Use-case analysis and modelling identified the major actors of the system, differentiating these from job roles or individuals, and investigated how they impacted principal system functions across several scenarios. This was further iteratively validated in a process that involved all stakeholders, including users, researchers and developers; this strengthened the cohesion of the project, provided a common visual language for communication and problem representation and led to an agreed requirements specification.

The main requirements elicitation methods used were semi-structured interviews and strictly non-participant observation of medical procedures; with appropriate permissions, the team observed such procedures as basic mammography, ultrasound-guided biopsy, breast MRI, reading of mammograms and other images, and X-ray examination of biopsy specimens. The first resulting requirements models (i.e. the use-case models) were established and then iteratively and incrementally presented to radiologists at Udine and CADe experts from Pisa and Sassari (with contributions from a further private hospital in Torino), then to radiologists at Cambridge, then to imaging specialists and finally to two project plenary meetings. In parallel, the team from Mirada Solutions constructed the 'acquisition system' prototype and defined related data structures. It was then possible to cross refer and thus to validate the data requirements emerging from the use-cases. The acquisition system remained an architectural component of the MammoGrid.

An analysis of project non-functional requirements (NFRs) was conducted including product-related, organization and process requirements, external constraints, such as confidentiality, and interface specifications. In this analysis, constraints on the process of mammogram study, such as usability, reliability, robustness and security were investigated and specified as a means to assess the degree of adherence to these requirements at implementation time. In addition, the impact of product-related NFRs on architecture selection and specification was investigated.

The MammoGrid project has certain non-standard characteristics such as:

- a wide diversity of backgrounds among problem domain specialists (radiologists, radiographers, epidemiologists, medical imaging experts);
- the application domain itself, roughly speaking, the construction of the evidence base for radiological practice in mammography;
- the geographically dispersed locations of the different parties involved in the project;

- the need to establish *ab initio* the use of a modelling language, UML [15] (in particular, use-cases) with corresponding modelling, validation, and requirements management tools.

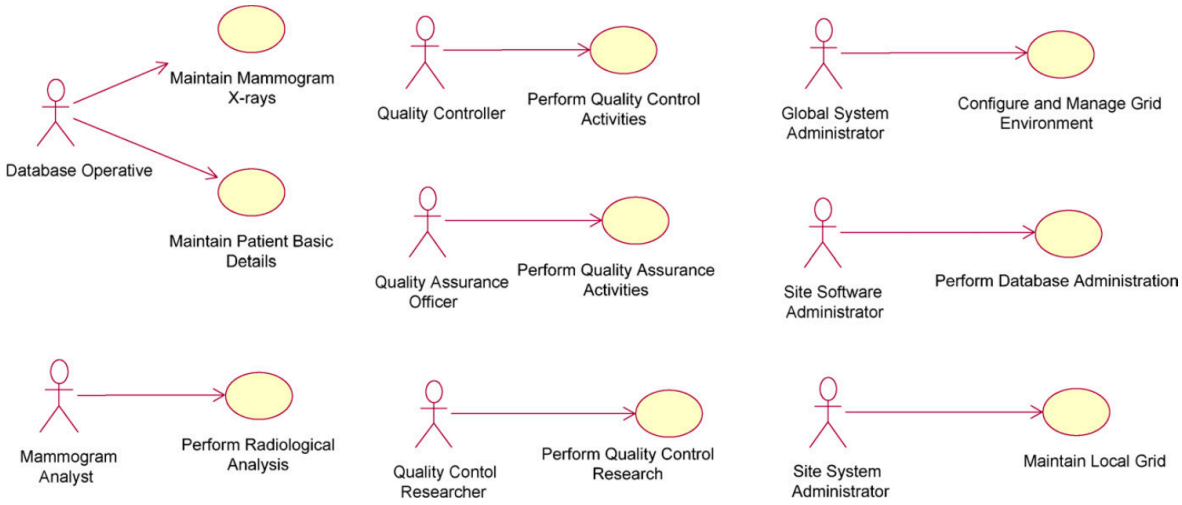


Fig. 1 – The MammoGrid system level use-case view.

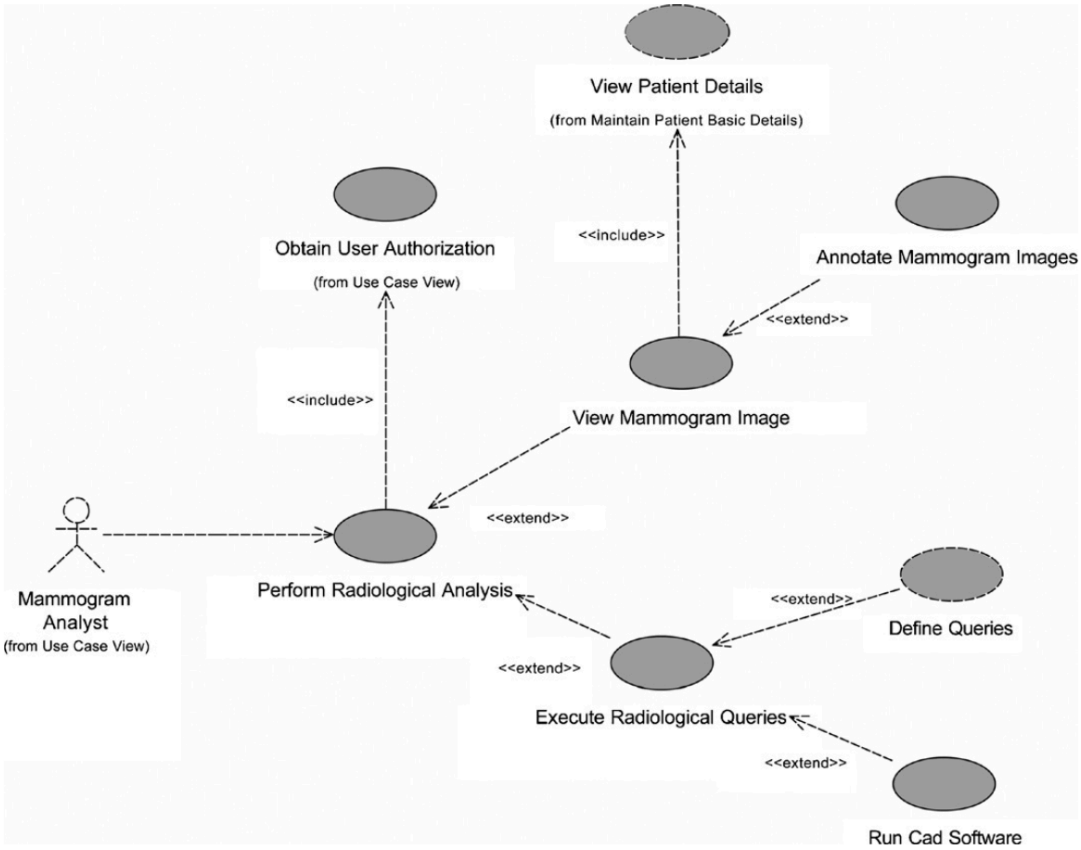


Fig. 2 – Use-case hierarchy and diagram.

5. The design and deployment phases

As set out in its goals, the MammoGrid project concentrated on applying existing Grid middleware rather than developing new Grid software: the design philosophy adopted in the project focused on services that address user requirements for collaborative mammogram analysis. One of the main deliverables of the project was an interface between the radiologist's image analysis workstation and the 'MammoGrid Information Infrastructure' (MII) based on the philosophy of a Grid. This enabled radiologists to query images across a widely distributed federated database of mammographic images and to perform epidemiological and CADe analyses on the sets of returned images. In delivering the MII, the MammoGrid project has customised and, where necessary, enhanced and complemented Grid software for the creation of a medical analysis platform. The approach that is being followed in the project is therefore twofold: to provide an MII based on a service-oriented architecture [16,17] and a metadata and query handler coupled to a 'front-end' to ensure that both patient data and images remain appropriately associated and that metadata based searches are effectively handled. The MII has been fully specified and a prototype delivered, in which a set of medical imaging services is implemented to manage the federation of distributed mammograms [18].

To encourage re-use the MammoGrid software was delivered through a set of evolving prototypes following a form of 'spiral model' [19] development (including 'stages' of planning, specification, evaluation, and development for each prototype version) in which the clinical user community provided input in each loop of the spiral. Release of the staged prototypes was planned to coincide with project milestones and the delivery of tested MammoGrid services. Involvement of the clinicians helped to maintain their engagement with the project at a stage when they could not yet draw benefit from any tangible system. This ensured commitment to project deliverables and enabled the software developers to gain a deeper understanding of the actual system requirements of the clinicians; these were important benefits of this design and its implementation approach. This strategy also enabled the project to cope easier with the multiple versions of the underlying Grid software that emerged during the lifetime of the project as well as with regular updates to the clinical workstation provided by Mirada Solutions.

The MammoGrid project has recently delivered its final proof-of-concept prototype enabling clinicians to store digitized mammograms along with appropriately anonymized patient metadata; the prototype provides controlled access to mammograms both locally and remotely stored. A typical database comprising several thousand mammograms has been created for user tests of clinicians' queries. The prototype comprises

- a high-quality clinician visualization workstation (used for data acquisition and inspection);
- an imaging standard–compliant interface to a set of medical services (annotation, security, image analysis, data storage and querying services) residing on a so-called 'Grid-box';
- secure access to a network of other Grid-boxes connected through Grids middleware.

6. Clinical evaluation of the MammoGrid prototype

The evaluation of the MammoGrid prototype was conducted in the final months of the project and was driven by assessing the achievement of the overall project objectives. The evaluation process therefore concentrated on the following aspects:

- The establishment and deployment of a MammoGrid virtual organization;
- The evaluation of the service oriented approach with the emphasis on clinical services provided by the Grid middleware layer;

- The use of the clinical workpackages (by senior radiologists) specified before the start of the project to drive the qualitative and quantitative evaluation of the implemented use-cases in order to provide
- Feedback for future 'healthgrid' projects on the clinical collaborative nature of the adopted approach in the form of lessons learnt.

The discussion below summarises the main outcomes of evaluating the MammoGrid's prototype in light of the above criteria.

To allow for the evaluation of the final prototype at the test clinical sites, which was the first attempt at studying the use of a Grid-based cross-national database by practicing radiologists, a MammoGrid Virtual Organization (MGVO) was established and deployed (as shown in Fig. 3). The MGVO was composed of three mammography centres – Addenbrookes Hospital, Udine Hospital, and Oxford University. These centres were autonomous and independent of each other with respect to their local data management and ownership. The Addenbrookes and Udine hospitals had locally managed databases of mammograms, with several thousand cases between them (see Table 1). As part of the MGVO, registered clinicians had access to (suitably anonymized) mammograms, results, diagnosis and imaging software from the other centres. Access was coordinated by the MGVOs central node at CERN.

In order to minimise development and maximise re-use of existing Grid software, the adopted middleware solution was the ALICE Environment (AliEn) [20] component of the EGEE-gLite middleware [21], i.e. the grid middleware of the EU-funded EGEE project [22]. The service-oriented approach adopted in MammoGrid permitted the interconnection of communicating entities, called services, which provided capabilities through exchange of messages. The services were 'orchestrated' in terms of service interactions: how services were discovered, how they were invoked, what could be invoked, the sequence of service invocations, and who could execute them.

The MammoGrid Services (MGS) are a set of services for managing mammography images and associated patient data on the grid. Fig. 4 illustrates the services that made up the MGVO (for simplicity, Oxford University has not been included).

The MGS are: (a) *Add* for uploading files (DICOM [23] images and structured reports) to the MGVO; (b) *Retrieve* for downloading files from the grid system; (c) *Query* for querying the federated database of mammograms; (d) *AddAlgorithm* for uploading executable code to the Grid; (e) *ExecuteAlgorithm* for executing grid-resident executable code on grid-resident files on the Grid system; (f) *Authenticate* for logging into the MGVO. See [18] for further details.

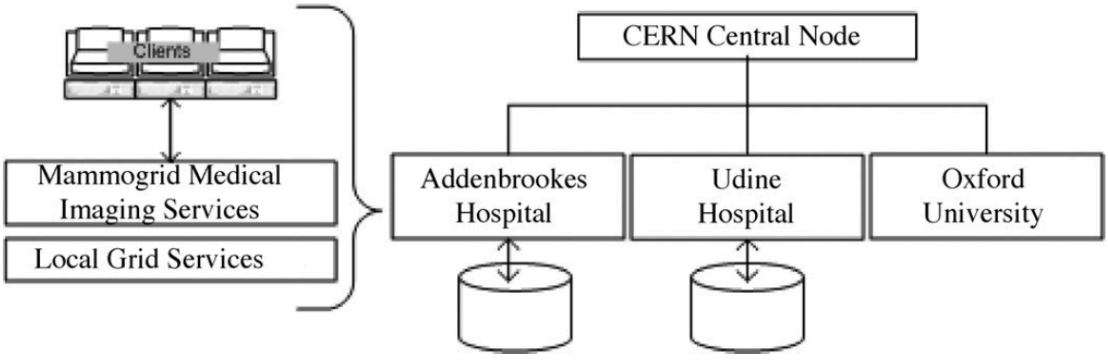


Fig. 3 – The MammoGrid virtual organization (MGVO).

Site	Number of patients	Number of image files	Number of SMF files	Associated database size (Mb)	File storage size (Gb)
Cambridge	1,423	9,716	4,815	14.0	260
Udine	1,479	17,285	8,634	23.5	220
Total	2,902	27,001	13,449	37.5	480

Table 1 – Virtual repository size of the MammoGrid prototype

Evaluation of the MammoGrid prototype took place using the MGVO over a set of clinical workpackages performed by senior radiologists at Addenbrookes and Udine hospitals. The MammoGrid Virtual Organization encompassed data accessible to the radiologists at the hospitals, as well as at Oxford University and CERN. The evaluation comprised the qualitative and, where possible, quantitative assessment of the use-cases captured during the requirements elicitation phase of the project (detailed in Section 4). The domain of the evaluation reflected on the key elements of the clinical workpackages, as identified in Section 2, and these are:

- *Quality control*: the effect of image variability, due to differences in acquisition parameters and processing algorithms, on clinical mammography;

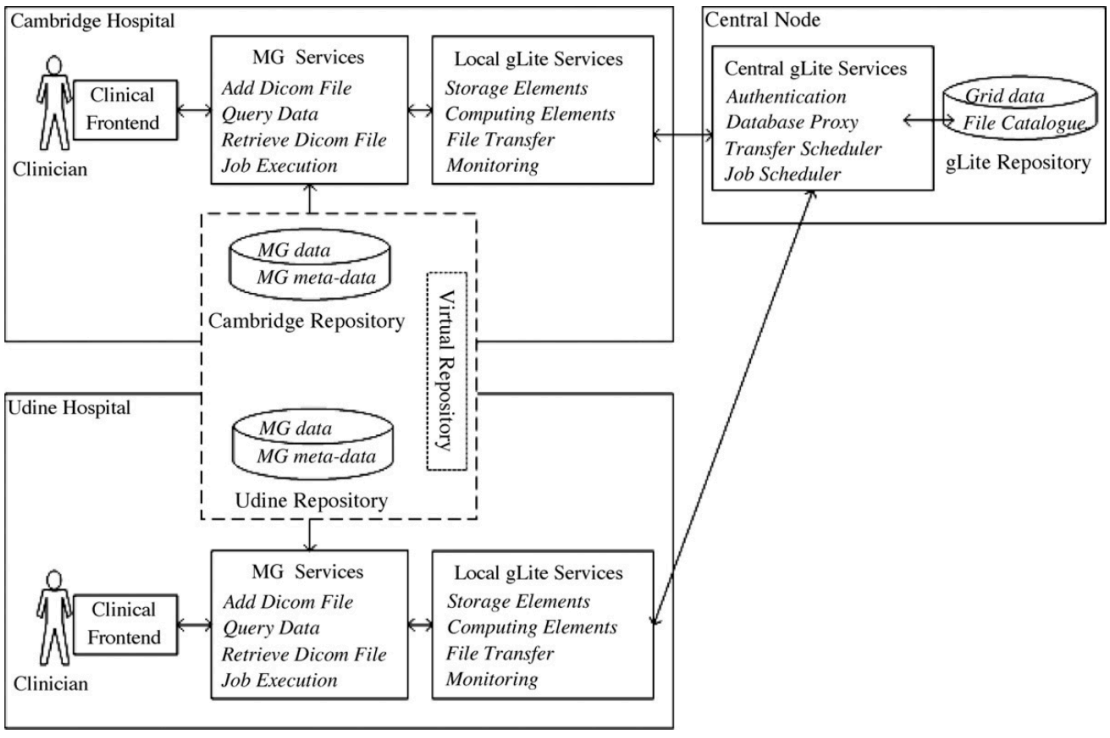


Fig. 4 – The MammoGrid services in the MGVO.

- *Epidemiological studies*: the effects of population variability, regional differences such as diet or body habitus and the relationship to mammographic density (a biomarker of breast cancer) which may be affected by such factors;
- *Support for radiologists*, in the form of tele-collaboration, second opinion, training and quality control of images.

During this clinical evaluation, the radiologists were able to view raw image data from each others' hospitals and were able to second-read Grid-resident mammograms and to separately annotate the images for combined diagnosis. This demonstrated the viability of distributed image analysis using the Grid and showed considerable promise for future health-based Grid applications. Despite the anticipated performance limitations that existing Grid software imposed on the system usage, the clinicians were able to discover new ways to collaborate using the virtual organization. These included the ability to perform queries over a virtual repository spanning data held in Addenbrookes and Udine hospitals and joint analyses thereof.

Following the 'Perform Radiological Analysis' use-case scenario shown in Fig. 2, clinicians defined their mammogram analysis in terms of queries they wished to be resolved across the collection of data repositories. Queries were categorized into simple queries (mainly against associated data stored in the database as simple attributes) and complex queries which required *derived data* to be interrogated or *special purpose algorithms* (e.g. for detection of abnormalities) to be executed on a (sub-)set of distributed images. One important result was that image and data distribution were transparent for radiologists, and hence complex queries were formulated and executed as if the associated data and images were locally resident. Queries were executed at the location where the relevant data resided, i.e. sub-queries were moved to the data, rather than large quantities of data being moved to the clinician, which could have been prohibitively expensive given the volume of the data involved. Fig. 5 illustrates how queries were handled in MammoGrid.

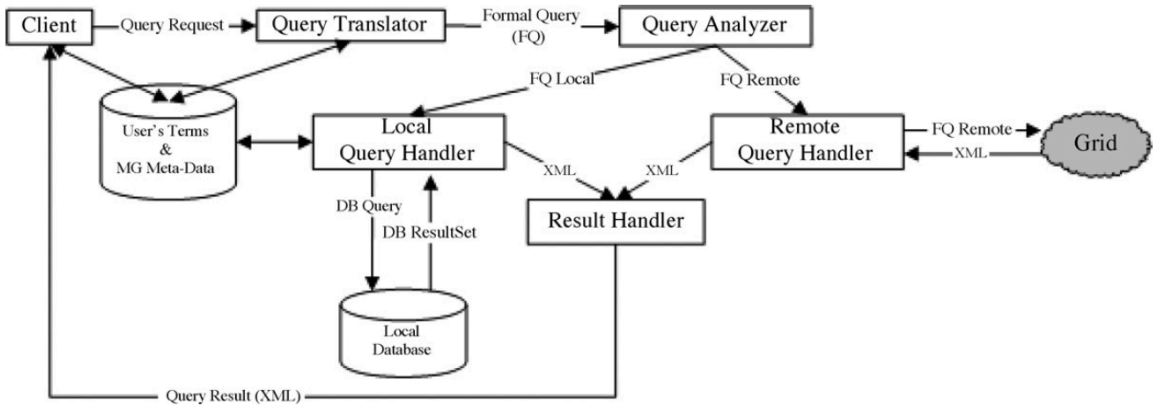


Fig. 5 – Clinical query handing in MammoGrid.

The Query Analyzer took a formal query representation and decomposed it into (a) a formal query for local processing, and (b) a formal query for remote processing. It then forwarded these decomposed queries to the Local Query Handler and the appropriate Remote Query Handler for the resolution of the request. The Local Query Handler generated the corresponding query language statements (e.g. SQL) in the query language of the associated Local DB. The result set was converted to XML and then routed to the Result Handler. The Remote Query Handler is a portal for propagating queries and their execution

results between sites. This handler forwarded the formal query for remote processing to the Query Analyzer of the remote site. The remote query result set was converted to XML and routed to the Result Handler. For detail see also [18]. At the time of writing this paper, the database is continuing to grow and currently holds.

The average processing time for the core services was: (1) add a 8 Mb DICOM file approximately 7 s; (2) retrieve a 8 Mb DICOM file from a remote site approximately 14s; (3) SMF workflow of ExecuteAlgorithm and Add around 200 s. The evaluation, carried out in mid 2005 on a *subset* of the currently available data revealed that for querying, see Table 2.

As a direct result of their satisfaction with the MammoGrid evaluation, clinicians continue in the process of scanning and annotating cases that contribute to several ongoing medical studies. These include (1) cancers versus control study: breast density study using SMF standard, (2) dose/density study: exploring the relationship between mammographic density, age, breast size and radiation dose, and (3) CADe and validation of SMF in association with CADe. These studies continue to show how health professionals can work together without co-locating. And, most importantly the collaborative approach pursued in MammoGrid has already identified new ways in which clinicians can work together using a common Grid-based repository which were hitherto not possible. For example, the use of the SMF [24] algorithm on data supported by MammoGrid and accessible to radiologists in Cambridge and Udine for the purposes of joint mammogram analysis has directly led to results being recently submitted to the European Radiology Journal [25].

In summary, during the final months of the project the clinicians have evaluated the MammoGrid prototype across two applications. *First*, the project has facilitated the use of the SMF software to measure breast density. The clinical project, designed jointly by Cambridge and Udine, explored the relationship between mammographic density, age, breast size, and radiation dose. In this project, breast density has been measured by SMF and compared with standard methods of visual assessment. Heights, weights, and mass indicators are used in an international comparison, but a richer dataset would be needed to study effects of lifestyle factors such as diet or HRT (Hormone Replacement Therapy) use between the two national populations. *Second*, the University of Udine led a project to validate the use of SMF in association with CADe from the CALMA project [9]. Cancers and benign lesions have been supplied from the clinical services of Udine and Cambridge to provide the benchmarking and the set of test cases. Cancer cases include women whose unaffected breast will serve the density study to provide cases for the CADe analysis from the affected side mammogram. MammoGrid has demonstrated that these new forms of clinical collaboration can be supported using the Grid [26].

Furthermore a strong collaboration has been established through the evaluation phase of the project between radiologists active in breast cancer research and academic computer scientists with expertise in the applications of grid computing. The success of the evaluation has led to interest from outside companies and hospitals, with one Spanish company, Maat GKnowledge [27], looking to deploy a commercial variant of the system in three hospitals of the Extremadura region in Spain. Maat GKnowledge aims to provide the Extremadura doctors with the ability to verify test results, to obtain second opinions and to make use of the clinical experience acquired by the hospitals involved in the MammoGrid project. They then aim to scale the system up and to expand it to other areas of Spain and then Europe. With the inclusion of new hospitals, it is proposed that the database will increase in coverage with clinical knowledge increasing in relevance and accuracy, and thus enabling larger and more refined epidemiological studies. Consequently, clinicians will be provided with a significant data set to serve better their investigations in the domain of cancer prevention, prediction and diagnoses. This is expected to result in improved research quality as well as improved citizen access to the latest

healthcare technologies. Further details of the clinical evaluation of MammoGrid and its exploitation plans can be found in [28].

Query	Cambridge	Udine	Number of images	Number of patients
By ID: Cambridge patient	2.654 s	2.563 s	8	1
By ID: Udine patient	2.844 s	3.225 s	16	1
All female	103 s	91 s	12,571	1,510
Age [50, 60] and image laterality = L	19.489 s	22.673 s	1,764	357

Table 2 – Data query performance of the MammoGrid prototype

7. Lessons learned

The nature of the project and its particular constraints of multi-disciplinarity, dispersed geographical development, large discrepancies in participants' domain knowledge (whether of software engineering techniques or of breast cancer screening practice), and the novelty of the Grid environment provide experiences from which other Grid-based medical informatics projects can benefit. We summarise below some of the main lessons that can be learned in this context.

First, the project was particularly fortunate with selection of its medical partners. In general, the medical environment is very risk-averse, conservative in nature and reluctant to adopt new technologies without significant evidence of tangible benefit. It is therefore important in Grid system prototyping to identify a suitable user community in which new technologies (such as Grid-resident medical databases) can be evaluated. In the case of MammoGrid we have had real commitment from the radiology community in the project's requirements definition and analysis, implementation and evaluation and this was crucial to the success of the project. The data samples used were of a sensitive nature and required both ethical clearance from participating institutions and anonymization of the data and even then strictly for only research use in the project. Realistically many ethical obstacles remain to be tackled before clinicians can share sensitive patient data between institutes, never mind across national boundaries.

Second, it has become clear from our experiences that Grid middleware technology itself is still evolving, and this suggests that there is a clear need for standardization to enable production-quality systems to be developed. Despite the availability of toolsets such as the Globus 4.0 [29] the development of applications that harness the power of the Grid, as yet, requires specialist skills and is thus costly in terms of manpower. Only with the arrival of stable middleware and packaged Grid services will the development of medical applications become viable.

Third, the performance of existing middleware is also somewhat limited; the MammoGrid project had therefore to circumvent some of the delivered Grid services to ensure adequate system performance for its prototype evaluation. For example, the database of medical images was taken out of the Grid software to provide adequate response for MammoGrid query handling. The EGEE [22] project is addressing these technological deficiencies and improved performance of the middleware should consequently be delivered in the coming years.

Fourth, Grid technology for medical informatics is still in its infancy and needs some proven examples of its applicability; MammoGrid is the first such exemplar in practice. Equally, awareness of Grid technology and its potential (and current limitations) must still be raised in the target user communities such as Health, Biomedicine, and more generally life sciences.

Fifth, the project has indicated that it is possible to use modelling techniques (such as use-cases from UML) in a widely distributed, multi-disciplinary software engineering problem domain, provided a very pragmatic approach is used, where adopting a certain modelling technique is, to some extent, independent from the software development life cycle model being applied. The MammoGrid project has benefited significantly in its coordination, communication and commitment by utilizing the use-case model as the *lingua franca* during user requirement analysis and system design rather than following the disciplines of RUP to the letter.

Sixth, the evolutionary approach to system development work packages has mitigated the effects of the project constraints of a highly dynamic research-oriented environment in which novices and specialists in software engineering have worked together even though they may have been geographically separated.

Further areas that might promote the use of rigorous software engineering disciplines in the design of Grid-based software services are that of model-driven engineering [30] and the use of architecture descriptions [31] as the basis for the generation of Grid-wide services. These aspects are, however, outside the scope of the current project.

8. Future directions and conclusions

The MammoGrid Virtual Oragnization (MGVO) is a distributed computing environment for harnessing the use of and access to massive amounts of mammography data across Europe. The MammoGrid approach used grid technologies, service-orientation, and database management techniques to federate distributed mammography databases allowing healthcare professionals to collaborate transparently without co-locating.

Furthermore, the MammoGrid project has delivered a Grid-enabled infrastructure which federates multiple mammogram databases across institutes. This permits clinicians to develop new common, collaborative and cooperative approaches to the analysis of mammography data. Using the MammoGrid they have been be able to quickly harness the use of massive amounts of medical image data to perform epidemiological studies, advanced image processing, radiographic education and ultimately, tele-diagnosis over communities of medical 'virtual organizations'. This was achieved through the use of Grid-aware services for managing (versions of) massively distributed files of mammograms, for handling the distributed execution of mammograms analysis software, for the development of imaging algorithms and for the sharing of resources between multiple collaborating medical centres.

In addition, the MammoGrid project has attracted attention as a paradigm for Grid-based imaging applications. While it has not solved all problems, the project has established an approach and a prototype platform for sharing medical data, especially images, across a Grid. In loose collaboration with a number of other European medical Grid projects (e.g. [7,11,12,32,33]), it is addressing the issues of informed consent and ethical approval, data protection, compliance with institutional, national and European regulations, and security [34]. In our view, the MammoGrid project paves the way for further research and development projects to meet the aims of the HealthGrid association [35] in the following respects:

- The identification of potential business models for medical Grid applications.

- Feedback to the Grid development community on the requirements of the pilot applications deployed by the European projects.
- Development of a systematic picture of the broad and specific requirements of physicians and other health workers when interacting with Grid applications.
- Dialogue with clinicians and those involved in medical research and Grid development to determine potential pilots.
- Interaction with clinicians and researchers to gain feedback from the pilots.
- Interaction with all relevant parties concerning legal and ethical issues identified by the pilots.
- Dissemination to the wider biomedical community on the outcome of the pilots.
- Interaction and exchange of results with similar groups worldwide.
- The formulation and specification of potential new applications in conjunction with the end user communities.

Recently, the Healthgrid association held its third annual international conference [36] at which the progress made in the spectrum of biomedical Grid projects was reviewed. The MammoGrid project provided important input to the ongoing debate on the role of Grids for (bio-)medical informatics. One very clear conclusion of the conference is the need to have greater involvement of the clinician community in the active use of medical informatics applications as demonstrated by MammoGrid. Finally, Grid computing is a promising distributed computing paradigm that can facilitate the management of federated medical images. This technology spans locations, organizations, architectures and has the potential to provide computing power, collaboration and information access to everyone connected to the Grid. Grid-based applications like the MammoGrid project benefit from this solution being based on open internet standards. These applications are potentially cross platform compatible, cross programming interoperable and widely accepted, deployed, and adopted.

Acknowledgements

The authors wish to thank their institute and the European Commission for support and to acknowledge the contribution of the following MammoGrid project members: Professor Roberto Amendolia (formerly of CERN), David Manset (at UWE for MammoGrid, now Maat GKnowledge), Dr Ruth Warren and Iqbal Warsi (Addenbrookes Hospital, Cambridge), Dr Chiara Del Frate and Professor Massimo Bazzocchi (Policlinico Universitario, Udine), Professor Sir Mike Brady and Chris Tromans (Oxford University), Martin Cordell (Mirada Solutions), Dr Piernicola Oliva (Sassari University), Drs Evelina Fantacci and Allessandra Retico (Pisa University) and Dr Jose Galvez and Dr Predrag Buncic (CERN).

References

[1] Foster, C. Kesselman, S. Tuecke, The anatomy of the grid: enabling scalable virtual organizations, Int. J. Supercomput. Appl. 15 (3) (2001).

[2] Global Grid Forum Data Access and Integration Services (GGF-DAIS) Working Group. http://www.cs.man.ac.uk/grid-db/.

[3] J.S. Silva, M.J. Ball, Prognosis for Year 2013, Int. J. Med. Infor. 66 (1–3) (2002) 45–49 (elsevier publishers).

[4] P. Ruotsalainen, A cross-platform model for secure electronic health record communication, Int. J. Med. Infor. 73 (3) (2004) 291–295

[5] S.R. Amendolia, M. Brady, R. McClatchey, M. Mulet-Parada, M. Odeh, T. Solomonides, MammoGrid: large-scale distributed mammogram analysis, Stud. Health Technol. Inform. 95 (2003) 194–199 (IOS Press: ISBN 1-58603-347-6).

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

[7] M. Brady, M. Gavaghan, A. Simpson, M. Mulet-Parada, R. Highnam, eDiamond: a grid-enabled federated database of annotated mammgrams, in: F. Berman (Ed.), Grid Computing: Making the Global Infrastructure a Reality, Wiley, 2003.

[8] National Digital Mammography Archive. See http://nscp01.physics.upenn.edu/ndma/.

[9] I. De Mitri, The MAGIC-5 project: medical applications on a grid infrastructure connection, Stud. Health Technol. Inform. 112 (2005) 157–166 (IOS Press: ISBN 1-58603-510-x, ISSN 0926-9630).

[10] [10] M. Ellisman, et al., BIRN: Biomedical Informatics Research Network, Stud. Health Technol. Inform. 112 (2005) 100–109 (IOS Press: ISBN ISBN 1-58603-510-x, ISSN 0926-9630).

[11] [11] J.L. Oliveira, et al., DiseaseCard: a web-based tool for the collaborative integration of genetic and medical information. Biological and medical data analysis, in: Proceedings of the 5th International Symposium, ISBMDA 2004,Barcelona,Spain,November18–19,2004.

[12] [12] I. Blanquer, et al., Clinical Decision Support Systems (CDSS) in GRID environments studies, Health Technol. Infor. 112 (2005) 80–89 (IOS Press: ISBN 1-58603-510-X, ISSN 0926-9630).

[13] [13] The Rational Unified Process Model. See http://www.rational.com.

[14] [14] M. Odeh, T. Hauer, R. McClatchey, A. Solomonides, Use-case driven approach in requirements engineering: the MammoGrid project., in: M.H. Hamza (Ed.), Proceedings of the 7th IASTED International Conference on Software Engineering and Applications, Marina del Rey, CA, USA: ACTA Press; November, 2003, pp. 562–567.

[15] [15] G. Booch, J. Rumbaugh, I. Jacobson, The Unified Modeling Language User Guide, Wesley Longman, Reading, MA, 1999.

[16] [16] Service-Oriented Architecture. See: http://www.service-architecture.com.

[17] [17] I. Foster, C. Kesselman, J. Nick, S. Tuecke, The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Open Grid Service Infrastructure WG, In: Global Grid Forum, June 22, 2002. See http://www.globus.org/research/papers/ogsa.pdf.

[18] [18] S. Amendolia, F. Estrella, C. del Frate, J. Galvez, W. Hassan, T. Hauer, et al., Deployment of a Grid-based medical imaging application, Stud. Health Technol. Infor. 112 (2005) 59–69 (IOS Press: ISBN 1-58603-510-X, ISSN 0926-9630).

[19] [19] B. Boehm, A spiral model of software development and enhancement, IEEE Comput. 21 (5) (1988) 61–72.

[20] P. Saiz, L. Aphecetche, P. Buncic, R. Piskac, J.-E. Revsbech, V. Sego, AliEn – ALICE environment on the GRID, Nucl. Instrum. Methods A 502 (2003) 437–440 (last accessed October 6, 2004) http://alien.cern.ch.

[21] gLite: Lightweight Middleware for Grid Computing. http://glite.web.cern.ch/glite.

[22] The Information Societies Technology project. In: EU-EGEE, EU Contract IST-2003-508833, 2003. http://www.eu-egee.org/ (last accessed October 6, 2004).

[23] DICOM: Digital Imaging and Communications in Medicine. http://medical.nema.org.

[24] R. Highnam, M. Brady, B. Shepstone, A representation for mammographic image processing, Med. Image Anal. 1 (1) (1996) 1–18.

[25] R. Warren, et al., A comparison of some anthropometric parameters between an Italian and a UK population: 'proof of principle' of a European project using MammoGrid. To be published by European Radiology, Springer-Verlag publishers, 2006.

[26] R. Warren, et al., A Prototype Distributed Mammographic Database for Europe. To be published by European Radiology, Springer-Verlag publishers.

[27] The Maat GKnowledge company. See http://www.maat-g.com.

[28] C. del Frate, et al., Final results from and exploitation plans for MammoGrid, in: Accepted for Publication at the 4th International HealthGrid Conference, Valencia, Spain, 2006.

[29] The Globus Toolkit 4.0. http://www.globus.org/toolkit/.

[30] A. Kleppe, J. Warmer, W. Bast, MDA Explained: The Model Driven Architecture™: Practice and Promise, Addison Wesley Professional (2003), and Rick Kazman, Steven G. Woods, S. Jeromy Carrière, Requirements for Integrating Software Architecture and Reengineering Models: CORUM II. Software Engineering Institute, Carnegie Mellon University.

[31] F. Oquendo, S. Cimpan, H. Verjus, The ArchWare ADL: Definition of the Abstract Syntax and Formal Semantics. ARCHWARE European RTD Project IST-2001-32360. See also http://www.arch-ware.org.

[32] S. Benkner, et al., GEMSS – Grid Infrastructure for Medical Service Provision. In Methods of Information In Medicine, vol. 44, no. 2, ISSN 0026-1270, Schattauer GMBH publishers.

[33] N. Shadbolt, P. Lewis, S. Dasmahapatra, D. Dupplaw, B. Hu, H. Lewis, MIAKT: Combining Grid and Web Services for Collaborative Medical Decision Making. UK e-Science All Hands Meeting (AHM) 2004, Nottingham, UK, 2004.

[34] V. Breton, K. Dean, T. Solomonides, The HealthGrid White Paper in 2005. In: From Grid to Healthgrid. Studies in Health Technology and Informatics, vol. 112, ISBN 1-58603-510-X, ISSN 0926-9630 IOS Press. See also http://www.heathgrid.org.

[35] S. Nørager, Y. Paindaveine, The HealthGrid Terms of Reference, EU Report Version 1.0, 20th September 2002.

[36] From Grid to HealthGrid. T. Solomonides, R. McClatchey, V. Breton, Y. Legre, S. Norager (Eds.), Studies in Health Technology and Informatics, vol. 112, ISBN 1-58603-510-X, ISSN 0926-9630 IOS Press (Proceedings the 3rd HealthGrid International Conference (HG'05), Oxford, UK, April 2005).

Added in proof

Additional acknowledgement

Thanks are extended to Dr. Jan Talmon and the Editorial Board for constructive comments in the preparation of this manuscript.

Summary points

What was known before the study

- The potential benefits of Grid-based applications deployment in distributed systems.
- The advantages of delivering stable, maintainable systems by using software engineering principles.
- The need for standardization of medical image quality and for consultation between radiologists to improve diagnosis.
- The nature of service-oriented architectures and research into their use in Grid systems.

What the study has added to our knowledge

- The practical issues involved in deploying Grid- based medical imaging applications including soft- ware architectures, clinician commitment and communication and the need for standardization.
- The advantages of following the use-case modelling techniques in engineering Grid medical solutions.
- The tangible benefits of using the SMF algorithm coupled with a service-oriented architecture for distributed mammogram analysis.
- The limitations of existing Grid technologies and approaches for mitigating these limitations in future healthgrid implementations.

A retrospective study of paediatric health and development following pre-implantation genetic diagnosis and screening

Mark Olive
University of the West of England, Bristol, UK
mark2.olive@uwe.ac.uk

Alison Lashwood
Centre for PGD, Guy's Hospital, London, UK
alison.lashwood@gstt.nhs.uk

Tony Solomonides
Biosciences, University of Exeter, UK
tony.solomonides@gmail.com

Abstract

Pre-implantation genetic diagnosis and screening (PGD and PGS) are treatments for patients that have (or are carriers of) an inherited genetic disorder, or who have had a history of miscarriage, problems with embryo implantation, etc. Often conducted alongside assisted reproductive technologies (ART), a number of embryos are produced, and the DNA and chromosomes of each are tested for various disorders by removing one or two cells for analysis.

A retrospective cross-sectional study looking at the health and development of children born following PGD and PGS is now underway, aided by an online system developed by the EuroPGDcode project. Data has been collected from a number of ART/PGD centres worldwide, and has been entered into this system. A number of complex queries have been constructed to interrogate the data; although retrospective and not case controlled, indications are that the birth abnormality rate is low at 1.42%. However the special care requirements of PGD infants was 22.9% and the incidence of health problems after birth was 22.3%.

In addition to statistical analysis of the data, a number of cases of particular interest have been identified. The online system provides the facility for the full details of these cases to be exported in a specially designed XML format for further analysis.

1. Introduction

Pre-implantation genetic diagnosis (PGD) has developed considerably in the last 20 years. First introduced for sexing embryos in the case of an X-linked genetic disorder in 1990 [1], this was subsequently followed by a live birth after PGD for the monogenic disorders cystic fibrosis [2] and Duchenne muscular dystrophy [3]. Munné et al [4] then reported the first case of using fluorescent in-situ hybridization (FISH) for a reciprocal chromosome translocation. Internationally, PGD is available for over 200 single gene and chromosomal disorders [5] and the technology diversified in 1999 to include pre-implantation genetic screening (PGS). PGS is aimed at improving the outcome of in vitro fertilisation (IVF) in sub-fertile couples, and also for the selection of human leucocyte antigen matched embryos (HLA) as a source of therapeutic stem cells for sick siblings [7]. Couples undertaking PGD are generally fertile and do not need assisted reproductive technologies (ART) to conceive. Many care for children with special needs and a some have medical problems themselves as a consequence of the genetic disorder that affects them and puts their offspring at risk. It is important therefore that these factors are considered during preparation for PGD.

PGD (and in particular PGS) remain somewhat controversial [6], and while evidence suggests that human embryo development in vitro is not affected by biopsy, confirmation can only be obtained by long term follow up. Another concern is what impact PGD might have on the subsequent health and development of children born after undergoing it. Long term follow up of such children has been recommended since the introduction of PGD, but as the number of children born remains small, international collaboration and standardised data collection is essential.

The European Society of Human Reproduction and Embryology (ESHRE) PGD consortium has completed a total of 10 data collection exercises since 1997, gathering referral data, biopsy, FISH and other data. The analysis of these has in turn lead to a number of high quality journal publications, as well as the creation of best practice guidelines. The impact of PGD/PGS on the success of IVF is understood better now than ever before, but many scientific, technical and ethico-legal questions remain. Among these is the question of the impact of PGD/PGS on the health (or morbidity) and development of children born after undergoing one of these procedures.

This paper describes the outcome of a study of children born after embryo blastocyst biopsy, with both PGD and PGS cycles included.

Follow up of babies born by PGD has been recommended since the early days of the technology [8,9,10,11] and is included in current guidelines. IVF and ICSI (intracytoplasmic sperm injections) have been available for over 30 years and follow up of children born has been long term and case controlled. As a result, 3 meta analyses [12,13,14] and a case controlled study of 3000 ICSI vs IVF infants [15] showed a relative risk of major abnormality of 1.24 in ART babies compared with spontaneously conceived infants. A major abnormality is considered one that has medical or social consequences and occurs in 2-3% of live births and 5% of 5 year olds [16]. Longer term studies on ICSI/IVF babies showed a relative increased risk of abnormality: ICSI (2.77) and IVF (1.8) [17].

An increase in imprinting disorders such as Beckwith Wiedemann (BWS), Angelmans and retinoblastoma has been recognised in ART babies [18]. A recent study [19] looking at over 15,000 babies born after IVF/ICSI concluded that there was 4.24% major abnormality rate with a 5 fold increase in BWS (0.04% vs 0.007%) and a 4 fold increase in retinoblastoma (0.03% vs 0.006%) over the norm.

In addition to the technology used in assisted reproductive technology (ART), PGD requires additional micro-manipulation procedures which may have an impact on paediatric outcome. Whilst embryos are created using standard ART, testing embryos requires embryonic tissue from biopsy at either polar body, blastomere or trophectoderm stages [5]. The preferred fertilisation method for PGD cases that require the use of polymerase chain reaction (PCR) is ICSI [5] to avoid the risk of contamination. IVF is acceptable for use when FISH analysis is to be used for chromosome rearrangements and embryo sex determination for X-linked disorders [20].

Data available from two studies reviewing 480 PGD babies [21] and adding new data to the previous study resulting in 576 PGD babies [22] reported a major abnormality rate of 1.6 and 1.9% respectively. In the latest ESHRE PGD Consortium report [23] which amalgamates nine data sets representing outcomes of PGD up to 2008, a total of 5047 babies have been born, with outcome data available on 4021 (79.7%). The total number of babies with minor and major malformations in this group was 154/4021 (3.83%); 84 major and 74 minor abnormalities (with some babies having more than 1). The incidence of neonatal complications was 10.3% and there was a total of 45 neonatal deaths (1.1%); 12 singletons, 30 twins and 3 triplets; further in-depth analysis of this data is now being undertaken. The abnormalities that were reported varied in severity and ranged from significant cardiac abnormalities to mild syndactyly. These outcomes are similar to those reported in the IVF/ICSI population [15] where neonatal complications occurred in 9% of cases, with 1% of cases resulting in neonatal death. Liebaers et al. [24] reported the first prospective case controlled study comparing 581 PGD birth outcomes with 2889 IVF and ICSI babies at 2 months of age. The rate of major abnormalities was not statistically different at 2.13 and 3.38% respectively, but the rate of perinatal death was higher in the PGD multiple pregnancies at 11.73 and 2.54% respectively.

Although limited, a few studies have investigated long-term growth and development of children born following PGD/PGS. Two studies with 49 and 102 children [25,26] have compared 2 year old PGD with ICSI and normally conceived children. These found that, although PGD babies were of lower birth weight, their linear growth compared well with normally conceived children, and the PGD children had the same incidence of congenital abnormality and childhood ill health as the control groups.

2. Factors influencing outcome

Maternal health. Women who have health-related problems associated with their genetic diagnosis should be referred to an obstetric or other relevant physician to discuss the impact of treatment and pregnancy. For example, women who are affected with myotonic dystrophy should be assessed before anaesthesia as they have an increased risk of arrhythmias, prolonged recovery from the anaesthesia and a risk of developing malignant hyperpyrexia [27]. Their myotonia often deteriorates during pregnancy and they are prone to obstetric complications including prolonged labour, placenta praevia and postpartum haemorrhage.

Multiple pregnancy. ART is associated with an increased risk of multiple births [28]. Attempts to increase the chance of pregnancy in PGD cycles by replacing more than one embryo have led to a high multiple pregnancy rate [5]. Babies born from multiple births have a higher risk of prematurity, low birth weight, neonatal mortality and neurological disability [29]. The issue of the number of embryos for transfer needs careful discussion in cases where there is a choice. A multiple pregnancy may have both clinical and social implications for a couple. Couples requesting PGD often also care for children with disabilities and special needs as a result of the genetic condition within the family, and a multiple pregnancy would be a significant additional burden. In addition, as part of recommended best practice, confirmatory

prenatal testing is advised following a successful PGD pregnancy [20] and although possible, prenatal testing is more complex in a multiple pregnancy.

The frequency of multiple pregnancies in PGD couples has generally reduced, from 25% in 1999-2003 to 20% in 2003–2004 (although in 2009, ESHRE PGD data still reported a multiple birth rate of 27%). Some studies in both PGD and ART are now demonstrating improvements in live birth outcome using single-blastocyst transfer [30]. The use of single-embryo transfer, especially in women younger than 36 years is resulting in fewer multiple pregnancies, without a reduction in the overall delivery rate. It has been demonstrated that selection of single embryos for transfer, with cryopreservation of surplus unaffected embryos, maintains a good pregnancy rate while reducing multiple births [30]. The implantation rate using cryopreserved biopsied PGD blastocysts is comparable with that obtained after using non-biopsied frozen IVF blastocysts. This is an important step towards encouraging couples to opt, where clinically indicated, for single-embryo transfer, which may have additional benefits for PGD couples.

Type of embryo biopsy. Several biopsy techniques are employed with blastomere biopsy used in 90% of cases [5]. Polar body biopsy can be used to assess maternal genotype or karyotype only. Trophectoderm biopsy, which provides a larger tissue sample on day 5, is used in only a few centres but there is no evidence that this increases treatment success rates [31].

Number of cells biopsied. One or two cells may be taken at embryo biopsy. A blastomere may not be representative of the embryo as a whole and mosaicism is known to occur in up to 50% of cleavage-stage embryos, but some embryos, initially mosaic for aneuploid cell lines, self correct with increasing cell division; the abnormal cell line is selected against and the resulting embryo becomes euploid. Centres that only use embryos where the two cells biopsied are concordant with a normal result will exclude a higher proportion of embryos owing to higher rates of false positive results. While the reliability of test results is of paramount importance, the aim is to identify sufficient embryos with transferable results. There is evidence to suggest that two-cell biopsy might reduce the number of embryos available for transfer even though the predictive value of such results would be higher [32].

3. The ESHRE PGD questionnaire

In 2005 the ESHRE PGD Consortium agreed to support a retrospective data collection of the babies born following PGD/PGS. 57 centres (members of ESHRE) were sent a questionnaire asking about their current provision for paediatric follow up and inviting them to participate in a retrospective data collection study. Ethical approval was obtained through the UK National Research Ethics Service[1] (NRES) and 6 centres that met both the NREC requirements and the inclusion criteria for the study were included. Written parental consent was obtained. Participating centres were required to have at least 10 PGD/PGS live births before 31/10/07, a patient population that could read English or a translated questionnaire (French, Flemish, Portuguese, Spanish, Czech or Turkish), and personnel to locally administer the questionnaire.

The questionnaires were comprised of 4 sections. Part 1 was completed by the participating centre and included technical information about the cycle, embryo biopsy, the reason for PGD/PGS and date of embryo transfer. Parts 2-4 were completed by the parents of the PGD/PGS children and included data relating to the pregnancy, birth, health and development of the children, and parental demographics.

The questionnaires were allocated a centre number and a case study number to ensure anonymity, and were returned to the local study administrator. The questionnaires were then translated back into English and forwarded to the central study coordinator at Guy's Hospital, London (UK).

Data was included from 400 questionnaires with 6 participating centres. 41 questionnaires were excluded, mainly because they related to children born outside the timeframe of data capture.

Table 1. Questionnaire statistics by centre

Centre	Sent	Returned	Excl.	Analysed
England	112	54 (45.5%)	3	51
Spain	265	249 (70.9%)	5	243
Czech Republic	137	92 (67.1%)	32	60
Portugal	10	10 (100%)		10
Turkey	16	16 (100%)		16
Belgium	27	21 (70.4%)	1	20

4. EuroPGDcode

During the collection of this PGD/PGS data, the EuroPGDcode project arose through an initiative from the ESHRE Classification of Infertility Taskforce (ECIT) for a common nomenclature, and was funded by a European Union grant (Executive Agency for Health and Consumers, contract A800103 2007-2009).

The methods of PGD/PGS data collection over the past 14 years have varied considerably between

[1] See http://nres.npsa.nhs.uk/

different laboratories and clinics. While the range of data that must be collected has long been established, the means of collection has varied from applications interfacing hospital systems through spreadsheets, text files and proprietary databases, complicating the process of combining these into a single, coherent data set for analysis. A major part of the EuroPGDcode project was the creation of a new system for the collection, storage and analysis of PGD/PGS questionnaire data (and other anticipated or derived data items from concurrent or future data collections), based on open source components and made available worldwide via a web-based interface.

An online system was also required due to the geographically distributed nature of the work; questionnaire data had to be interpreted and entered at ART/PGD centres, queries were constructed and modified by informaticians at the University of the West of England, then executed by PGD specialists at Guy's Hospital. Further local analysis also had to be facilitated, with input from medical statisticians.

Another goal of the project was the creation of a prototype XML structure for representing the PGD/PGS cycle data, various details relating to the babies that were born (such as their health and development), parental demographics, etc.

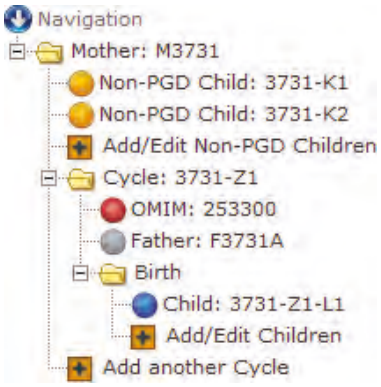


Figure 1. A tree view of a PGD/PGS cycle record

A relational database was constructed, with a web-based front end for the input, viewing and analysis of data. As the data from a questionnaire is entered, a graphical, interactive tree is created to visualise the record - this also corresponds to the structure of the XML document that can be generated.

Queries can be constructed through an integral query editor, or (for expert users) entered directly in SQL. These can then be 'bookmarked' and stored for specific or all users. A variety of complex queries have been constructed, and the output of these can be combined and interrogated by additional tools locally.

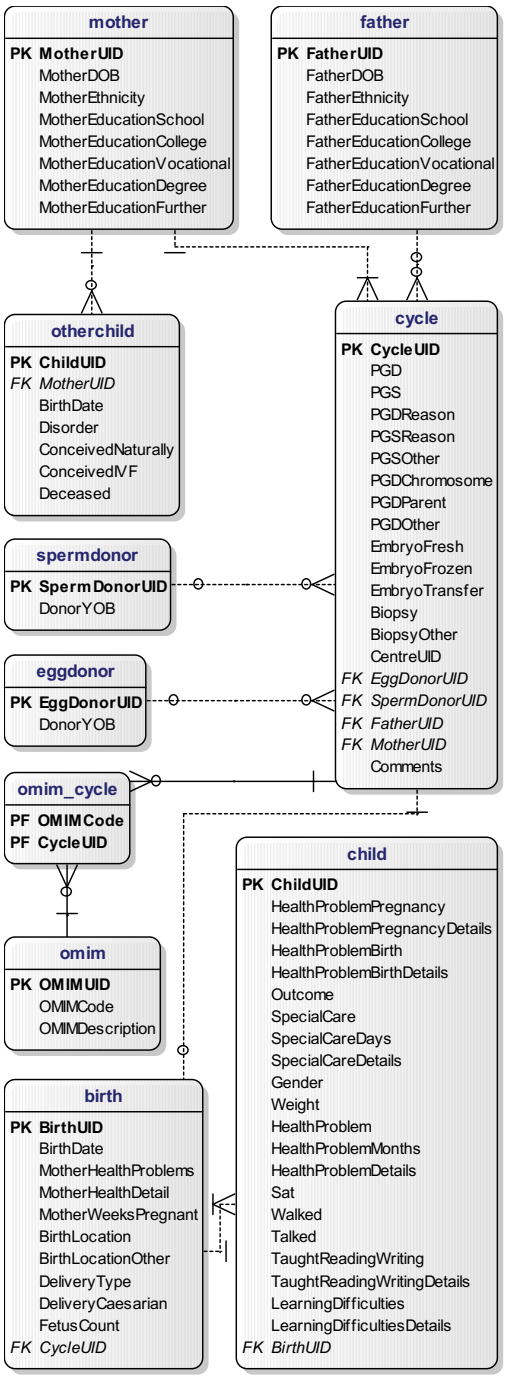


Figure 2. The relational database structure

In addition to statistical analyses, the full records of individuals of interest, which can include cross-referenced data gathered from multiple questionnaires, can be exported as XML documents. These validate against a schema, and the XML format defined serves

as a first step towards the ESHRE goal of creating a standardised format for the exchange of infertility data.

Figure 3. The initial user interface screen

5. Results

Cycle details. 112 cycles were undertaken for PGD (28%), 287 for PGS (71.7%) and one for both. One cell biopsy was used in 54 PGD cycles and in 125 PGS cycles; two cell biopsy was used in 52 PGD cycles and 162 PGS cycles and blastocyst biopsy was done in 6 PGD cycles; the PGD/PGS case was recorded as a one cell biopsy. Fresh embryos were used in 107 PGD cycles and 284 PGS and 1 cycle that was for both; frozen embryos were used in 5 PGD and 3 PGS cycles. This compares with the proportion of fresh/frozen cycles in the ESHRE Data 10 collection [23]. The date of embryo transfer was checked for validity against the date of birth by the online system. 11 mothers in this cohort had babies born from more than one PGD cycle. The number of cycles for any given parent did not exceed 2, and the Mean number of cycles was 1.03 for mothers and 1.02 for fathers (in 12 cases, no father details were recorded). No PGD cycles were undertaken for HLA matching or for social sex selection. 52 cases were because of a chromosomal abnormality, 46 for a monogenic disorder and 14 cycles were sex selection for X-linked inheritance. The parental origin of the monogenic or chromosomal disorder was recorded in 106/112 cycles; maternal in 48 cycles, paternal in 39 cycles and from both in 19 cycles. The PGD/PGS case was for a monogenic disorder and raised maternal age (RMA). Indications for PGS included 98 cycles for repeated ART failure, 58 for miscarriage after ART, 52 for RMA and infertility and 5 for RMA only, 7 for recurrent miscarriage, 7 for previous aneuploidy, 43 due to male factor infertility, 8 with egg donation and 18 did not complete.

Parental demographics. The range of birth years for mothers was 1958 to 1981 (aged 28 to 51 at the time of data collection), with the Mean being 1970 (age 41). Day and month of birth were not input in order to preserve patient anonymity. For fathers, the range was 1936 to 1981, with a Mean of 1967. The vast majority of mothers and fathers self identified as Caucasian at 98.46 and 98.64% respectively; 0.26% of mothers were African (0.54% of fathers), and 0.27% of fathers were Asian (no mothers).

Pregnancy and birth. Details of maternal and fetal health in pregnancy were collected for 400 deliveries. 297 (75.25%) women reported no problems. 4 women had ovarian hyper-stimulation syndrome, 2 of whom had related ovarian torsion, 13 experienced bleeding, 5 placenta praevia, 15 had symptoms associated with onset of premature labour, 11 gestational diabetes, 4 pre-eclampsia, 4 non related raised blood pressure and 4 with renal problems. The Mean gestation of delivery varied depending on the parity; for singletons, twins and triplets the (estimated) Mean gestation was 38.7, 35.9 and 33 weeks respectively. Mode of delivery indicated that 145 (36.34%) births were delivered spontaneously, with 38 (9.52%) requiring assisted delivery (ventouse or forceps) and 216 (54.13%) by caesarian section. The incidence of either planned or emergency caesarian section was higher in multiple births at 76.4% and 100% in twins and triplets respectively compared with 47.4% in singleton deliveries. 82.1% of babies were born $\geq$ 36 weeks gestation; 89.5% singletons, 58% twins; 14.5% of deliveries were between 30-36 weeks gestation and 3.2% $\leq$ 30 weeks.

Health of babies at delivery. A total of 487/494 (98.58%) babies were reported by parents as having no congenital abnormalities at birth. 7 babies (1.42%) had 9 congenital abnormalities including tongue tie, talipes, clicky hip, undescended testes, cardiac abnormalities and hypospadias. Special care was required for 3/7 babies with abnormalities. Neonatal problems were recorded in 129/494 babies (26%) with 113 babies (87.6%); 44 singletons (14.3%), 63 twins (44.9%) and 6 triplets (100%) requiring special care. The Mean number of days spent in special care was 18.9 and the most common reasons for special care were low birth weight, prematurity and respiratory problems. Since birth, 378/494 (76.7%) babies had no major illness or operations; 115 (23.3%) recorded illnesses, 71 were singletons, 43 twins and 1 triplet. The range of

illnesses can be broadly categorized into respiratory, ENT, gastrointestinal, genito-urinary, orthopaedic, renal, neurological, ophthalmic, infection, and allergy related problems. The incidence of most illnesses/problems was confined to 1 or 2 individual cases. More frequent respiratory illnesses were asthma (9), pneumonia (8), and bronchitis/bronchiolitis (7). Surgery was done in 8 cases for adenotonsillectomy and 3 children had hearing loss. Six children had general allergies with 5 being lactose intolerant. Genitourinary problems included 5 children with inguinal hernia, 3 with undescended testes and 2 with spermatic cysts.

Development after birth. At data collection, the Mean age of children was 5, and the range was 1 to 11 years old. Parents were asked to give the age at which their child reached certain motor milestones. For being able to sit, entires were made for 441/494 children; 425 children (99.37%) sat before 10 months with 16 (3.54%) sitting after 10 months. Parents recorded the age of walking unaided in 473/494 cases; 468 babies (98.94%) were walking unaided by 20 months whilst 5 (1.05%) walked after 20 months. 11 children who had started school and had begun formal teaching had learning difficulties. Of these, 4 were twins, 2 triplets and 5 singletons. Birth weights ranged from 0.815 kg to 3.26 kg (Mean 2.1916). Three singletons were born at > 38 weeks gestation, 2 singleton and 3 twins at > 36 weeks gestation, 2 triplets at 33-36 weeks gestation and 1 twin at 25 weeks gestation.

6. Conclusion

Centres from England, Spain, the Czech Republic, Portugal, Turkey and Belgium returned data on 400 deliveries; 112 PGD cycles, 288 PGS cycles and one cycle with both. These resulted in 494 live babies born between December 1999 and October 2007. There were 308 (62.3%) singletons, 180 (36.5%) twins and 6 (1.2%) triplets. Mean birth weights were 3.17, 2.41 and 1.67 KGs respectively. 77.2% were born at ≥ 36 weeks, 19.3% between 30-36 weeks and 3.45% ≤ 30 weeks gestation. Neonatal problems occurred in 129 babies (26%), 16 of which required no special care. 113 babies (22.9%) required special care (Mean 18.9 days). Nine abnormalities were recorded in 7 babies (1.42%) including tongue tie, cardiac anomalies, hip dysplasia, talipes, hypospadias and undescended testes. Since birth 378 (76.7%) babies were recorded as having no health problems. The rate of recorded health problems for singletons and twins was 23.1% and 23.9%, Mean ages 15.2 and 16.9 months respectively.

The analysis of the data collected is ongoing. Several cases of special interest have been identified (primarily the children with abnormalities); the full records of these children and their siblings can be exported as XML records for further analysis. Complementary data from other sources, such as electronic health records (EHRs) and electronic integrated care pathways (eICPs), can then be gathered in order to build up a full case history; in previous papers we have described a method that involves a generic ICP ontology, and a method to identify portions of an EHR that correspond to periods of interest [33, 34].

Another area being looked at are the differences in the health and development of children born after PGD when compared with PGS, if any.

This study has demonstrated that completed parental questionnaires can provide valuable information about the long term health and development of PGD/PGS children. Various methods, including an integrated searchable database of OMIM (Online Mendelian Inheritance in Man) codes for genetic disorders, were employed by the online system to standardise responses, and the web-based nature of the system can facilitate data capture from a wide demographic population. Although retrospective and not case controlled the data indicates that the birth abnormality rate is low at 1.42%. However, 22.9% of PGD infants required special care, and the incidence of health problems after birth was 22.3%.

7. Acknowledgements

The authors would like to thank Gary Harton (Reprogenetics, USA), Dr Joanne Traeger-Synodinos (University of Athens, Greece) and Dr Frances Flinter (Guy's Hospital, London, UK), authors of preceding papers. EuroPGDcode was funded in part by the EU Executive Agency for Health and Consumers (EAHC), and supported by the ESHRE PGD Consortium.

8. References

[1] Handyside A H, Kontogianni E H, Hardy K & Winston R. "Pregnancies from biopsied human preimplantation embryos sexed by Y specific DNA amplification". *Nature*, Macmillan Publishers, 1990;224:768-70.

[2] Handyside A H, Lesko J G, Tarin J J, Winston R & Hughes M R. "Birth of a normal girl after in vitro fertilization and preimplantation diagnostic testing for cystic fibrosis". *N Engl J Med*, Massachusetts Medical Society, 1992;327:905–9.

[3] Liu J, Lissens W, Van Broeckhoven C, Löfgren A, Camus M, Liebaers I & Van Steirteghem A. "Normal pregnancy after preimplantation DNA diagnosis of a dystrophin gene deletion". *Prenat Diagn* 1995;15:351-8

[4] Munné S, Morrison L, Fung J, Márquez C, Weier U, Bahçe M et al. "Spontaneous Abortions Are Reduced

After Preconception Diagnosis of Translocations". *J Assist Reprod Genet*, Plenum Publishing, 1998;15(5):290-6

[5] Goossens V, Harton G, Moutou C, Traeger-Synodinos J, Van Rij M, Harper J C. "ESHRE PGD Consortium data collection IX: cycles from January to December 2006 with pregnancy follow-up to October 2007". *Hum Reprod*, 2009;24:1786–810

[6] Kalfoglou A L, Scott J & Hudson K. "PGD patients and providers attitudes to the use and regulation of PGD". *Reprod Biomed Online*, 2005;11(4):486-96

[7] Verlinsky Y, Rechitsky S, Schoolcraft W, Strom C, Kuliev A. "Preimplantation genetic diagnosis for Fanconi anemia combined with HLA matching". *JAMA*, 2001;285:3130–3

[8] Verlinsky Y, Cohen J, Munne S, Gianaroli L, Simpson J L, Ferraretti A P et al. "Over a decade of experience with preimplantation genetic diagnosis: a multicenter report". *Fertil Steril*, 1994;82:292–4

[9] Simpson J L & Liebaers I. "Assessing congenital anomalies after preimplantation genetic diagnosis". *J Assist Reprod Genet*, 1996;13(2):170-6

[10] Baruch S, Kaufman D & Hudson K L. "Genetic testing of embryos: practices and perspectives of U.S. IVF clinics". *Fertil Steril*, 2008;89:1053–8

[11] Human Genetics Commission. "Making Babies: reproductive decisions and genetic technologies". *Department of Health*, 2006

[12] Hansen M et al. "Assisted reproductive technologies and the risk of birth defects: a systematic review". *Hum Reprod*, 2005;20:328-38

[13] Lie R T et al. "Birth defects in children conceived by ICSI compared to children conceived by other IVF methods: a meta analysis". *Int J Epidemiol.* 2005;34:696-701

[14] Rimm A A et al. "A meta analysis of controlled studies comparing major malformation rates in IVF and ICSI infants with naturally conceived children". *J Assist Reprod Genet*. 2004;21:437-43

[15] Bonduelle M, Liebaers I, Deketelaere V, Derde M P, Camus M, Devroey P et al. "Neonatal data on a cohort of 2889 infants born after ICSI (1991–1999) and of 1995 infants born after IVF (1983–1999)". *Hum Reprod*, 2002;17:671–94

[16] Stevenson R E. "Malformations and Related Anomalies". In *Human Malformations and Related Anomalies, 2nd Edition*, Oxford University Press, 2006

[17] Bonduelle M et al. "A multi centre cohort study of the physical health of 5 year old children conceived after ICSI, in vitro fertilisation and natural conception". *Hum Reprod*, 2005;20:413-419

[18] Sutcliffe A et al. "Assisted reproductive therapies and imprinting disorders- a preliminary British survey". *Hum Reprod*, 2006;21:1009-1011

[19] Viot G B, Epelboin S & Olivennes F. "Is there an increased risk of congenital malformations after Assisted Reproductive Technologies?". *European Human Genetics Conference*, Sweden, June 2010

[20] Thornhill A R, de Die-Smulders C E, Geraedts J P, Harper J C, Harton G L, Lavery S A et al. "ESHRE PGD Consortium Best practice guidelines for clinical preimplantation genetic diagnosis (PGD) and preimplantation genetic screening (PGS)". *Hum Reprod*, 2005;20:35–48

[21] Tur Kaspa I, Horwitz Z, Ginsberg J, Cieslak S, Rechitsky Y & Verlinsky Y. "Clinical Outcome of PGD". *Fertil Steril*, 2005;84(suppl.1):599

[22] Horwitz A et al. "No increased birth defects in 576 liveborn babies after PGD". *Reprod Biomed Online*, 2005;10(Suppl. 2):0–50

[23] Harper J C, Coonen E, De Ryke M, Harton G, Moutou C, Pehlivan T et al. "ESHRE PGD consortium data collection X: cycles from January to December 2007 with pregnancy follow up to October 2008". *Human Reproduction*, 2010;25(11):2685-2707

[24] Liebaers I, Desmyttere S, Verpoest W, De Rycke M, Staessen C, Sermon K et al. "Report on a consecutive series of 581 children born after blastomere biopsy for preimplantation genetic diagnosis". *Hum Reprod*, 2010;25(1):275-82

[25] Banerjee I, Shevlin M, Taranissi M, Thornhill A, Abdalla H, Ozturk O et al. "Health of children conceived after PGD: a preliminary outcome study". *Reprod Biomed Online*, 2008;16:376–81

[26] Desmyttere S, De Schepper J, Nekkebroeck J, De Vos A, De Rycke M, Staessen C et al. "Two-year auxological and medical outcome of singletons born after embryo biopsy applied in preimplantation genetic diagnosis or preimplantation genetic screening". *Hum Reprod*, 2009;24:470–6

[27] de Die-Smulders C E, Höweler C J, Thijs C, Mirandolle J F, Anten H B, Smeets H J et al. "Age and causes of death in adult-onset myotonic dystrophy". *Brain*, 1998;121(Pt 8):1557–63

[28] Braude P. "One Child at a Time: Reducing Multiple Births after IVF". *Human Fertilisation and Embryology Authority*, October 2006

[29] Sutcliffe A & Ludwig M. "Outcome of assisted reproduction". *Lancet*, 2007;370:351-59

[30] Khalaf Y, El Toukhy T, Coomarasamy A, Kamal A, Bolton V & Braude P. "Selective single blastocyst transfer reduces the multiple pregnancy rate and increases pregnancy rates: a pre and postintervention study". *BJOG*, 2008;115:385–90

[31] McArthur S J, Leigh D, Marshall J T, de Boer K A & Jansen R P. "Pregnancies and live births after trophectoderm biopsy and preimplantation genetic testing of human blastocysts". *Fertil Steril*, 2005;84;1628–36

[32] Los F J, Van Opstal D & van de Berg C. "The development of cytogenetically normal, abnormal and mosaic embryos: a theoretical model". *Hum Reprod Update*, 2004;10:79–94

[33] Olive M, Lashwood A & Solomonides T. "Care pathway records with ontologies: potential uses in medical research and health care". *Intl J Care Pathw* 2011;15:15-17

[34] Olive M, Lashwood A & Solomonides T. "Care pathway records and variance data: enabling research through the use of ontologies". *CBMS 2010*, Perth, Australia

The Healthgrid White Paper

Vincent BRETON[1], Kevin DEAN[2] and Tony SOLOMONIDES[3], Editors
(on behalf of the Healthgrid White Paper collaboration*)

([1]) CNRS-IN2P3, LPC
([2]) Internet Business Solutions Group, Cisco Systems
([3])CEMS, University of the West of England, Bristol

Abstract

Over the last four years, a community of researchers working on Grid and High Performance Computing technologies began 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 16–17, 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 29–30 , 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**			
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

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 [1–3]. 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 roundthe-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 pseudonymization 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 semiautomatic 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 24 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 selfdiscoverable 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 lifecritical 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, postprocessing 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, multicentre - 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 nonrigid 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.

Mammogram 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, timevarying 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 treatmentplanning 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-theart 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 peroperative 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 robotarm 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 opensource data.

The problems range from knowledge-representation and integration, to distributed systems search and access control, to data mining and knowledge management, to realtime 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.

Figure. Concrete Structure of a Grid for Rare Diseases.

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.

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 (panEuropean) 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-centre approaches that take into account environmental and other factors
- Availability of evidence-based medicine
- Sophisticated statistical analysis and modelling
- 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 defence 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 defences 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 metaanalyses.

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.

Biosocial Hierarchy	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 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]).

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.

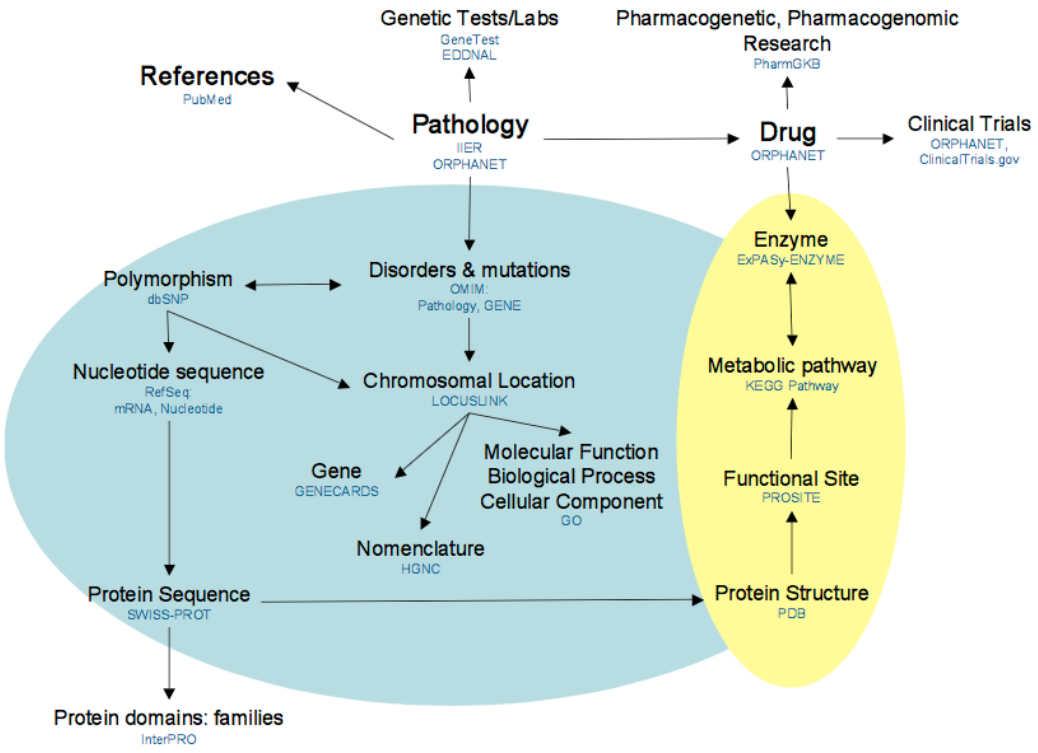


Figure 1. A conceptual framework for the study of genetic disorders.

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 (Inter-PRO) 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 highlevel 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 Treat-ments, 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 pro-ject” 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 341342.
[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 privac.y: 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, MayJune 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 (*)
European Commission -DG Information Society and Media, Office BU 31 6/41, B-1049 Brussels sofie.norager@cec.eu.int
(*) 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

SHARE road map for healthgrids: Methodology

Mark Olive[a], Hanene Rahmouni[a], Tony Solomonides[a],*, Vincent Breton[b], Yannick Legré[c], Ignacio Blanquer[d], Vicente Hernandez[d]

[a] *Biomedical Informatics Group, Bristol Institute of Technology, UWE Bristol, Coldharbour Lane, Bristol BS16 1QY, UK*
[b] *Laboratoire de Physique Corpusculaire, CNRS-IN2P3, Campus de Cézeaux, 63177 Aubière Cedex, France*
[c] *HealthGrid, 36 rue Charles de Montesquieu, 63430 Pont du château, France*
[d] *Universidad Politécnica de Valencia, Camino de Vera s/n, E-46022 Valencia, Spain*

(*) *Corresponding author*. Tel.: +44 1173283149; E-mail address: Tony.Solomonides@uwe.ac.uk

A B S T R A C T

The SHARE[1] project (http://www.eu-share.org) was asked to identify the key developments needed to achieve wide adoption and deployment of healthgrids throughout Europe. The project was asked to organise these as milestones on a road map, so that all technical advances, social actions, economic investments and ethical or legal initiatives necessary for healthgrids would be seen together in a single coherent document. The full road map includes an extensive analysis of several case studies exploring their technical requirements, full discussion of the ethical, legal, social and economic issues which may impede early deployment, and concludes with an attempt to reconcile the tensions between technological developments and regulatory frameworks. This paper has been restricted to the technical aspects of the project.

SHARE built on the work of the 'HealthGrid' initiative so we begin by, reviewing work carried out in various European healthgrid projects and report on joint work with numerous European collaborators. Following many successful healthgrid projects, HealthGrid published a 'White Paper' which establishes the foundations, potential scope and prospects of an approach to health informatics based on a grid infrastructure. The White Paper demonstrates the ways in which the HealthGrid approach supports many modern trends in medicine and healthcare, such as evidence-based practice, integration across levels, from molecules and cells, through tissues and organs to the whole person and community, and the promise of individualised healthcare. SHARE was funded by the European Commission to define a research roadmap for a 'Healthgrid for Europe', to be seen as the preferred infrastructure for biomedical and healthcare projects in the European Research Area.

1. Introduction

A 'grid' – not *the* grid – is now understood to mean an Internet-like infrastructure which extends the concept of the Internet in several significant ways:

- like the Internet, a grid would provide access to information services but in addition would provide pooled storage,
- processing power and collaboration in so-called 'virtual organisations' (VOs);
- use of a grid will be reciprocal—while a user subscribes and takes advantage of services provided by a grid, the user's resources are pooled and are available to all grid subscribers;
- the process is transparent—the grid allocates resources and provides an interface to services which give the appearance that the user is accessing just one powerful machine.

Major IT companies have agreed to develop web services as the technology to enable the deployment of services on the Internet. It has been also adopted by the Open Grid Forum which is the acknowledged body to propose and develop standards for grid technology.

Moreover, web service technology provides the bridge between the grid world and the Semantic Web which is about common formats for the interchange of data and about language for recording how the data relates to real world objects. Although many current grid infrastructures do not offer a web service interface to their services, we will concentrate our 'state of the art' on web services because it is the relevant technology for the future. We will then go on to discuss the status of existing grid infrastructures, the technologies they use and the services they offer.

The initial idea behind web services was to enable the World Wide Web increasingly to support real applications and a means of communication among them. Thus, a set of standards and protocols have been proposed to allow interaction between distant machines over a network. These interactions are made possible through the use of standardised interfaces which describe the available operations in a service, the nature and form of messages exchanged (requests and responses), and the physical location of the service on the network. A language, web service description language (WSDL) has been devised to describe such interfaces.

One of the subtle advantages of a so-called Service Oriented Architecture (SOA) is that it leads to loosely coupled components that may be substituted by better services so long as they comply with the interface specification. Thus in a grid, services can be offered in a way that approximates an ideal marketplace, a market with near perfect information.

Grid technology has been identified as one of the key technologies to enable and support the 'European Research Area' The impact of this concept is expected to reach far beyond eScience, to eBusiness, eGovernment, and eHealth. However, a major challenge is to take the technology out of the laboratory to the citizen. A *healthgrid* is an environment in which data of medical interest can be stored and made easily available to different actors in the healthcare system, physicians, allied professions, healthcare centres, administrators and, of course, patients and citizens in general. Such an environment has to offer all appropriate guarantees in terms of data protection, respect for ethics and observance of regulations; it has to support the notion of 'duty of care' and may have to deal with 'freedom of information issues'. Working across member states, it may have to support negotiation and policy bridging.

Early grid projects, while encompassing potential applications to the life sciences, did not address the specificities of an e-infrastructure for health, such as the deployment of grid nodes in clinical centres and in healthcare administrations, the connection of individual physicians to the grid and the strict regulations ruling the access to personal data. However, a community of researchers did emerge with an awareness of these issues and an interest in tackling them.

2. The HealthGrid initiative

Many pioneering projects in the application of grid technologies to health and biomedical research have been completed, and the technology to address high-level requirements in a grid environment has been under development and making good progress. Because these projects had a finite lifetime and the ambition for healthgrids required a sustained effort over a much longer period, and besides because there was an obvious need for these projects to cross-fertilise, the 'HealthGrid initiative', represented by the HealthGrid association (http://www.healthgrid.org), was initiated to bring the necessary long-term continuity. Its goal has been to encourage and support collaboration between autonomous projects in such a way as to ensure that requirements really are met, and that the wheel, so to speak, is not re-invented repeatedly at the expense of other necessary work.

Writing about the HealthGrid initiative very soon after its inception, this community identified a number of objectives [1]:

- Identification of potential business models for medical grid applications.
- Feedback to the grid development community on the requirements of the pilot applications deployed by the European projects.
- Development of a systematic picture of the broad and specific requirements of physicians and other health workers when interacting with grid applications.
- Dialogue with clinicians and those involved in medical research and grid development to determine potential pilots.
- Interaction with clinicians and researchers to gain feedback from the pilots.
- Interaction with all relevant parties concerning legal and ethical issues identified by the pilots.
- Dissemination to the wider biomedical community on the outcome of the pilots.
- Interaction and exchange of results with similar groups worldwide.
- The formulation and specification of potential new applications in conjunction with the end user communities.

Apart from research, where the value of grid computing is well established, a healthgrid may be deployed to support the full range of healthcare activities, from screening and diagnosis, through treatment planning, to epidemiology and public health. For example, anticipating that population trends, air pollution and global warming may lead, through extremes of heat, to increased risks to the elderly, we may deploy a monitoring service to track conditions and medical episodes in hot summers. Patients' medical data would have to be stored in a local database in each healthcare centre. These databases would have to be federated and would essentially share the same logical model. Secure access to data would have to be granted to authorised individuals or services, which would therefore need to be authenticated: only partial views of local data would be available to external services and patient data would have to be anonymised or at least pseudonymised.

Given the source of the concept of grid in the physical sciences, many of these requirements were not a central concern to grid developers in general. Indeed, even today, when these requirements have been fed through to the middleware services community, they are not a priority for mainstream grid developers. Thus HealthGrid has been actively involved in the definition of requirements relevant to the development and deployment of grids for health and was among the first to identify the need for a specialist middleware layer, between the generic grid infrastructure and middleware and the medical or health applications.

Among data related requirements, the need for suitable access to biological and medical image data arose in several early projects, but for the most part these are present in other fields of application also. Looking to security requirements, most of these are special to the medical field: anonymous or private login to public and private databases; guaranteed privacy, including anonymisation, pseudonymisation and encryption as necessary; legal requirements, especially in relation to data protection, and dynamic negotiation of security and trust policies while applications remain live. Most administrative requirements are common to medicine and eScience, although the flexibility of 'virtual grids', i.e. the ability to define sub-grids with restrictions on data storage and data access and also on computing power, is more obviously required in healthcare. Medical applications also require access to small data subsets, like image slices and model geometry. At the (batch) job level, medical applications need an understanding of job failure and the means to retrieve the situation.

Requirements of this kind have been addressed in a number of projects. Among the examples we shall mention are MammoGrid [3], Health-e-Child [4] and Virtual Physiological Human (*VPH*) [5]. The former constructed a grid

database of standardised mammogram files and associated patient data to enable radiologists in Cambridge, UK, and Udine, Italy, to request second opinion and computer-aided detection services. The project also enabled a further study of an epidemiological nature on breast density as a risk factor. In the larger Health-e-Child 'integrated project' radiologists are working with oncologists, cardiologists and rheumatologists to identify early imaging signs of conditions that may have a strong genetic component. The knowledge thus obtained should reduce the need for genetic maps to be obtained on an indiscriminate basis. *VPH* requires collaborative research to create a methodological and technological framework that once established will enable the investigation of the human body as a single complex system. *VPH* will provide a framework within which observations made in laboratories and hospitals may be collected, catalogued, organised, shared and combined in any possible way. It should also allow experts to collaboratively analyse observations and develop systemic hypotheses that involve the knowledge of multiple scientific disciplines, and to interconnect predictive models defined at different scale, with different methods, and with different levels of detail, into systemic networks that provide concretisation to those verifiable systemic hypotheses.

In another European project, Wide In Silico Docking On Malaria (*WISDOM*) tens of millions of molecular docking experiments have been carried out to help identify potential antigens for the malaria parasite. The experiment uses large-scale virtual screening techniques to select molecular fragments for further investigation in the development of pharmaceuticals for neglected diseases. The economic dynamics in this area are telling: only about 1% of drugs developed in the last quarter century have been aimed at tropical diseases, and yet these are major killers in the third world, with mortality in excess of 14 million per annum. One of the relevant applications which are well suited to the grid is molecule 'docking' and this has been successfully applied, e.g. to proteins of the malaria parasite. This work continues with increasingly promising results and has now been extended to other diseases also, most notably avian influenza strain H5N1.

Meanwhile, the results of several major studies of the interface between bioinformatics and medical informatics had been published with a remarkable promise of synergy between the two disciplines, leading to what had already begun to be referred to as 'personalised medicine' [6,7]. From the point of view of HealthGrid, this made clear the need to unify the field and to put its various elements in perspective: how would they – improved evidence bases, imaging, genetic information, pharmacology, epidemiology – fit together, what was their relative importance in the unfolding programme of work?

3. The White Paper: from grid to HealthGrid

Thus, the next step for the HealthGrid community was to try to systematise the concepts, requirements, scope and possibilities of grid technology in the life sciences. The White Paper [8] defines the concept of a healthgrid more precisely than before:

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.

The ultimate goal for eHealth in Europe would be the creation of a single healthgrid, i.e. a grid comprising all eHealth resources, incorporating a 'principle of subsidiarity' of independent nodes of the healthgrid as a means of implementing all the legal, ethical, regulatory and negotiation requirements. We may anticipate, however, the development path to proceed through specific healthgrids with perhaps rudimentary inter-grid interaction and interoperational capabilities. Thus, we may identify a need to map future research and advice on research policy, so as to bring diverse initiatives to the point of convergence.

Healthgrid applications address both individualised healthcare – diagnosis and treatment – and epidemiology with a view to public health. Individualised healthcare is improved by the efficient and secure combination of immediate availability of personal clinical information and widespread availability of advanced services for diagnosis

and therapy. Epidemiology healthgrids combine the information from a wide population to extract knowledge that can lead to the discovery of new correlations between symptoms, diseases, genetic features and other clinical data. With this broad range of application in mind, the issues below are identified as key features of our analysis.

• *Business case, trust and continuity issues*: healthgrids are data- and collaboration-grids, but health institutions are reluctant to let information flow outside institutional boundaries. Large-scale deployment, which would make an attractive business opportunity, requires 'security', using the word inclusively, to be scaled up to a very high level of confidence. Data storage remains the responsibility of a hospital, yet business opportunities may arise from data sharing and processing applications; this degree of federation of databases introduces additional complexity.

• *Biomedical issues*: 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 explore sources of diseases, risk populations, evolution of diseases or suitable proteins to fight against specific diseases. 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, including those for complex and multimedia information.

• *Security issues*: These flow naturally from the nature of medical data and from business requirements. Security in grid infrastructures is currently adequate for research platforms, but not for real healthcare applications. Biomedical information must be carefully managed to maintain its integrity and to avoid privacy leakages. Secure transmission must be complemented with secure storage, with strictly controlled authenticated and authorised access. Automatic pseudo/anonymisation is necessary for a 'production' healthgrid.

• *Management issues*: The central concept of a 'virtual organisation' (VO) at the heart of eScience, which gave rise to grids, is very apt for Healthgrid, but additional flexibility is needed to structure and to control VOs in the large, including, for example, the meta-level of a VO of VOs. The management of resources has to be more precise and dynamic, depending on many criteria such as urgency, medical protocols, users' authorisation or other administrative policies.

Current examples of healthgrids span a wide range. At one end, we find the classic 'high-throughput' approach of numerical simulation of organs obtained from a patients' data and used to aid understanding or to improve the design of medical devices; this leads to patient-customised approaches at least at research-level in areas such as radiotherapy, craniofacial surgery and neurosurgery [9]. Other healthgrids deal with large-scale information processing, such as medical imaging. Breast cancer imaging has been the focus of several successful grid projects and eHealth projects suitable for migration to a Healthgrid. 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 [10]. 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 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.

A very helpful account of the state of the art at the end of 2006 was given by Groen and Goldstein [11] with an expanded version promised in a text book to follow. Among the many initiatives and links they mention, it is useful to single out the Open Grid Forum's Life Sciences Grid—Research Group (cf [12]). This emphasis reflects SHARE's anticipation that healthgrids will be deployed in the service of biomedical research before they are adopted in healthcare as such.

4. The SHARE project: from White Paper to Road Map

In the White Paper, the HealthGrid community expressed its commitment to engage with and support modern trends in medical practice, especially 'evidence-based medicine' as an integrative principle, to be applied across the dimensions of individual through to public health, diagnosis through treatment to prevention, from molecules through cells, tissues and organs to individuals and populations. In order to do this, it had to address the question how to collect, organise, and distribute the 'evidence'; this might be 'gold standard' evidence, i.e. peer reviewed knowledge from published research, or it might be more tentative, yet to be confirmed knowledge from practice, and, in addition, would entail knowledge of the individual patient as a whole person. The community also had to address the issues of law, regulation and ethics, and issues about crossing legal and cultural boundaries, finding ways to express these in terms that translate to technology—security, trust, encryption, pseudonymisation. Then it had to consider how the services of the Healthgrid middleware would satisfy these requirements; and, if it was to succeed in the real world, how to make the business case for Healthgrid to hard-pressed health services across Europe while they are struggling with their own modernisation programmes [2].

The vision of health that informs the thinking of the White Paper and the work of HealthGrid since its publication has been defined in the 'Action Plan for a European e-Health Area' [13] as follows:

> "... the application of information and communications technologies across the whole range of functions that affect the health sector. e-Health tools or 'solutions' include products, systems and services that go beyond simply Internet-based applications. They include tools for both health authorities and professionals as well as personalised health systems for patients and citizens. Examples include health information networks, electronic health records, telemedicine services, personal wearable and portable communicable systems, health portals, and many other information and communication technology-based tools assisting prevention, diagnosis, treatment, health monitoring, and lifestyle management."

The 'vertical integration' implicit in this visionary statement can be translated into more concrete terms by mapping it to its human subjects, their pathologies and the implicit disciplines. The relationships between the different ontological and epistemological levels and the various modalities of data have been captured by Fernando Martin-Sánchez et al. (cf [6]) in the schematic diagram (Fig. 1).

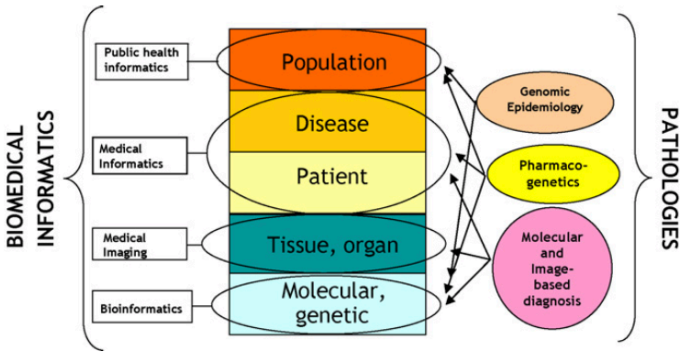


Fig. 1 – Levels of biosocial organisation, disciplines, pathologies and informatics (courtesy Fernando Martín Sánchez).

In the light of the White Paper and its impact, the EC has funded a 'specific support action' project, SHARE, to explore exactly what it would mean to realise the vision of the White Paper, investigate the issues that arise and define a roadmap for research and technology which would lead to wide deployment and adoption of healthgrids in the next 10 years. To be more precise, based on the assumption that Healthgrid will be the infrastructure of choice for biomedical and eHealth applications within the next 10 years, the two objectives of the project are

- a roadmap for research and technology to allow a wide deployment and adoption of healthgrids both in the shorter term (3–5 years) and in the longer term (up to 10 years); and
- a complementary and integrated roadmap for e-Health research and technology development (RTD) policy relating to grid deployment, as a basis for improving coordination among funding bodies, health policy makers and leaders of grid initiatives, avoiding legislative barriers and other foreseeable obstacles.

Thus the project had to address the questions, What research and development needs to be done now? and What are the right initiatives in eHealth RTD policy relating to grid deployment?—with all that implies in terms of coordination of strategy, programme funding and support for innovation.

In summary, therefore, the project has sought to define a comprehensive and detailed European research and development roadmap, covering both technology and policy aspects, to guide and promote beneficial EU-wide uptake of healthgrid technologies, and their applications into health research and into healthcare service provision.

5. Technical road map: step one

SHARE has defined a technical roadmap and a separate ethical, legal and socio-economic (ELSE) issue conceptual map, both informed by experiences in Healthgrid projects, merging them into an initial integrated roadmap. We present here the technical road map component, which is comprised of seven milestones representing the key technological, deployment and standard challenges for future healthgrid research. Two technical milestones have been defined for the implementation, development and testing of healthgrid services, two milestones as examples of grid standards that must be defined for the medical domain, and three deployment milestones increasing in complexity and scope (Fig. 2).

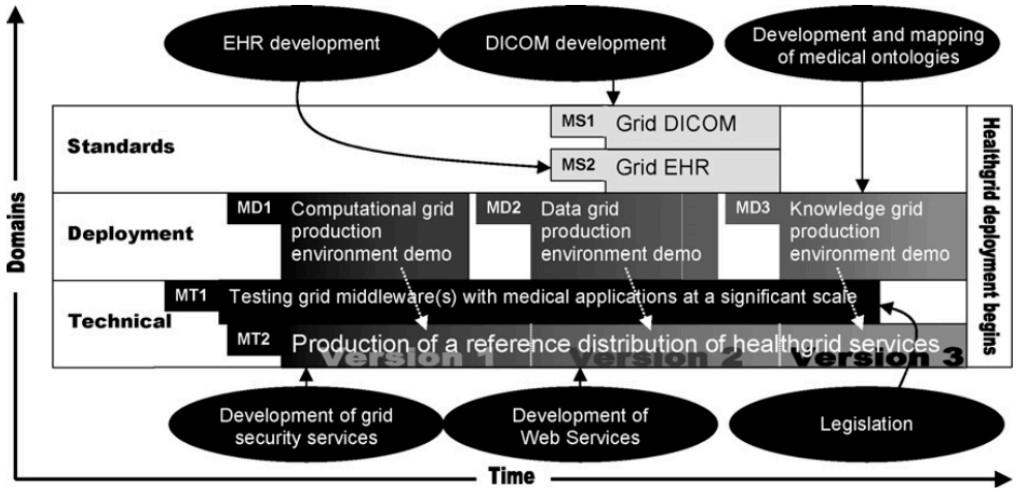


Fig. 2 – A diagrammatic representation of the technical road map, with milestones and external influences.

MT1. Before deployment can begin, the first step should be to begin the testing of grid middleware(s) with medical applications for scalability and robustness. This should begin at an early stage, as any deficiencies in this

area will only hamper deployment and the development of a reference distribution. It is anticipated that this will be an ongoing activity, with different generations of grid operating systems offering newer, faster and more stable capabilities.

A key issue for this milestone will be the robustness of grid solutions based on web services. Scalability, particularly regarding medical applications, is still a concern for grid middleware based on the Open Grid Services Architecture (OGSA) such as GT4 [14] and GRIA [15]. Middleware such as gLite and Unicore on the other hand have been deployed on large-scale infrastructures in Europe and have demonstrated their scalability and robustness, but are still awaiting migration to web services.

MD1. The first deployment step will be the rollout of a computational grid production environment demonstrator for medical research. This would seem to be an achievable goal in the reasonably near future given that there have already been successful deployments of computational grid applications (the WISDOM data challenges, for example) on general purpose grid infrastructures such as DEISA [16] and EGEE [17]. Apart from such 'innovative medicine' applications, there are examples of computational grid applications in healthcare delivery: vascular surgery, radiotherapy planning, optimal drug delivery, monitoring the effects of a heat wave or modelling an epidemic of antibiotic resistant bacteria in a nosocomial setting. However, convincing healthcare management of the benefits of deploying a computational grid on a hospital or clinic IT infrastructure, which would not be composed of dedicated grid nodes and may already be working near capacity, is a real concern. Many medical centres may simply not have the necessary bandwidth or storage capabilities to make best use of grid technology, or may not have appropriate equipment to capture data in digital form. The management and configuration of a grid is rather complex and may require significant investment in manpower and training.

Installation of grid nodes behind hospital firewalls is incompatible with the present security model, which required inbound connections. Secure services for data management are under development that could allow firewall rules to be relaxed, enabling connections to grid infrastructures. New architectures and designs should be defined that will minimise the volume of data leaving hospital borders. Ideally, the grid node would be located outside the firewall with only anonymised or pseudonymised data being stored on the grid. However, even with the most stringent pseudonymisation and de-identification techniques there is still some risk of unauthorised re-identification by a person with sufficient knowledge from other sources. There are therefore legal and ethical implications of storing even anonymised personal data on the grid, and further investigation will be required to determine if this is possible with current national and European policies and legislation.

MT2. Starting with MD1 and ending when production deployment begins, this milestone is the development of a reference distribution of grid services, using standard web service technology and allowing the secured manipulation of distributed data. An important consideration for this milestone will be the level of security required when dealing with distributed medical data.

Key grid infrastructures such as EGEE have been developed for scientific communities (such as high energy physics), and while providing an appropriate level of complexity to those using and administering grid nodes in that environment, the installation and management of grid notes in hospitals and medical research centres will need to be considerably more user friendly, with an easy-to-use user interface.

Emerging web services technologies such as the WS-addressing standard and the WSRF specification must be respected and promoted in the development of healthgrids, and already have several widely used implementations. They provide a standard and interoperable way for implementing state in web services, and separate state information from operations. Industry partners have recognised this, but have concerns about the level of tool support for these standards [18].

The EPSRC funded IBHIS project [19] found that web service description languages and registries are not yet mature enough, particularly WSDL, which describes how to access web services, and UDDI, which provides a registry

for service discovery. For example, the UDDI registry was only searchable using keywords, but to identify all relevant data sources ontology-based searching would be required. These technologies are currently not flexible enough, cannot be used for semantic queries or descriptions (e.g. a description of the function a service provides, or the meaning of parameter names) or non-functional descriptions such as quality of service and performance levels. The development of WSDL and UDDI is ongoing, with OWL-S extensions to UDDI to facilitate semantic searching, upcoming versions of UDDI promising to address other limitations, and WSDL 2.0 promising to support semantic descriptions and include non-functional requirements.

Considerable development is occurring in the area of web services:

• A Negotiation Description Language (NDL) can be used to construct supply chains of services to achieve the desired goal. NDL could be extended to 'many to many' negotiations, where all participants in the chain can obtain information about suppliers and react accordingly. NDLs can also contain information other than technical conformance, such as non-functional legal and cost conformances.

• An ISO-9126 compliant automated just-in-time service quality assessment tool was being developed by members involved with IBHIS to select the 'best' service. This assessment is not trivial as many of the quality characteristics being weighed can conflict, so that a rise in one would result in the decline of another (e.g. as usability increases, security may decrease).

Security is not an option but strict requirement for healthgrids. Security is an issue at all technical levels: networks need to provide protocols for secure data transfer, the grid infrastructure needs to provide secure mechanisms for access, authentication, and authorisation, as well as sites for secure data storage. The grid operating system needs to provide access control to individual files stored on the grid. High level services need to properly manage legal issues related to the protection of medical data.

The security offered by the existing infrastructures does not yet allow the manipulation of medical data. Important progress is being made in terms of fine-grained access control and data encryption. Some prototype services are under development but they are not yet fully deployed. A specific security feature implemented by hospitals is a restriction in the access to the Internet. Installation of grid elements behind the hospital firewall is incompatible with the present security model where outbound connection is not allowed through this firewall. Ideally, the grid node would be located outside the firewall with only anonymised or pseudonymised data being stored on the grid. However, given the legal and ethical implications of storing any personal data on the grid, even if it is fully anonymised, further investigation will be required to determine if this is possible with current national and European policies and legislation. For example, even with the most stringent pseudonymisation and de-identification techniques there is still some risk of unauthorised re-identification by a person with sufficient information from other sources.

Revocation of credentials and how to provide temporary access to data is still an open issue, and an important one for healthgrids. There are a number of situations where users would temporarily require access to data that they would not normally have access to, such as a visiting expert to a breast cancer unit being shown an unusual case. Certificate authorisation servers have been developed in both 'pull' mode (VOLDAP, GridSite LDAP, and VOMS-httpd), in which sites periodically pull a list of valid members from a central service, and 'push' mode (VOMS attribute certificates), in which users obtain a short-lived attribute certificate that they present to sites to prove their membership. However, both of these would leave a window where revoked or expired credentials could be used to gain unauthorised access. Several healthgrid projects have suggested that the data itself should have a 'lifetime'—users with temporary access should not be able to access the data (or a copy of the data) once their credentials have expired. This would, for example, be in accord with the fifth principle of the UK Data Protection Act, 1998; a form of Digital Rights Management (DRM) has been proposed as the means to provide this restriction.

MD2. Although several prototype data grids for medical research have been demonstrated by healthgrid projects, developing and maintaining a production quality data grid will require a number of issues relating to the distributed storage of medical data to be resolved. In European grid infrastructures, the distributed storage of medical images has been hampered by the limited data management services available, and so the continuation of improvements in this area will be important for the adoption of grids by the medical community. High speed links between data providers and consumers will be a prerequisite, particularly given the high volume of data predicted.

Many legal and ethical issues will need to be resolved, such as the ownership of patient data, ethical control of information, the patient's right to access or be informed about data that concerns them, as well as local, national, and European level legislation governing the use of patent data and IT.

Another important concern for this milestone will be the integration of heterogeneous data from multiple sources. While mechanisms for data integration have been demonstrated by previous projects, biomedical data can be exceptionally varied including images with associated metadata and free form text or hand written notes from patient records. There is also the issue of how to deal with missing, inaccurate or obsolete data.

MS1 and MS2. The use of computer-based tools for clinical research has led to the definition of standards for the exchange of data in many areas. However, such standards are in many cases not universal, with different disciplines and countries adopting different standards. The exchange of data between bioinformatics and medical informatics is an area where standards are particularly limited.

Medical imaging is an exemplary case, in which the adoption of DICOM for the acquisition, connection and storage of medical images has been accepted worldwide. Medical records are another area where standardisation would have clear benefits, with HL7 being the favoured standard for the exchange of data. However, previous standards such as CEN/TC251 EN13606 focused more on the storage and structuring of clinical records and have prevented a wider uptake of HL7. The adoption of both DICOM and HL7 has increased due to initiatives such as IHE (Integrating the Healthcare Enterprise), which promotes the coordinated use of DICOM and HL7 by publishing best practice guidelines. A particularly important consideration for both of these standards is their compatibility with grid technologies, and how they could be implemented on a healthgrid. Both DICOM and HL7 developers are just starting to study the interface between their standards and web services technology.

An observation on collaboration through grids. Many successful examples of computational and data grids have also entailed a significant extra dimension, that of collaboration. This has been found to be of value in both research and model healthcare delivery applications. An exemplary case is that of MammoGrid [20], a project in which Italian and English radiologists were able to collaborate in diagnostic studies and, even more successfully, in epidemiology [21]. The project sought to replicate the 'workflow' of a real breast screening department by means of communication and exchange of selected images through a VPN-like secure grid. In this project, a proprietary standard, SMFTM was used to normalise mammograms so that they could be compared as like-with-like. Day to day activities were imitated as much as possible. Thus, where further tests are conducted as part of the recall process, readers can see what their decisions amounted to in practice. Double reading accords opportunities for readers to compare their decisions with those of their colleagues for the same set of cases (double reading is not blind, and readers may make notes for each other). Co-located readers can walk down the hall to ask a colleague what they meant by this annotation or whether they had seen and ignored or missed this feature. Commonly, disagreements between readers are arbitrated by a third reader or by discussion. A grid-enabled virtual meeting between readers could provide the medium where these issues can be discussed, but calls for a very high degree of acceptance of the technology on the part of the users. Problems may not be resolved as quickly, and the process may well become more formal. Even questions of clarification may have connotations of professional competence.

The visibility of readers' decisions serves to alert readers to occasions where colleagues are not aligned with a local understanding about what constitutes an appropriately recallable presentation. Physicians and other

healthcare workers "typically assess the adequacy of medical information on the basis of the perceived credibility of the source" [22]. In other words, healthcare workers develop a sense of the trustworthiness of the people (and machines) they work with and evaluate the meaning, status and quality of data in accordance. How can a reader who lacks knowledge of the (local) conditions of a mammogram's production read that mammogram confidently? (i.e. in accordance with their sense of professional competence) How can an 'unknown' reader be trusted to have read mammograms in an accountably acceptable manner? One project does not answer all these questions, of course, but MammoGrid was able to establish a context within which realistic workflow, in a simulated routine screening,

MD3. After the issues with the distributed storage and querying of medical data have been resolved, the next task will be to deploy services that can build relationships between data items, and will provide appropriate representation to medical researchers. Particularly given that there have been no successful deployments of knowledge grids for medical research to date, this will pose a significant challenge. The data concerned can be extremely varied in nature, structure, format and volume. Depending on the area of research, the synthesis of knowledge from data could require sophisticated data mining, integrated disease modelling and medical image processing applications, and may also involve the use of techniques from artificial intelligence to derive relationships between data from different sources and in different contexts.

The development of medical ontologies and the mapping between ontologies will be particularly important for the successful deployment of knowledge grids. The standardisation of interfaces can dramatically increase interoperability between biomedical resources, and by operating on standardised data formats they can more easily be integrated into complete bioinformatics experiments by eliminating the restructuring of data between each service. The construction of standardised data formats can be improved by defining a domain ontology that covers the concepts used within a given domain. These ontologies will allow relationships between concepts and nuances in meaning to be captured, greatly enhancing the opportunities for communication, knowledge sharing and reuse, and machine reasoning.

An ontology is the systematic description of a given phenomenon: it often includes a controlled vocabulary and relationships, captures nuances in meaning and enables knowledge sharing and reuse. From an agreed ontology it is possible to define a common data model that describes the format of the data used by all the services. Another benefit of ontologies is that they can also help to provide useful high-level functionalities based on machine reasoning. The field of machine reasoning would be further enhanced if the functionalities of services themselves were described in an ontology, and not only the data they operate on. Examples of such functionalities are automatic service discovery, invocation and composition. The Semantic Web Activity at the World Wide Web Consortium is dedicated to these topics. In particular the Health Care and Life Sciences Interest Group (HCLSIG) is relevant for healthgrids. An ontology is only as valuable as the general support it has, as its primary purpose is to facilitate a common understanding of terms. As an ontology is adopted more widely, this increases the possibilities for a resource that supports it. Within life science and healthcare sector, the highest degree of general support is arguably the Open Biomedical Ontologies (OBO).

Semantic Web technologies are just emerging in the field of medical research and healthcare. Open issues include how to integrate biomedical data using ontologies, how to combine different initiatives and how to employ advanced, semantic reasoning techniques for analysing medical data. The majority of the biomedical applications currently using ontologies mostly deal with decision support, namely assisting health professionals in disease diagnosis, staging or therapy planning via preliminary detection services. Breast cancer diagnosis and treatment is one of the most advanced domains with the development of a Breast Cancer Imaging Ontology. Anatomy is another field where ontological approaches have been explored. The interest is focused on the representation of anatomical terminology and classification of surgical procedures, extraction of heart anatomical features, etc. The GALEN model aims at developing advanced terminology systems for clinical information systems. The ontology for the GALEN

model is designed to be re-usable and application independent. It is intended to serve not only for the classification of surgical procedures but also for a wide variety of other applications—electronic healthcare records (EHCRs), clinical user interfaces, decision support systems, knowledge access systems, and natural language processing.

Ontology approaches are also under development in the cardiological domain. For instance, NOESIS aims at developing a platform for wide scale integration and visual representation of medical intelligence for research and cure of cardiac and cardiovascular diseases.

In most cases, biomedical ontologies function as terminology vocabularies, containing the domain knowledge required to build the classes, rules and relationships according to which the several concepts interact with each other. The Unified Medical Language System (UMLS) facilitates the development of computer systems that behave as if they "understand" the language of biomedicine and health. Developers use UMLS to build or enhance systems that create, process, retrieve, and integrate biomedical and health data and information. A very good example is the NCI Thesaurus which is a public domain description logic-based terminology produced by the American National Cancer Institute. This thesaurus implements rich semantic interrelationships between the nodes of its taxonomies. The semantic relationships in the thesaurus are intended to facilitate translational research and to support NCI bioinformatics infrastructure.

6. Conclusion

In total, SHARE predicts that the journey from a sustainable computing grid to a generalised knowledge grid should take from 7 to 15 years. However, the transition to data grid may not be as simple as the success of projects such as BIRN and BRIDGES would suggest, and the transition to knowledge grid and its later generalisation will be breaking new ground. It is therefore possible that this timescale will be multiplied by several times; it has been suggested that a more realistic timeframe might be 15 to 25 to 50 years (Fig. 3).

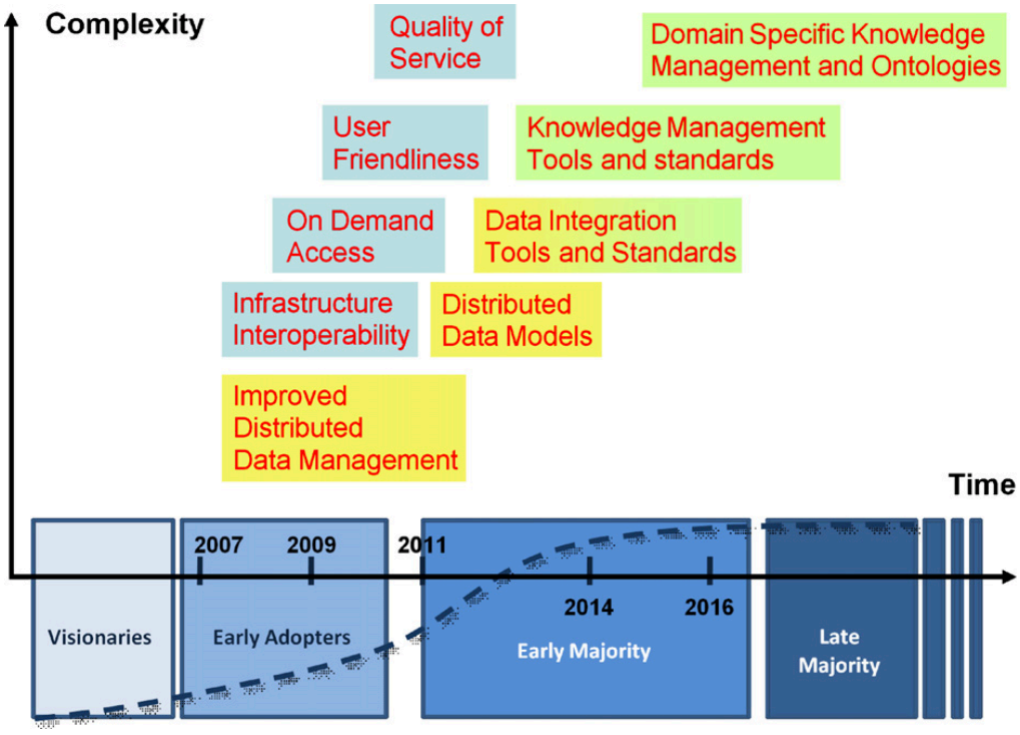


Fig. 3 – Complexity vs time: adoption curve for HealthGrid technologies.

The distinction made in innovation studies between 'visionaries', 'early adopters', 'early majority' and 'late majority' is reflected here in Fig. 3. Even for early adopters, infrastructure interoperability and distributed data management are necessary; on demand access, 'user friendliness' and quality of service are at the first point of inflection, before rapid expansion, followed by sophisticated AI tools in the later stages, where a second inflection occurs and the technologies become routinely accepted.

Certain specific features of the community, such as issues of patient ownership of her/his data and the tension between hospitals' IT policies and the requirements of grids, will continue to prove troublesome unless addressed with political will. Another non-functional obstacle is the drag on technology transfer between EC projects. E.g. there is a need for HealthGrid projects to begin thinking about data curation and digital libraries, but researchers and providers have not come together to explore this need. We hope to publish a further article to cover these softer issues in relation to the technological advances we have canvassed here.

Acknowledgements

The authors acknowledge the European Commission for funding the project SHARE: EU-FP6-2005-IST 027694. The work reported here has been carried out in collaboration with many colleagues in the HealthGrid community (www.healthgrid.org). Thanks are also due to numerous other colleagues in Europe, the US and Asia for many helpful discussions.

References

[1] V. Breton, A. E. Solomonides, R. H. McClatchey, A perspective on the HealthGrid initiative, in: Second International Workshop on Biomedical Computations on the Grid held at the Fourth IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2004), Chicago, USA, April, 2004.

[2] V. Breton, I. Blanquer, V. Hernandez, Y. Legré, A. E. Solomonides, Challenges and Opportunities in Healthgrids, in: Proceedings of HealthGrid 2006, vol. 120, IOS Press Studies in Health Technology and Informatics, 2006.

[3] C. delFrate, *et al*, Final Results and Exploitation Plans for MammoGrid, in: Challenges and Opportunities in Healthgrids, Proceedings of HealthGrid 2006, vol. 120, IOS Press Studies in Health Technology and Informatics, 2006.

[4] J. Freund, *et al*, Health-e-Child: an integrated biomedical platform for grid-based paediatric applications, in: Challenges and Opportunities of Healthgrids, Proceedings of HealthGrid 2006, vol. 120, IOS Press Studies in Health Technology and Informatics, 2006.

[5] STEP Consortium, Seeding the Europhysiome: A Roadmap for the Virtual Physiological Human. http://www. europhysiome.org and http://www.biomedtown.org/ biomed town/STEP/Reception/step presentations/ RoadMap/vph roadmap v2b.pdf.

[6] F. Martin-Sanchez, V. Maojo, G. Lopez-Campos, Integrating genomics into health information systems, in: Methods of Information in Medicine, vol. 41, 2002, pp. 25–30.

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

[8] V. Breton, K. Dean, T. Solomonides, From grid to HealthGrid, in: Proceedings of HealthGrid 2005, vol. 112, IOS Press Studies in Health Technology and Informatics, 2005.

[9] GEMSS Project, Grid Enabled Medical Simulation Services. www.gemss.de.

[10] R. D. Stephens, A. J. Robinson, C. A. Goble, myGrid: personalised bioinformatics on the information grid, Bioinformatics 19 (Suppl. 1) (2003).

[11] P. Groen, D. Goldstein, Grid computing, health grids and EHR systems Virtual Medical Worlds, vol. 26, 2006, December http://www.hoise.com/vmw/07/articles/vmw/LV-VM-01-07-36.html.

[12] OpenGridForum Life Sciences Grid—Research Group at http://forge.gridforum.org/projects/lsg-rg.

[13] European Commission e-Health—making healthcare for European citizens: an action plan for a European e-Health Area COM, 2004, p. 356, Final this 'Communication from the Commission' is at: http://europa.eu.int/information society/doc/qualif/health/COM 2004 0356 F EN ACTE.pdf.

[14] The Globus Alliance Globus Toolkit 4. http://www.globus.org/.

[15] GRIA Grid Resources for Industrial Applications. http://www.gridstart.org/factsheets/GRIA factsheet.pdf.

[16] DEISA Distributed European Infrastructure for Supercomputing Applications. http://www.deisa.org/.

[17] EGEE Enabling Grids for E-sciencE. http://www.eu-egee.org/.

[18] SHARE Technology Baseline Report. http://eu-share.org/about-share/deliverables-and-documents.html.

[19] D. Budgen, M. Turner, I. Kotsiopoulos, F. Zhu, M. Russell, M. Rigby, K. Bennett, P. Brereton, J. Keane, P. Layzell, Managing Healthcare Information: The Role of the Broker, in: From Grid to HealthGrid, Proceedings of HealthGrid 2005, vol. 112, IOS Press Studies in Health Technology and Informatics, 2005.

[20] R. Warren, A. E. Solomonides, C. delFrate, *et al*, MammoGrid a prototype distributed mammographic database for Europe, Clinical Radiology 62 (November (11)) (2007) 1044–1051.

[21] R. Warren, D. Thompson, C. delFrate, *et al*, A comparison of some anthropometric parameters between an Italian and a UK population: "proof of principle" of a European project using MammoGrid, Clinical Radiology 62 (November (11)) (2007) 1052–1060.

[22] A. Cicourel, The Integration of Distributed Knowledge in Collaborative Medical Diagnosis, in: J. Galegher, R. Kraut, C. Egido (Eds.), Intellectual Teamwork: Social and Technological Foundations of Cooperative Work, Lawrence Erlbaum Associates, 1990.

Added in proof prior to publication

Summary points

What was already known on this topic

- Several successful project have been completed in the field of grid computing for health ("healthgrids"). These associated themselves with an effort to synthesise the potential and the prospects of healthgrids through the HealthGrid association.
- HealthGrid published its White Paper in 2005 setting out a vision of what can be accomplished through appropriate adaptations of grid computing in biomedical research and healthcare applications.
- The SHARE project was specifically funded to develop a "road map" for healthgrids which would see the realisation of the vision in the following decade or so and so also address the aspirations of the European Union's "Action Plan for a European e-Health Area".

What this study has added

- A state of the art report led to an outline roadmap with only the most basic of milestones identified. Development then followed separate paths roughly along the lines of functional and non-functional requirements, thought of as technical and standards on one hand and as ethical, legal, social and economic ('ELSE') issues on the other.
- These ideas were concurrently tested in two detailed use cases, innovative medicine and epidemiology, and against three actual projects. These three distinct streams fed into the development of the second road map, a significantly more sophisticated and differentiated document.
- An agreed roadmap was finally generated after several presentations to expert groups and workshops.
- In developing the roadmap it became necessary to differentiate between data, computing and collaboration grids. These vary in their technical requirements and to some extent in the ELSE issues they give rise to.
- While it is clear that healthgrids have immense potential in biomedical research, it is equally clear that applications in healthcare will present significant challenges to regulatory regimes. In the latter context, creative compromises will be necessary but do appear possible; moreover, the technology to automate certain aspects of regulatory compliance have the potential to be incorporated in healthgrids.

SHARE, from Vision to Road Map: Technical Steps

Mark OLIVE [a], Hanene RAHMOUNI [a], Tony SOLOMONIDES [a,1], Vincent BRETON [b], Yannick LEGRÉ [b], Ignacio BLANQUER [c] and Vicente HERNANDEZ [c]
[a] *CCCS / CEMS Faculty / UWE, Bristol / Coldharbour Lane / Bristol BS16 1QY / UK*
[b] *LPC / CNRS-IN2P3 / Campus de Cézeaux / 63177 Aubière Cedex / France*
[c] *Universidad Politécnica de Valencia / Camino de Vera s/n / E-46022 Valencia / Spain*

Abstract. We present the 'HealthGrid' initiative and review work carried out in various European healthgrid projects and report on joint work with numerous European collaborators. Since the European Commission's Information Society Technologies programme funded the first grid-based health and medical projects, the HealthGrid movement has flourished in Europe. Many projects have now been completed and 'HealthGrid' consulted a number of experts to compile and publish a 'White Paper' which establishes the foundations, potential scope and prospects of an approach to health informatics based on a grid infrastructure. The White Paper demonstrates the ways in which the healthgrid approach supports modern trends in medicine and healthcare, such as evidence-based practice and information integration. With a second generation of projects now funded, the EC has commissioned a study to define a research roadmap for a 'healthgrid for Europe' as the preferred infrastructure for medical and health care projects in the European Research Area.

Keywords. healthgrid, e-health, grid applications

1. The Healthgrid Initiative

'Grid' has been identified as one of the key technologies to support the European Research Area. The impact of this concept is expected to reach far beyond eScience, to eBusiness, eGovernment, and eHealth, but a major challenge is to take the technology out of the laboratory to the citizen. A *healthgrid* is an environment in which medical data can be stored and made available to all actors in the healthcare system, doctors, allied professions, healthcare centres, administrators and, of course, patients and citizens in general. Such an environment has to offer all appropriate guarantees in terms of data protection, respect for ethics and observance of regulations; it has to support the notion of 'duty of care' and may have to deal with 'freedom of information' issues. Working across member states, it may have to support negotiation and policy bridging.

Pioneering projects in the application of grid technologies to the health area have been completed, and the technology to address high level requirements in a grid environment has been under development and making good progress. Because these projects had a finite lifetime and the vision required a sustained effort over an extended

[1] Corresponding Author: Tony.Solomonides@uwe.ac.uk. The work reported here has been carried out in collaboration with many colleagues in the HealthGrid community. Thanks are due to these and numerous other colleagues in Europe, the US and Asia for helpful discussions.

period, and besides because there was an obvious need for these projects to cross-fertilise, the 'HealthGrid initiative', represented by the HealthGrid association (http://www.healthgrid.org), was launched to bring the necessary long-term continuity. Its goal is to encourage and support collaboration between autonomous projects in such a way as to ensure that requirements really are met and that the wheel, so to speak, is not re-invented repeatedly at the expense of other necessary work.
Writing about the healthgrid initiative very soon after its inception, this community identified a number of objectives [1]: identification of potential business models for medical grid applications; feedback to the grid development community on the requirements of the pilot applications deployed by the European projects; development of a systematic picture of the broad and specific requirements of physicians and other health workers when interacting with grid applications; dialogue with clinicians and those involved in medical research and grid development to determine potential pilots; interaction with clinicians and researchers to gain feedback from the pilots; interaction with all relevant parties concerning legal and ethical issues identified by the pilots; dissemination to the wider biomedical community on the outcome of the pilots; interaction and exchange of results with similar groups worldwide; and the formulation and specification of potential new applications with the help of the end user communities.

The grid concept is rooted in the physical sciences and these considerations were not a central concern to general grid developers. Even today these requirements are not a priority for developers, even though they have been fed through to the middleware services community. Thus HealthGrid identified the need for a specialist middleware layer, between the generic grid infrastructure and the medical or health applications.

Among data related requirements, the need for suitable access to biological and medical image data arose in several early projects, but for the most part these are present in other fields of application also. Looking to security requirements, most of these are special to the medical field: anonymous or private login to public and private databases; guaranteed privacy, including anonymization, pseudonymization and encryption as necessary; legal requirements, especially in relation to data protection, and dynamic negotiation of security and trust policies while applications remain live. Medical applications also require access to small data subsets, like image slices and model geometry. At the (batch) job level, medical applications need an understanding of job failure and means to retrieve the situation.

2. The White Paper: From Grid to Healthgrid

The next step for the HealthGrid community was to try to systematize the concepts, requirements, scope and possibilities of grid technology in the life sciences. The White Paper [2] defines the concept of a healthgrid more precisely than before: ... grid infrastructures comprising applications, services or middleware components that deal with the specific problems arising in the processing of biomedical data.

The ultimate goal for eHealth in Europe may be the creation of a single healthgrid incorporating a 'principle of subsidiarity' for independent nodes of the healthgrid as a means of implementing all the legal, ethical, regulatory and negotiation requirements. We may anticipate, however, the development path to proceed through specific healthgrids with perhaps rudimentary inter-grid interaction/interoperational capabilities. We may therefore identify a need to map future research and advice on research policy, so as to bring diverse initiatives to the point of convergence.

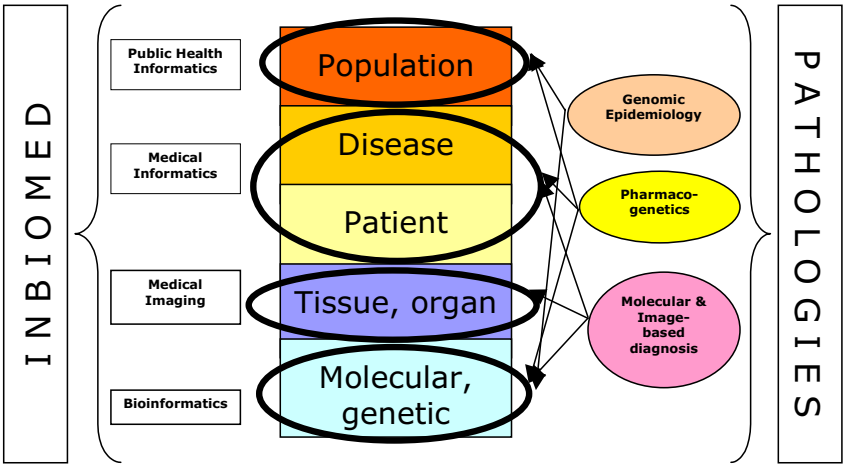
Healthgrid applications address both individualised healthcare – diagnosis and treatment - and epidemiology with a view to public health. Individualized healthcare is improved by the efficient and secure combination of immediate availability of personal clinical information and widespread availability of advanced services for diagnosis and therapy. Epidemiology healthgrids combine the information from a wide population to extract knowledge that can lead to the discovery of new correlations between symptoms, diseases, genetic features and other clinical data. With this broad range of application in mind, the issues below are identified as key features of our analysis.

- Business case, trust and continuity issues: healthgrids are data- and collaboration grids, but healthcare organizations are required by law to maintain control of their patients' records. Deployment on a scale to make an attractive business opportunity requires a high level security and compliance.
- Biomedical issues: Distributed databases and data mining are important tools for many biomedical applications in fields such epidemiology, drug design and even diagnosis. Expert system services running on the grid must be able to interrogate large distributed databases to explore sources of diseases, risk populations, evolution of diseases or suitable proteins to fight against specific diseases.
- Security issues: These flow naturally from the nature of medical data and from business requirements. Security in grid infrastructures is currently adequate only for research platforms.
- Management issues: The central concept of a 'virtual organisation' (VO) at the heart of eScience, which gave rise to grids, is very apt for healthgrid, but additional flexibility is needed to structure and to control VOs on a broader scale, including, for example, the meta-level of a VO of VOs.

We illustrate the concept of healthgrid with some prototypical examples: GEMSS [3] used a 'high-throughput' numerical simulation of organs obtained from a patients' data and used these to aid understanding or to improve the design of medical devices, with patient-customized approaches at research-level in areas such as radiotherapy, cranio-facial surgery and neurosurgery. MammoGrid [4] created a database of standardized mammogram files and associated patient data to enable radiologists in the UK and in Italy to request second opinion and computer-aided detection services. The project also enabled a further study of an epidemiological nature on breast density as a risk factor. In Health-e-Child [5] radiologists are working with oncologists, cardiologists and rheumatologists to identify early imaging signs of conditions that may have a strong genetic component, possibly reducing the need for genetic maps to be obtained on an indiscriminate basis. In Wide In Silico Docking On Malaria (WISDOM) [6] tens of millions of molecular docking experiments have been used to help identify potential antigens for the malaria parasite. The experiment uses large scale virtual screening techniques to select molecular fragments for further investigation in the development of pharmaceuticals for neglected diseases. The economic dynamics in this area are telling: only about 1% of drugs developed in the last quarter century have been aimed at tropical diseases, and yet these are major killers in the third world, with mortality in excess of 14 million per annum. Meanwhile, the results of several major studies of the interface between bioinformatics and medical informatics had been published with a remarkable promise of synergy between the two disciplines, leading to what had already begun to be referred to as 'personalised medicine'. [7,8]

3. The SHARE Project: From White Paper to Road Map

The vision of health that informs the thinking of the White Paper is reflected in European thinking [9] and is depicted in a map of the relationships between the different ontological and epistemological levels and the various modalities of data have been captured by Fernando Martin-Sánchez (cf [1]) in the schematic diagram below.



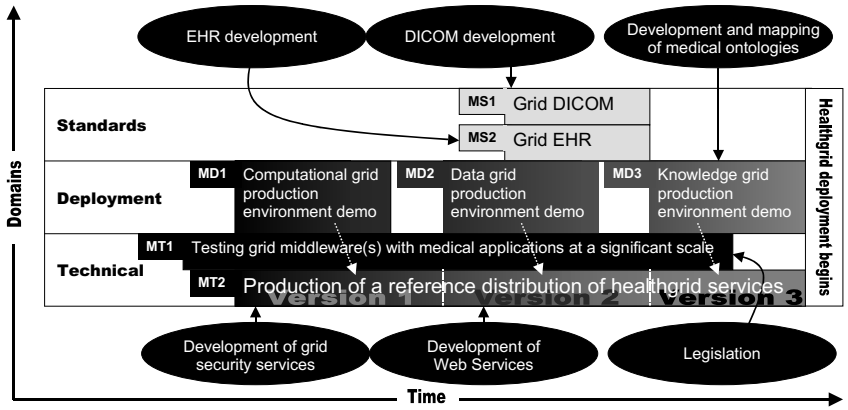
Disciplines, levels of being and pathology diagnostics (F. Martin-Sánchez)

In the White Paper, the HealthGrid community expressed its commitment to engage with and support modern trends in medical practice, especially 'evidence-based medicine' as an integrative principle, to be applied across the dimensions of individual through to public health, diagnosis through treatment to prevention, from molecules through cells, tissues and organs to individuals and populations.

In view of the impact of the White Paper, the EC has funded the project SHARE [10] to explore exactly what it would mean to realise the vision of the White Paper, investigate the issues that arise and define a roadmap for research and technology which would lead to wide deployment and adoption of healthgrids in the next ten years. Thus the project must address the questions, *What research and development needs to be done now?* and *What are the right initiatives in eHealth RTD policy relating to grid deployment?*, with all that implies in terms of coordination of strategy, programme funding and support for innovation. Thus the project will define a comprehensive European research and development roadmap, covering both policy and technology, to guide and promote beneficial EU-wide uptake of healthgrid technologies.

4. Technical Road Map: Step One

SHARE has defined a preliminary technical road map (see figure below) with two technical milestones for appropriate development of healthgrid services (**MT1**,**2**), two milestones as examples of grid standards for the medical domain (**MS1**,**2**), and three deployment milestones of increasing complexity and scope (**MD1**,**2**,**3**).



SHARE technical roadmap diagram, showing milestones and external influences

MT1 Before deployment can begin, grid middleware must be tested with medical applications for scalability and robustness. This must begin at an early stage. It is anticipated that this will be an ongoing activity, with different generations of grid operating systems offering newer, faster and more stable capabilities. A key issue is the robustness of web-services based grid solutions. Scalability, especially of medical applications, is still a concern for grid middleware based on the Open Grid Services Architecture (OGSA), while other exemplars, such as gLite and Unicore, have been deployed on large scale infrastructures in Europe with good scalability and robustness, but are still awaiting migration to web services.

MD1 The first deployment step will be the rollout of a computational grid production environment demonstrator for medical research. This should be a near-term achievable goal given that there have already been successful deployments of computational grid applications (cf. WISDOM) on general purpose grid infrastructures such as DEISA and EGEE. However, it is a wholly different matter to create a convincing business case for healthcare management.

MT2 starts with **MD1** and ends when production deployment begins. This milestone is the development of a reference distribution of grid services, using standard web service technology and allowing secure manipulation of distributed data. Standards in emerging Web services technologies (WSx and the WSRF specification) will facilitate interoperability between healthgrids built using different underlying tools. A precise, well documented set of requirements is needed to describe the security features and obligation policies at different levels of abstraction in the middleware.

MD2 Developing and maintaining a production-quality data grid will require resolution of several issues relating to the distributed storage of medical data. In European grid infrastructures, storage of medical images has been hampered by limited data management services. High speed links will be an obvious prerequisite.

MS1 & **MS2** The use of computer-based tools for clinical research has led to the definition of standards for the exchange of data in many areas but their adoption has not been universal. The exchange of data between bioinformatics and medical informatics is an area where standards are particularly limited. By contrast, in medical imaging the adoption of DICOM for the storage and transmission of medical images

has been accepted worldwide. In medical records HL7 is the emerging standard. For both of these standards, there is a question of compatibility with grid technologies.

MD3 The final milestone is the deployment of services whose purpose is to sustain a knowledge grid for medical research. Beginning with a single domain of application, the development of medical ontologies will allow relationships between concepts and nuances in meaning to be captured, thus enabling knowledge sharing and management.

5. Conclusions and future work

Certain specific features of the community, such as issues of patient ownership of her/his data and the tension between hospitals' IT policies and the requirements of grids, will continue to prove troublesome unless addressed with political will. Another non-functional obstacle is the drag on technology transfer between EC projects. E.g. there is a need for healthgrid projects to begin thinking about data curation and digital libraries, but researchers and providers have not come together to explore this need.

SHARE predicts that it may take ten to fifteen years from a sustainable computing grid to a generalised knowledge grid. However, the transition to data grid may not be as simple as the success of special projects suggests and the transition to knowledge grids will be breaking new ground. It has been suggested that a more realistic timeframe might be twenty to forty years. As a next step, SHARE will focus on the large number of Ethical, Legal and Socio-Economic issues related to healthgrids. These will be integrated with the technical roadmap to recommend both technical and policy actions.

References

[1] V. Breton, A. E. Solomonides & R.H. McClatchey. *A Perspective on the Healthgrid Initiative*. Second International Workshop on Biomedical Computations on the Grid, the Fourth IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2004), Chicago, USA, April 2004

[2] V. Breton, K. Dean & T. Solomonides (Eds.) *The Healthgrid White Paper*. From Grid to Healthgrid: Proceedings of Healthgrid 2005: pp. 249-321, IOS Press, Studies in Health Technology and Informatics, Vol 112, 2005.

[3] GEMSS Project. *Grid Enabled Medical Simulation Services*. Available at: http://www.gemss.de

[4] C. del Frate et al. *Final Results and Exploitation Plans for MammoGrid*. Challenges and Opportunities of Healthgrids, Proceedings of HealthGrid 2006, IOS Press, Studies in Health Technology and Informatics Vol 120, 2006.

[5] J. Freund et al. *Health-e-Child: An Integrated Biomedical Platform for Grid-Based Paediatric Applications*. Challenges and Opportunities of Healthgrids, Proceedings of HealthGrid 2006, IOS Press, Studies in Health Technology and Informatics Vol 120, 2006.

[6] Jacq, N., Salzemann, J., Jacq, F., Legré, Y., Medernach, E., Montagnat, J., Maaß, A., Reichstadt, M., Schwichtenberg, H., Sridhar, M., Kasam, V., Zimmermann, M., Hofmann, M., Breton, V., *Grid-enabled Virtual Screening Against Malaria*, Journal of Grid Computing (to appear 2007).

[7] A. Sousa Pereira, V. Maojo, F. Martin-Sanchez, A. Babic & S. Goes. *The INFOGENMED Project*. ICBME 2002, Singapore, 2002.

[8] BIOINFOMED Project. *Synergy between Medical Informatics and Bioinformatics: Facilitating Genomic Medicine for Future Healthcare*, BIOINFOMED Study Report White Paper, 2003.

[9] European Commission. *e-Health – making healthcare for European citizens: An action plan for a European e-Health Area*. COM(2004) 356 final. A 'Communication from the Commission', available at: http://europa.eu.int/information_society/doc/qualif/health/COM_2004_0356_F_EN_ACTE.pdf

[10] V. Breton, I. Blanquer, V. Hernandez, Y. Legré & A. E. Solomonides. *Proposing a Road Map for Healthgrids*. Challenges and Opportunities of Healthgrids, Proceedings of HealthGrid 2006, IOS Press, Studies in Health Technology and Informatics Vol 120, 2006.

Paper H: Due to publisher restrictions the paper is not available in this document.

SHARE: A European Healthgrid Roadmap, which was originally published as a report by the European Commission (*SHARE the Journey: A European Healthgrid Roadmap*) and subsequently re-edited for inclusion as Chapter 1 of the **Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare** (IGI Global, 2009) has been withheld because of publisher's restrictions.

Compliance and creativity in grid computing

Anthony E. Solomonides
CEMS Faculty, University of the West of England, Bristol, BS16 1QY, UK
tony.solomonides@uwe.ac.uk

Abstract
Grid computing ("the grid") is a promising new technology to enhance the services already offered by the internet. This new paradigm offers rapid computation, large scale data storage and flexible collaboration by harnessing together the power of a large number of commodity computers or clusters of other basic machines. The grid was devised for use in scientific fields, such as particle physics and bioinformatics, in which large volumes of data, or very rapid processing, or both, are necessary. Unsurprisingly, the grid has also been used in a number of ambitious medical and healthcare applications. While these initial exemplars have been restricted to the research domain, there is a great deal of interest in real world applications. However, there is some tension between the spirit of the grid paradigm and the requirements of medical or healthcare applications. The grid maximises its flexibility and minimises its overheads by requesting computations to be carried out at the most appropriate node in the network; it stores data at the most convenient node according to performance criteria. On the other hand, a hospital or other healthcare institution is required to maintain control of its confidential patient data and to remain accountable for its use at all times. Despite this apparent conflict in requirements, we suggest that certain characteristics of the grid provide the means to resolve the problem: in the spirit of this paradigm in which "virtual organisations" arise ad hoc, "grid services" may negotiate ethical, legal and regulatory compliance according to agreed policy.

Introduction: the computing context

I will introduce some of the issues that concern us through examples from several recent projects in the field of 'healthgrid'. I will first motivate the concept of grid computing. 'Distributed computer systems' predate even the internet and the World Wide Web ('the web'). By means of a network of interconnections, computers are able to share a workload that would ordinarily be beyond the capacity of any one of them; they may also distribute data to different locations according to need or frequency of use. On the other hand, since the explosion of the web in every conceivable statistic – users, nodes, volume of information – we are familiar with its ability to serve information and misinformation in equal measure. The grid combines the technical features of distributed systems and the web, but efforts are also being made to ensure that it is not beset by the same problems of abuse, misuse and contamination as the web has been.

The ideal grid, envisaged as a servant of a new paradigm of scientific research called 'e-science', would provide transparent processing power, storage capacity and communication channels for scientists who may from time to time join the grid, do some work and then leave, so that the alliances they form in their scientific endeavours might be described as 'virtual organizations' or VOs for short. Different sciences have different needs, and the grid concept has become differentiated: particle physics generates enormous amounts of data which must be kept, but not necessarily instantly processed; on the other hand, data in bioinformatics is not large by comparison – it is, of course, in plain terms, large – but requires intensive processing. In extending the application of grid computing to e-health, another feature becomes pre-eminently necessary: that of collaboration.

An important consequence of the fluidity of collaboration in grid computing has been in the choice of 'architecture' for grid systems. 'Architecture' is used loosely in computer systems to describe the manner in which hardware and software have been assembled together to achieve a desired goal. Favoured also in the commercial application of the web, the so-called 'Service-Oriented Architecture' has been widely adopted in grid applications. In effect, it means that needed services – software applications – once constructed, are provided with a description in an agreed language and made available to be 'discovered' by other services that need them. A 'service economy' is thus created in which both *ad hoc* and systematic collaborations can take place.

Compared with data from physics or astronomy, medical data is less voluminous, but requires much more careful handling. Among the services it therefore calls for are 'fine grained' access control – e.g. through authorization and authentication of users – and privacy protection through anonymization or pseudonymization of individual data or 'outlier' detection and disguise in statistical data. There are, of course, many more specialist medical services, as our examples below reveal. It is a current requirement in the United States, for example, that if head images are communicated outside the team immediately caring for a patient, all facial features which might identify the patient must be removed.

Breast Cancer and MammoGrid

Breast cancer is arguably the most pressing threat to women's health. For example, in the UK, more than one in four female cancers occur in the breast and these account for 18% of deaths from cancer in women. Coupled with the statistic that about one in four deaths in general are due to cancer, this suggests that nearly 5% of female deaths are due to breast cancer. While risk of breast cancer to age 50 is 1 in 50, risk to age 70 increases to 1 in 15 and lifetime risk has been calculated as 1 in 9. The problem of breast cancer is best illustrated through comparison with lung cancer which also accounted for 18% of female cancer deaths in 1999. In recent years, almost three times as many women have been diagnosed with breast cancer as with lung cancer. However, the five year survival rate from breast cancer stands at 73%, while the lung cancer figure is 5%. This is testament to the effectiveness of modern treatments, provided breast cancer is diagnosed sufficiently early. These statistics are echoed in other countries. The lifetime risk of breast cancer in the USA has been estimated as 1 in 8. Here also incidence has increased but mortality decreased in the past twenty years. Twenty years ago breast cancer was almost unknown in Japan but its incidence now approaches Western levels. (For a world-wide picture, see [1].)

The statistics of breast cancer diagnosis and survival appear to be a powerful argument in favour of a universal screening programme. However, a number of issues of efficacy and cost effectiveness limit the scope of most screening programmes. The method of choice in breast cancer screening is mammography (breast X-ray); for precise location of lesions and 'staging' (establishing how advanced the disease is) ultrasound and MRI may be used. A significant difficulty lies in the typical composition of the female breast, which changes dramatically over the lifetime of a woman, with the most drastic change taking place around the menopause. In younger women, the breast consists of around 80% glandular tissue which is dense and largely X-ray opaque. The remaining 20% is mainly fat. In the years leading up to the menopause, this ratio is typically reversed. Thus in women under 50, signs of malignancy are far more difficult to discern in mammograms than they are in post-menopausal women. Consequently, most screening programmes, including the UK's, only apply to women over 50.

The increasing use of electronic formats for radiological images, including mammography, together with the fast, secure transmission of images and patient data, potentially enables

many hospitals and imaging centres throughout Europe to be linked together to form a single grid-based "virtual organization". It is not yet precisely understood what advantages might accrue to radiologists working in such virtual organizations, as the technological possibilities are co-evolving with an appreciation of potential uses; but one that is generally agreed is the creation of huge "federated" databases of mammograms, which appear to the user to be a single database but are in fact retained and curated in the centres that generated them. Each image in such a database would have linked to it a large set of relevant information, known as metadata, about the woman whose mammogram it is. Levels of access to the images and metadata in the database would vary among authorized users according to their "certificated rights": healthcare professionals might have access to essentially all of it, whereas, e.g., administrators, epidemiologists and researchers would have limited access, protecting patient privacy and in accordance with European legislation.

The Fifth Framework EU-funded MammoGrid project (2002-05) [3] aimed to apply the grid concept to mammography, including services for the standardization of mammograms, computer-aided detection (CADe) of salient features, especially masses and 'microcalcifications', quality control of imaging, and epidemiological research including broader aspects of patient data. In doing so, it attempted to create a paradigm for practical, grid-based healthcare-oriented projects, particularly those which rely on imaging, where there are large volumes of data with complex structures. Clinicians rarely analyse single images in isolation but rather in a series or in the context of metadata. Metadata that may be required are clinically relevant factors such as patient age, exogenous hormone exposure, family and clinical history; for the population, natural anatomical and physiological variations; and for the technology, image acquisition parameters, including breast compression and exposure data.

As a research project, MammoGrid encompassed three selected clinical problems:

i Quality control: the effect on clinical mammography of image variability due to differences in acquisition parameters and processing algorithms;

ii Epidemiological studies: the effects of population variability, regional differences such as diet or body habitus and the relationship to mammographic density (a potential biomarker of breast cancer) which may be affected by such factors;

iii Support for radiologists, in the form of tele-collaboration, second opinion, training and quality control of images.

The MammoGrid proof-of-concept prototype enables clinicians to store digitized mammograms along with appropriately anonymized patient metadata; the prototype provides controlled access to mammograms both locally and remotely stored. A typical database comprising several thousand mammograms has been created for user tests of clinicians' queries. The prototype comprises (a) a high-quality clinician visualization workstation (used for data acquisition and inspection); (b) an interface to a set of medical services (annotation, security, image analysis, data storage and queries) accessed through a so-called *GridBox*; and (c) secure access to a network of other *GridBoxes* connected through grid middleware. The *GridBoxes* may therefore be seen as gateways to the grid.

The prototype provides a medical information infrastructure delivered in a service-based grid framework. It encompasses geographical regions with different clinical protocols and diagnostic procedures, as well as lifestyles and dietary patterns. The system allows, among other things, mammogram data mining for knowledge discovery, diverse and complex epidemiological studies, statistical analyses and CADe; it also permits the deployment of

different versions of the image standardization software and other services, for quality control and comparative study.

It was always the intention of MammoGrid to get rapid feedback from a real clinical community about the use of such a simple grid platform to inform the next generation of grid projects in healthcare. In fact, a Spanish company has already entered into negotiations to commercialize the project and to deliver a real, MammoGrid-based radiology service in the region of Extremadura. Thus, many ideas which came up as questions, issues or obstacles in research, must be solved in a real-life system within the next two or three years.

We may now imaginatively consider what may happen in the course of a consultation and diagnosis using the MammoGrid system. A patient is seen and mammograms are taken. The radiologist is sufficiently concerned about the appearance of one of these that she wishes to investigate further. In the absence of any other method, she may refer the patient for a biopsy, an invasive procedure; however, she also knows that in the majority of cases, the initial diagnosis turns out to have been a false positive, so the patient has been put through a lot of anxiety and physical trauma unnecessarily. Given the degree of uncertainty, a cautious radiologist may seek a second opinion: how can the MammoGrid system support her? She may invoke a CADe service; the best among these can identify features which are not visible to the naked eye. Another possibility is to seek out similar images from the grid database of mammograms and examine the history to see what has happened in those other cases. However, since each mammogram is taken under different conditions, according to the judgement of a radiographer ('radiologic technician') it is not possible to compare them as they are. Fortunately, a service exists which standardizes and summarizes the images, provided certain parameters are available – the type of X-ray machine and its settings when the mammograms were taken. Perhaps at this particular moment the radiologist's workstation is already working at full capacity because of other imaging tasks, so it is necessary for the image to be transmitted to a different node for processing. Since our grid is distributed across Europe, it now matters whether the node which will perform the standardization is in the same country or not. Let us suppose that it is a different country. A conservative outcome is to ensure that, provided the regulatory conditions in the country of origin and in the country where the processing will take place are mutually compatible (i.e. logically consistent, capable of simultaneous satisfaction) that they are both complied with. If one set requires encryption, say, but the other does not, the data must be encrypted. If both sets of regulations allow the image to be transmitted unencrypted but one country requires all associated data transmitted with the image to be pseudonymized, this must be done. These are human decisions, but it is clear that they can be automated. Where will responsibility lie if something goes wrong in this process? In any case, the story has further ramifications: the whole idea of MammoGrid is to build up a rich enough database of images and case histories to provide a sound basis both for diagnostic comparison and for epidemiology. Once standardized and returned, is the image now to be stored and made available to others for comparative use, or is it to remain outside the system. This is now a question of informed consent. Will a service, in the sense we have already used the term, be trusted to determine whether such informed consent as the patient has given covers this question?

We now consider the comparison the radiologist wanted to make – the reason for standardizing the image to begin with. The intention is to find images which are sufficiently similar and whose associated history gives an indication of the associated risk. For example, if from among the ten most similar instances, seven turn out to be malignant, there would be good reason to proceed to the more invasive stage of investigation. But how is the database to be queried so as to suggest valid comparisons? Clearly, this goes beyond image similarity. The risks for a childless woman of 65 are very different from a 50-year old mother of three.

Image similarity would not be sufficient to warrant a comparison. Thus we must transmit, as part of the database query, data that potentially identify the patient; and the result of the query may provide data which potentially identify patients. On a need-to-know basis, the radiologist has to know details of the cases, but not necessarily the names of the patients, although it would not be difficult to imagine a case where the name reveals something about ethnic background and this turns out to be significant. In a fully deployed system, there may be relevant cases and images from several countries; the system must be capable of 'policy bridging', as described above, to ensure that all regulatory conditions are met. Indeed, if the impact of including a case from one particular country would be to render the comparison less useful overall, perhaps the system should be able to reject that particular case – in other words, to apply a criterion which maximizes the information obtained subject to satisfaction of applicable laws and regulations – where the 'applicable set' is itself a variable.

Evidence-Based and Individualized Medicine

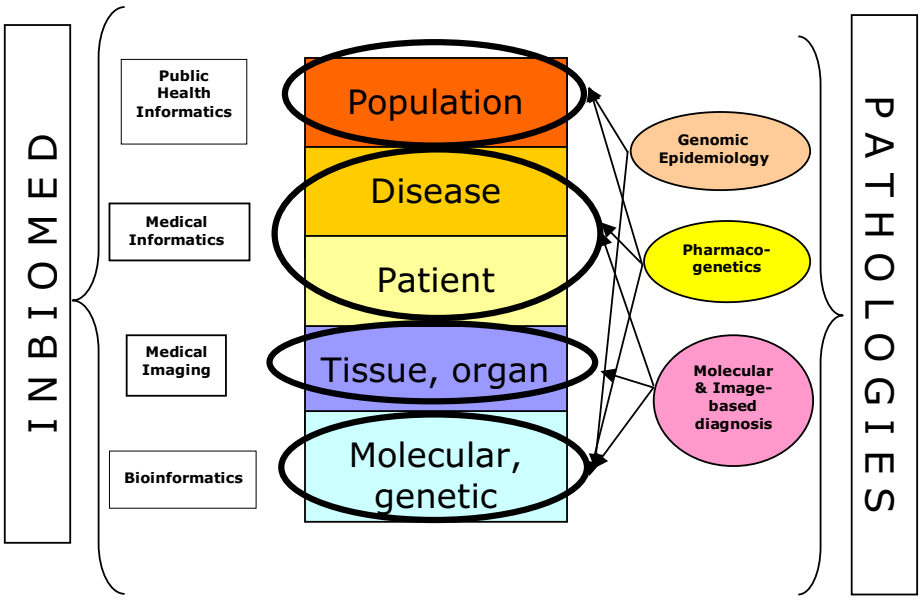
Hitherto, I have given a 'naïve' account of one system and its approach to diagnosis. How is such a system to fit into the modern conception of evidence-based medicine, i.e. medicine that is based on scientific results, rather than on the doctor's intuition, personal knowledge and craft skill? Evidence-based practice rests on three pillars: medical knowledge, as much as possible based on 'gold standard' (double-blind, controlled) clinical trials whose results have been peer reviewed and then published; knowledge of the patient, as complete as the record allows; and knowledge of the resources, procedures and protocols available in the setting where the encounter with the patient is taking place.

There is a very extensive literature on knowledge management and the difficulties and opportunities it presents. Some work currently undertaken in the healthgrid context, such as on ontologies and on knowledge representation, is relevant here. A development which is bringing economics into conflict with the traditional approach to the establishment and dissemination of knowledge is online publication of research results. While in medicine at present this is restricted to electronic publication of papers that have already been peer reviewed and are in the pipeline for printing in a journal, in other fields of science, notably physics, immediate online publication of un-peer reviewed results so that they can be viewed and critically assessed is now common. In another field, the journal *Nature* recently conducted a comparative study of errors in *Wikipedia* and in the *Encyclopaedia Britannica*; the results were equivocal, leading some to argue that an online, user-managed encyclopaedia is less error prone, although there have been many hacking attacks on *Wikipedia*. In the case of medicine, not only malicious postings, but poor research may have serious results. The American Medical Informatics Association is currently promoting the concept of a world bank of clinical trials. Here it may be said that the traditional approach to knowledge has failed; negative results are often not published and, as certain legal cases have brought to light, even results suggestive of risks are kept under wraps. Another practice that would benefit from being documented is the effective prescription of certain drugs beyond their designed purpose or licence, where nevertheless anecdotal clinical evidence has led practitioners to believe they are effective.

However, the MammoGrid application we have described above (and other similar projects) takes us a step further in the direction of 'dynamic' construction of knowledge. If images and histories are to be used as part of the diagnostic knowledge in new cases, it is imperative that they are collected with as much care and rigour as the cases in a controlled trial. Therefore, it is essential to know the 'provenance' of the data with precise details of how it has been handled (e.g. if standardized and subjected to CADe, which algorithms were used, set to what parameters, by whom, and if capture and interpretation were subject to

appropriate practice standards). I have labelled this set of issues "the question of practice-based evidence for evidence-based practice". If this were to be accepted as an appropriate source of diagnostic information, the underlying grid services which maintain it would have to make quality judgements without human intervention.

A major breakthrough in healthcare is anticipated from the association of genetic data with medical knowledge. In the healthgrid research community we have a map that has become almost an article of faith:



Disciplines, levels of being and pathology diagnostics (acknowledgement: F. Martin-Sánchez)

This view of the 'life' is in fact shared by many different disciplines, system biology being the most obvious among them. Drug development is increasingly driven by a molecular view of the world, using a variety of models to understand both how drugs act and how their action may be enhanced, inhibited or frustrated. This usually means understanding what proteins are present and, therefore, which genes code for those proteins. In the foreseeable future, we may anticipate certain drugs to be available in subtypes to account for the specific genetic endowment of the patient.

This would suggest that genetic information would have to be accessed routinely in the course of healthcare. Viewing this as part of the information held on a patient raises a number of difficult problems. Among these are the predictive value and the shared nature of genetic information. Knowing a person's genome could mean knowing what diseases they may or may not be susceptible to. Knowing one person's genetic map also reveals that of his or her siblings' in large measure. This introduces a range of questions, from confidentiality to 'duty of care' issues. If physicians will be held liable both for what they do and what they do not do, is it necessary for the underlying knowledge technology to 'be aware' and to inform them of the possibilities?

The grid could provide the infrastructure for a complete 'electronic health record' with opportunities to link both traditional patient data and genetic information to bring us closer to

the ideal of genomic medicine. Among many questions being investigated in current projects is a set concerning development and illness in childhood, especially conditions in which genetic predisposition is at least suspected and in the diagnosis of which imaging is also essential. Physicians want to know how certain genes impact the development of diseases and radiologists want to know what the earliest imaging signs are that are indicative of a disease. For example, the Health-e-Child project [5] is investigating paediatric rheumatology, cardiac dysmorphology and childhood brain tumours using this approach. Consider its aims:

i To gain a comprehensive view of a child's health by vertically integrating biomedical data, information, and knowledge, that spans the entire spectrum from genetic to clinical to epidemiological;

ii To develop a biomedical information platform, supported by sophisticated and robust search, optimization, and matching techniques for heterogeneous information, empowered by the Grid;

iii To build enabling tools and services on top of the Health-e-Child platform, that will lead to innovative and better healthcare solutions in Europe:
   - Integrated disease models exploiting all available information levels;
   - Database-guided biomedical decision support systems provisioning novel clinical practices and personalized healthcare for children;
   - Large-scale, cross-modality, and longitudinal information fusion and data mining for biomedical knowledge discovery.

With major companies looking to translate research results into products, successful outcomes from this and other projects would bring the scenario described above closer to reality.

Next Steps

The SHARE project, a so-called 'specific support action' within the European Information Societies Technology programme, will over the two years 2006-2007 be seeking to define a research road map that will allow not only the technology to be developed but the social issues also to be addressed, with the goal of establishing a healthgrid as the infrastructure of choice for European biomedical activity in the next ten years. The SHARE collaboration includes both computer scientists, experts on social requirements and medical law specialists. The project begins with the fundamental assumption that technical and social requirements must be addressed concurrently. It has identified these challenges to the modernization of health systems [7]:

- creating and populating, connecting and understanding patient records across organization boundaries and, in due course, across different national health systems;
- increasing the openness and accessibility of systems - e.g. providing patients with ownership of their healthcare record - while
- ensuring privacy, confidentiality and ethical compliance in the socio-legal plane, and
- maintaining data integrity, security and authenticity (e.g. provenance and semantics) in the technical plane;
- providing appropriate levels of authorization and authentication of users across all the services and the citizen;
- discovering, grading and certificating trustworthy sources of knowledge and case information to guide future action; finally,

– winning the trust and commitment of the medical professions at a time of immense change and economic pressure.

At present it seems unlikely that technology will be allowed to determine answers to questions of a legal nature, much less so of an ethical nature. Yet the extent to which we trust financial affairs to the internet and the extent to which we have allowed privacy to be invaded by online transactions, 'cookies' and preference tracking (to say nothing of store loyalty schemes) [8] suggests that we may be more flexible in our attitudes that our legal attitudes may imply. Indeed, as far as personal data are concerned, the financial analogy has been made before in the concept of a personal data bank. Would patients be less trusting of a 'bank' with their health record than they are with their money?

I have argued that 'healthgrid', the augmented application of grid computing to health, presents an opportunity to review not only information technology for health – a major enough task – but also our approach to the complex issues of ethical, legal and regulatory compliance as mediated by the technology. The case in favour of the technology, in terms of improved information and knowledge for clinicians, patients, public health officials, administrators and governments, is not difficult to make. The need for ethical and legal safeguards cannot be circumvented, but in itself this may prove an insuperable obstacle for the deployment of the new technology. One way forward is to analyse precisely these 'social' requirements and enhance the technology with the means to apply them automatically with minimal human intervention.

Acknowledgements

I am deeply indebted to my colleagues, in chronological order, on the MammoGrid, SHARE and Health-e-Child projects; also to the HealthGrid organization and the co-authors of the 'White Paper'; finally to the HealinG consortium. In relation to legal, ethical, security and trust issues, I must particularly thank Brecht Claerhout, Jean Herveg and James Lawford Davies — though I alone am responsible for my mistakes!

References

[1] *Frequency of Cancers Around the World*, The Scientist, Vol. 17, Cancer Supplement, 22 09 2003, at: http://www.the-scientist.com/article/display/14131/

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

[3] Warren R *et al.*, *A comparison of some anthropometric parameters between an Italian and a UK population: 'proof of principle' of a European project using MammoGrid* To appear, 2006.

[4] Warren R *et al.*, *A Prototype Distributed Mammographic Database for Europe* To appear, 2006.

[5] The Information Societies Technology Integrated Project: *Health-e-Child – An integrated platform for European paediatrics based on a grid-enabled network of leading clinical centres*, EU Contract Number IST-2005-027749.

[6] The Information Societies Technology Specific Support Action: *SHARE*, EU Contract Number IST-2005-027694.

[7] SHARE Consortium, *The Healthgrid Framework*, not yet published.

[8] David H Freedman, *Why Privacy Won't Matter*, Newsweek (International Edition), 3rd April 2006; at http://www.msnbc.msn.com/id/12017579/site/newsweek/.

Patient-Centered Outcomes Research in Practice: The CAPriCORN Infrastructure

Anthony Solomonides[a,*], Satyender Goel[b], Denise Hynes[c,l], Jonathan C Silverstein[a], Bala Hota[d1], William Trick[e], Francisco Angulo[e], Ron Price[f], Eugene Sadhu[c], Susan Zelisko[f], James Fischer[c], Brian Furner[g2], Andrew Hamilton[h], Jasmin Phua[i], Wendy Brown[j], Samuel F Hohmann[k,d2], David Meltzer[g1], Elizabeth Tarlov[c,l], Frances M Weaver[f,l], Helen Zhang[e], Thomas Concannon[m], Abel Kho[b,*]

*[a] Center for Biomedical Research Informatics, NorthShore University HealthSystem; [b] Feinberg School of Medicine, Northwestern University; [c] University of Illinois, Chicago; [d] Rush University Medical Center—[1] Department of Medicine, —[2] Dept of Health Systems Management; [e] Cook County Health and Hospital Systems; [f] Loyola University Health System; [g] University of Chicago—[1] Medicine, —[2] Center for Research Informatics; [h] Alliance of Chicago Community Health Services; [i] Medical Research Analytics and Informatics Alliance; [j] VA Jesse Brown Hospital; [k] Universities Healthsystem Consortium; [l] VA Edward Hines Hospital; [m] RAND Corporation. *Corresponding authors*

Abstract

CAPriCORN, the Chicago Area Patient Centered Outcomes Research Network, is one the eleven PCORI-funded Clinical Data Research Networks. A collaboration of six academic medical centers, a Chicago public hospital, two VA hospitals and a network of federally qualified health centers, CAPriCORN addresses the needs of a diverse community and overlapping populations. In order to capture complete medical records without compromising patient privacy and confidentiality, the network has devised policies and mechanisms for patient consultation, central IRB approval, de-identification, de-duplication, and integration of patient data by study cohort, randomization and sampling, re-identification for consent by providers and patients, and communication with patients to elicit patient-reported outcomes through validated instruments. The paper describes these policies and mechanisms and discusses two case studies to prove the feasibility and effectiveness of the network.

Keywords:

Patient-Centered Outcomes Research; Comparative Effectiveness Research; Electronic Health Records; Data Collection, —Linkage , —Aggregation, —Sets; Deidentification, Re-identification; Consent.

Introduction

PCOR, CER and PCORnet

The Patient-Centered Outcomes Research Institute (PCORI) was established following the US Patient Protection and Affordable Care Act in 2010. Its mission is to advance and support Patient-Centered Outcomes Research (PCOR), which

[…] helps people and their caregivers communicate and make informed healthcare decisions, allowing their voices to be heard in assessing the value of healthcare options. [1]

In particular, PCOR:

- Encompasses comparative effectiveness research (CER) on interventions to inform decision making.
- Addresses individuals' (especially patients' and caregivers') preferences and autonomy.
- Studies a diversity of settings and populations.
- Seeks to balance stakeholders' concerns, including burden to individuals and availability of resources.

One of the principal means by which PCORI has sought to achieve these goals is by supporting eleven Clinical Data Research Networks (CDRN) and eighteen Patient-Powered Research Networks (PPRN). Both kinds of research network are seen as infrastructure-building projects, with specific structural, process and outcome goals to prove the feasibility and usefulness of the networks. CDRNs are focused on major academic medical centers; apart from demonstration of viable infrastructures, they are expected to demonstrate their value by conducting research in a number of specific conditions. Each network has had to nominate the conditions on which it will work. However, longer term sustainability for the infrastructure can only be achieved through manifest success in these early studies, by proving to the research community that the network represents a valuable resource that is worth both exploiting and supporting through further funded studies and grant proposals. PPRNs are focused on specific conditions, some relatively common and some rare, that are of particular concern to patients, carers, and patient advocacy organizations. Many have formed around existing formal or informal networks of support and advocacy groups.

Overarching the CDRNs and PPRNs, PCORI has established a supra-network, PCORnet, that acts as collaboration venue, clearing house, and policy-development body on behalf of all. Best conceived of as a network of networks, it ensures that the infrastructures created by the different CDRNs and PPRNs will remain interoperable and responsive both to researchers' needs and to the expectations of patients, carers and advocates.

CAPriCORN

One of the CDRNs, CAPriCORN, represents a remarkable alliance of Chicago institutions collaborating in recognition of the need for pre-competitive comparative effectiveness research (CER) in their highly diverse community—diverse both in the type of institutions involved and, importantly, in the populations they serve. It is not altogether typical of CDRNs, although it naturally shares many characteristics. Some of its unique features have provided a model for collaboration in environments where, for example, patient populations at different institutions overlap, as they do within a city setting, where nevertheless a full picture of each patient's health record is necessary for meaningful research results.

Data for sharing within CAPriCORN—and in the wider community at a later stage—will be in a HIPAA-compliant, de-identified format. Two working groups, Informatics WG and Ethics and Regulatory WG, have devised a federated data architecture, a data model with appropriate standards, and a designed data flow engineered to ensure that no protected

health information (PHI) is released other than under strictly controlled conditions, at the same time as maintaining the research value of the data that is released. De-identified data will be released on a study-by-study basis. A statistically benchmarked process is used to generate a pseudonymous identity for each patient in such a way that patients' records that are distributed across different providers in the network can be matched and integrated, not by being brought together into a single central database, but in a virtual repository – by allowing distributed queries across the different systems through the validated mechanism of PopMedNet [4, 5]. Consent will be sought when access to PHI or directly to the patient for patient-reported outcomes is necessary.

Methods

Population

CAPriCORN comprises a network of six academic medical centers (University of Chicago, University of Illinois, Chicago, Loyola University, NorthShore University HealthSystem, Northwestern University and Rush University Health), the Alliance of Chicago's Federally Qualified Health Centers, a major public hospital, Cook County Hospital, and two Veterans Affairs hospitals, VA Edward Hines and VA Jesse Brown. Geographically, these serve the greater Chicago metropolitan area and are available to a total population of approximately 9.5 million. (In addition to these "data-providing" institutions, 22 other organizations contribute research, patient advocacy, and infrastructure services to CAPriCORN. Their role is described below.)

At the time of proposal submission, CAPriCORN institutions among them held 2,860,000 covered lives in electronic health records. A preliminary analysis of seven of the ten institutions indicated 6,923,111 patients, of whom 1,465,285 were registered with a primary care provider; however, after de-duplication, the numbers were 5,741,268 and 1,242,380 unique patients respectively. Thus some 20.6% of patients are associated with more than one institution, and even among the primary populations, there are 18% of patients with more than one PCP registration. This appears to be symptomatic of deprivation in the inner city, where of economic necessity individuals move opportunistically from provider to provider.

The racial breakdown of the primary population is 47.5% Caucasian, 27.9% African American and 14.9 Hispanic, with just over 9% in other categories. Of this population, 59.3% are female, 40.7% male. The mean age is 50 with a standard deviation of 17.9.

De-identification and De-duplication

While fragmented care may be suboptimal, research on comparative effectiveness of treatments requires as accurate and as complete a record of each patient's health status and episodes of illness as can be reconstructed, if meaningful and valid results are to be achieved. With multiple records for up to 20% of patients, de-duplication is strongly indicated. The means of achieving this lie in a particular method of de-identification.

In the US context, there is currently little prospect of a single unique patient identification code. Where health information exchanges have been instituted, it has been necessary to implement an "enterprise master patient index" (EMPI), but even these are rare because of a number of concerns, principally privacy and security, and economics and sustainability. Nevertheless, prior experience was sufficiently encouraging to suggest that a specific design and implementation in the Chicago area would be worthwhile. This prior knowledge and experience provided a fundamental cornerstone for the CAPriCORN network.

The de-identification algorithm is due to Kho *et al* [2, 3]. It uses a set of strictly personal identifiers, i.e. including nothing that may be institution-specific, to generate up to 17 different combination strings and uses a statistically selected subset of these to construct a "hash-ID". As its name implies, the hashing algorithm is not reversible, but its high specificity allows patients who have multiple records to be discovered, albeit anonymously.

Organizational Design

As a project, CAPriCORN is led by a Principal Investigator at the Chicago Community Trust, an organization focused on civic leadership and philanthropy. A Steering Committee is the decision making body, whose composition has been designed around the natural concerns of a network to conduct and facilitate patient-centered outcomes and comparative effectiveness research across a number of healthcare institutions, and also reflects the underlying architectural design of the infrastructure and the projected governance and regulatory framework of that infrastructure.

Clinical Data Research Networks are by definition intended to be open to external collaboration, are explicitly designed to be open to patient concerns, and are subject to all the normal ethical and regulatory processes that apply to human subjects and social science research. These are, respectively, reflected in the network's External Researcher Committee, Patient and Clinician Advisory Committee, and Chicago Area Institutional Review Board (CHAIRb). All these figure in the definition of processes and workflows for patient and carer consultation, for the triage of internal and external research proposals, for the handling of data requests, for the release of data, and for the consenting process prior to any re-identification of and contact with patients.

Critical to the infrastructural design are two "honest broker" roles in the network. Other than in very specific, precisely defined circumstances involving only consented patients, these organizations hold no protected personal health information (PHI) but handle the "de-identifiers", principally the hash-IDs for de-duplication, and subsequent to the definition of specific condition cohorts, a second level of pseudonymization, the cluster-IDs, which are randomly generated "per study, per hash-ID" thus avoiding any unintended crosstalk between independent studies.

The principles, some explicit and some implicit, that have guided this design are:

- All studies, including those that have been submitted as "proof of principle" for the network, along with new and external proposals, will be subject to triage by the Patient and Clinician Advisory and External Researcher committees, then subject to review by CHAIRb, with the ultimate decision resting with the Steering Committee.
- All PHI will be held at institutions, benefiting from all the protections (firewalls, authorizations, etc.) that each applies to its own patient data.
- The data to be collected by an honest broker will be strictly non-PHI and will be minimal with respect to any cohort identification needs (all that is needed, but no more).
- Identifiers will be hashed into pseudonymous "hash-IDs" for the purpose of de-duplication. Honest Broker 1 (HB1) will provide institutions with a

unique "hash seed" that each will use to de-identify through hashing its own patients.

- The second honest broker, HB2, will use the hash-IDs provided by institutions to identify "duplication" and determine the set of institutions to which each patient corresponds. HB2 then generates a random identifier, the cluster-ID, for each unique patient in the given cohort. At this point, if considered necessary, the institutions themselves may be pseudonymized. (No PHI will flow to HB2.)
- Patients' records may only be linked through the hash-ID. Cohort identification for specific studies and specific (non-PHI) data requests from sites for the purpose of constructing aggregate records may be conducted only by means of a distributed query mechanism (currently, PopMedNet [4, 5]) which allows queries to be inspected and vetted prior to execution and results from queries to be examined prior to release.
- All studies that require access to PHI must identify a co-investigator at each site.
- Provider consent to approach patients to consent for particular studies will be requested, and subsequent patient consent will be sought, according to institutional rules and norms.
- Randomization of patients for consent will be done anonymously both in respect to patients and institutions.

As noted above, these principles are visible in the organizational structure of the network, but they are also evident in the architectural design of the infrastructure.

Network Architecture

The architecture of the network is depicted in Fig. 1 below. The processes represented by the various flows in this diagram are detailed below.

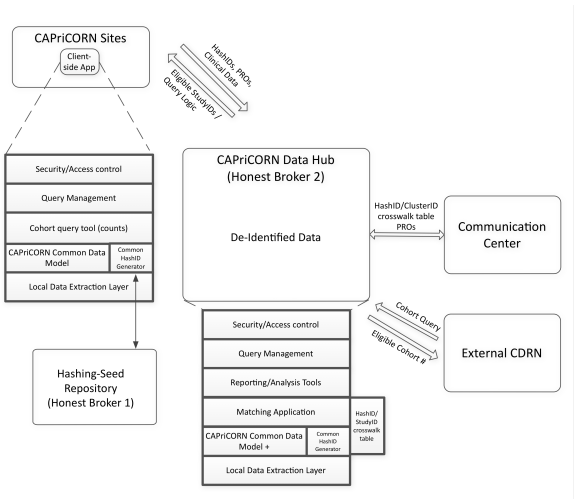


Figure 1 – A schematic diagram of the network displaying the two "honest broker" roles, the institutional repositories and the central "data hub" which hosts the matching and distributed query services.

CAPriCORN has developed a data model and data standards, together with "extract-transform-load" processes for its institutional data marts. The data model is effectively based on a star schema with the concept *Encounter* at its center, so that data can be understood at a transactional level. A data dictionary has been adopted showing domains and variables within them (apart from patient demographics, radiating out from encounters are diagnoses, medications, procedures, vital signs, laboratory results, and some additional local variables). Standards and terminologies are indicated for values in each category. The degree of privacy restriction for each variable (within-institution, within-CAPriCORN, within-PCORnet) is also indicated.

Each institution has established a data mart (or other local database) which, notwithstanding the differences in platforms, precisely matches the CAPriCORN data model. Thus, although local adaptations of SQL queries will be necessary, the essential logic of queries submitted to the "data hub", i.e. the distributed query service, will remain unaltered, as required by PCORnet for its greater vision of seamless patient-centered, comparative effectiveness research.

A Communication Center is also being established to facilitate the process of re-identification of patients for provider consent to approach patients and for patient consent to participate in survey research (patient-reported outcomes, or PROs) and intervention studies. Each institution's processes are respected, and no pre-consent PHI flows through the center.

Process Description

1. HB1 hosts a stand-alone, generic hashing-seed generator application; it generates a SEED and passes it automatically to all participating institutions.
2. Each INSTITUTION uses the SEED and a set identifiers to generate a set of multiple *hashes* for each patient on record:

   [SSN, FirstName, LastName, DoB, Gender] ⊗ SEED
   → { hashes }

   from which a unique hash-ID is generated and cross-linked to the patient's MRN for internal identification.

This is per patient; […] signifies a vector of personal data.

Hash-IDs can be used within each INSTITUTION locally, if desired.

3. For each STUDY, every INSTITUTION runs the appropriate phenotyping algorithm to select its subpopulation of all unique patients who satisfy the cohort criteria. The hash-IDs along with all the hashes are returned to HB2.
4. For each study, HB2 collects all hashed data and de-duplicates, storing the result in a vector as follows:

   { (institutionID=1) : hash-ID$_1$ }
   $\Diamond_{\text{hash-ID}} \cdots \Diamond_{\text{hash-ID}}$
   { (institutionID=10) : hash-ID$_{10}$ }
   → hash-ID : institutionVector

where $\Diamond_{\text{hash-ID}}$ represent the join on hash-ID. The patient's hash-ID and institutionVector now appear thus:

	Institutions									
Disease D	AL	CC	UC	UI	LU	NS	NU	RU	VH	VJ
hash-ID										
xyz123	0	0	0	1	0	0	0	1	0	0

This represents the following "facts": The patient whose hash-ID is "xyz123" has been identified as having disease D and having partial records at UI and RU. We note that

(i) the hash-ID is in reality a more complex object (cf. [2]);

(ii) this may not be the complete record for this patient.

5. The five collections { hash-ID }, one for each study, are returned to all the institutions for cohort verification.

This is necessary, because, for example, a patient with an anemia record at one hospital (RU) may turn out to have a record at another hospital (UI) that does not mention anemia. Nevertheless, a complete record for that patient must include the partial records from both institutions.

6. Each institution checks the lists against its reference hash-ID list and so completes each patient's record if necessary.

For the sake of illustration, suppose now that we have found the patient above has also been seen at yet another hospital (CC) for an unrelated condition. The vector now becomes:

Disease D	AL	CC	UC	UI	LU	NS	NU	RU	VH	VJ

hash-ID										
xyz123	0	1	0	1	0	0	0	1	0	0

We can now confidently compile a complete record of the patient.

7. At this point, HB2, as an honest broker, must do two more de-identification steps:
   a. first, to disguise the institutions, and
   b. second, to replace hash-IDs with non-derived ids for the patients; these are the cluster-IDs.

For the first step, HB2 randomly assigns pseudonyms to the institutions, say:

AL	CC	UC	UI	LU	NS	NU	RU	VH	VJ
ff	*dd*	*aa*	*jj*	*bb*	*ii*	*cc*	*ee*	*hh*	*gg*

and these are then indexed as:

aa	*bb*	*cc*	*dd*	*ee*	*ff*	*gg*	*hh*	*ii*	*jj*
UC	LU	NU	CC	RU	AL	VJ	VH	NS	UI

The example patient now appears as:

Disease D	*aa*	*bb*	*cc*	*dd*	*ee*	*ff*	*gg*	*hh*	*ii*	*jj*

hash-ID										
xyz123	0	0	0	1	1	0	0	0	0	1

   c. The hash-IDs for each study cohort can now be replaced with unique cluster-IDs.

Our example patient now appears as:

Disease D	*aa*	*bb*	*cc*	*dd*	*ee*	*ff*	*gg*	*hh*	*ii*	*jj*

cluster-ID										
D-900093	0	0	0	1	1	0	0	0	0	1

Now, only possession of the table converting hash-IDs to cluster-IDs can enable anyone to re-identify the patient.

Distributed Queries

With cohort cluster-IDs collected, HB2 can route data requests through the distributed query service to the institutional data marts (IDMs). Locally, each institution will have the opportunity to determine if the proposed query against its IDM is acceptable, allow it to execute, and even then scrutinize the results before releasing them. Both in sending the requests and as results are received, HB2 can match cluster-IDs to hash-IDs, so that even a clinician researcher working on a project in their own specialty may be able to view expanded records of their own patients without recognizing them as their own. This provides a very high standard of de-identification.

Re-identification

Once particular studies based on entire cohorts are launched, it is likely that re-identification of subsets of patients may prove necessary. Having received approval both from the Steering Committee (with advice from PCAC and ERC) and permission to proceed from CHAIRb, a researcher may request the Communication Center to randomly select a possibly weighted sample from across institutional or other populations for re-identification. The researcher will also be able to submit, through HB2, a data request for controls. It is possible, subject to CHAIRb's approval, for institutional processes to be employed to gain provider consent and from there patient consent to participate in a study. Given the cluster-IDs of the patients in the study group, the Communication Center can alert institutions to the hash-IDs of patients to be approached for re-identification. In some cases, the Communication Center will also provide institutions with the means to collect patient-reported outcomes.

In the case of patients attending multiple institutions, which institution (or more precisely, which provider) should consent the patient for an identified study may be complex. A variety of algorithmic approaches is possible, including some that may work well but are computationally expensive. This may take the form of querying the system for the number of encounters at each institution in the last year (complex, but likely to reflect the patient's expectation) or it may suffice to look where the patient is registered for primary care (inexpensive, but may be irrelevant). The present ruling of CHAIRb only constrains the approach to be through a provider who is actually involved in the patient's care.

Results

At this, approximately halfway point in the project, achievements across a number of fronts include:

- Establishment of a sound governance structure, including a common central IRB, with data use and business associate agreements in place.
- Establishment and launch of a Patient and Clinician Advisory Committee with a clear role in the review, triage and approval of new research proposals and a comprehensive manual for its operations.
- Agreed design for the technological infrastructure, including a data model designed for ease of distributed query as well as with model evolution in mind.
- Agreed processes and workflows now increasingly described and approved in protocols.
- Preliminary tests of the de-identification process and the distributed query machinery.

- Preliminary phenotyping in all five study cohorts proposed at project submission (see below). Preparatory phenotyping for a number of other studies, including incidental findings in osteoporosis, the national aspirin trial, bariatric surgery, antibiotics and childhood weight, bisphosphonates, and others.
- The de-identification and de-duplication processes in CAPriCORN are increasingly being looked at as a model to be replicated across other CDRNs.

The internal organization of the network lends itself well to establishing CAPriCORN as a corporate entity; this would no doubt present new challenges, but is under consideration.

Discussion

The data model has been deployed at institutions to construct a data mart. Based on model variables, five phenotyping algorithms have been devised and tested at multiple sites to identify overweight and obese patients (as required of all CDRNs); ambulatory patients suffering from asthma and in-patients with anemia (the two common disease cohorts); and patients with recurrent *Clostridium difficile* infection (RCDI) and sickle-cell disease sufferers (the two rare conditions).

In preparation for all these studies (and other anticipated future studies, including the PCORnet-inspired Aspirin trial and various collaborations with other CDRNs and PPRNs) the central IRB, CHAIRb, has already reviewed a Master Protocol which serves as a prefix to all specific study protocols.

Extract-Transform-Load (ETL) processes have been under-taken against a number of different proprietory EHR systems. Some of these have been shared publicly (e.g. through an EHR vendor's community sharing portal, thus conforming with requirements of commercial confidentiality). ETL logic has been shared among all data-contributing sites to ensure compatibility.

The CAPriCORN data model is a superset of the PCORnet common data model against which external requests will be formulated. This makes for a straightforward mapping of data and requests from PCORnet to CAPriCORN. Additional data models that are influencing the central PCORnet design, such as (Mini-)Sentinel, OMOP, i2b2 and others have also been studied with a view to establishing correspondences should the opportunity of collaboration make a translation between CAPriCORN and another data model desirable.

Among the proposed cohort studies, the case of RCDI provides a convenient example of a hard test-case for the infrastructure. The study has not yet been completed, but based on data stored according to the data model and addressing queries to pre-existing institutional data warehouses rather than the institutional data marts, accurate cohort counts have been achieved.

Index cases of CDiff infection have been identified, either by the presence of a diagnosis code or by laboratory test results. The first difficulty arises in recognizing resolved CDiff infection: how to differentiate between refractory and recurrent infection. If there is no encounter with CDiff code, laboratory test or relevant medication within eighteen days of date of diagnosis or of positive test result, the infection is assumed to have cleared. Any further infection in 18 to 56 days post index date is recorded as recurrence. Infections later than 56 days are considered new rather than recurrent.

One of the key challenges to CAPriCORN's distributed architecture will be in the identification of recurrence across institutions. This has not yet been attempted, but will be among the first studies that the system will address. The cohort is anticipated to be relatively small and the cases of patients moving from one institution to another while at risk of recurrence of CDiff should be fewer still, so that discovery of such cases will represent success with truly rare events.

Conclusion

Along with ten other CDRNs, CAPriCORN is at about the halfway point of its "Phase I" life span and is ready to test its systems with real use cases. The infrastructure has been designed to allow for evolution in the data model and increasing complexity of queries in future. Five submitted cohort studies are currently being processed through stages of the CAPriCORN workflow, and a number of new study proposals are being prepared.

The processes of de-identification, matching and de-duplication, cohort identification, record linkage and aggregation, and the distributed query mechanism have been described. It is possible to randomize and re-identify securely, and to extract matched controls through records in the IDMs.

Sustainability of the architecture will be demonstrated through a number of additional research studies that had not been considered at the proposal stage. These are also providing a valuable challenge to CAPriCORN's proposal triage, patient-centeredness, and external researcher engagement workflows.

References

[1] Patient Centered Outcomes Research Institute. PCOR. http://www.pcori.org/content/research-we-support (Accessed on December 22nd 2014.)

[2] Abel N. Kho, John P. Cashy, *et al*. Design and Implementation of a Privacy Preserving Electronic Health Record Linkage Tool in Chicago. (Under review at JAMIA.)

[3] HealthLNK Data Repository. www.healthlnk.org (Accessed on December 22nd 2014.)

[4] Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, Brown JS. Design of a national distributed health data network. Ann Intern Med 2009; 151:341-4.

[5] PopMedNet. http://www.popmednet.org (Accessed on April 9th 2015).

Acknowledgements

The authors are principal informaticians on the CAPriCORN project. Acknowledgement is due to the overall project PI, Terry Mazany of the Chicago Community Trust for his leadership, and to site PIs: other than those listed as co-authors, Fred Rachman of the Alliance, Raj Shah at Rush, and Brian Schmitt at VA Hines; also Jerry Krishnan who was the original site PI at UIC. Doriane Miller and Madeleine Shalowitz direct the PCAC and have provided great clarity on patient engagement. Jonathan Tobin (NYC CDRN) has been a great supporter with ideas and proposals in the External Researcher Committee. Last but not least, John Collins and Shelly Sital at the Illinois Medical District Commission have provided a consistently high standard of project support.

Address for correspondence

Anthony Solomonides, Center for Biomedical Research Informatics, NorthShore University HealthSystem, 1001 University Place, Evanston, IL 60201, USA

asolomonides@northshore.org

Abel Kho, Division of General Internal Medicine, Northwestern University Feinberg School of Medicine, Rubloff Building 10th Floor, 750 N Lake Shore Drive, Chicago IL 60611, USA

Abel.Kho@nmff.org

The Learning Patient in the Learning Health System

Anthony Solomonides[a]

[a] *Department of Family Medicine, NorthShore University HealthSystem, Evanston, Illinois, USA*

Abstract

Recent experience has shown that stakeholders can be powerful allies when they are authentically integrated into the research process. Patients, advocates and caregivers bring a wealth of views and viewpoints to bear on research questions. While reimbursement models often inhibit research engagement by physicians, we argue that patients can be a valuable resource. That we often forget this seems all the more paradoxical in view of our oft-asserted "patient-centered"-ness. We consistently neglect the patient, and more broadly the stakeholder, as a resource in research. Is there another way to look at this? We suggest "imagineering" in its pre-Disney sense: we can imagine a world in which patients play a much bigger role in the management of their health records—bypassing the fraught issues of ownership and custody—and a world in which patients have a means of subscribing to research as part of the management of their record. This would mean having options to receive bulletins about projects and results, information about upcoming studies, with the opportunity to choose studies in which to participate, perhaps subject to screening by a physician. Beyond this, for some it may mean engagement in the research process – formulation of research questions and goals, or participation in analysis in the spirit of "citizen science".

Keywords:

Patient participation, patient-centered outcomes research, informed consent.

Introduction: The Patient in the Learning Health System

This vision paper aims to bring together a number of parallel currents of thought in healthcare, biomedical research, informatics, and recent trends in consenting and institutional review. Its convergent vision is to relate these strands: the Learning Health System community [1]; the Patient-Centered Outcomes Research paradigm [2]; the Health Data Bank concept [3]; the move towards reform of the consent (or e-consent) and institutional review board processes [4, 5]; and patient engagement strategies through education to play a part in their own and others' care, and by extension, in research. [6, 7] The principal goal of this paper is to demonstrate that the necessary elements are available and that integration is necessary to make the vision a reality.

Values

The Learning Health System (LHS) originated in the work of Charles Friedman, and matured into a fully fledged idea his time as Chief Technology Officer at the Office of the National Coordinator for Health IT. His joint paper with Adam K. Wong and David Blumenthal [8] in Science Translational Medicine provided exemplars rather than a succinct definition of the Learning Health System and made heavy use of the "Meaningful Use" paradigm that really aimed at the conceptually simpler level of Health IT dissemination and adoption. Nevertheless, the LHS movement has since expanded to encompass many aspects of biomedical research activity.

Among its ten "core values" [9] are three that are particularly pertinent to the vision presented here. These are listed here retaining the original numbering of core values: The LHS will be

> **1. Person-Focused:** … *informing [individuals of] choices about health and healthcare. … through strategies that engage individuals, families, groups, communities, and the general population, …*

The LHS will be characterized by:

> 3. **Inclusiveness:** *Every individual and organization committed to improving the health of individuals, communities, and diverse populations, who abides by the governance of the LHS, is invited and encouraged to participate.*

and by

> 8. **Cooperative and Participatory Leadership:** *… [through] a multi-stakeholder collaboration across the public and private sectors including patients, consumers, caregivers, and families, in addition to other stakeholders. … Bold leadership and strong user participation are essential keys to unlocking the potential of the LHS.*

Our proposal espouses all the core values of the LHS; these three are highlighted here simply because of their particular relevance. In arguing for stakeholders—patients, caregivers, advocates—to be involved in both healthcare and research processes, we acknowledge that this can only happen if there is a commitment to inclusiveness and to participatory direction of activities. There is already experience, not least from PCORI-funded work through PCORnet [10], that stakeholders can make a significant contribution to the generation of relevant research questions.

Learning

Learning in the context of self-management has considerable history and literature, both promoting the concept and skeptical of it. It is sometimes said that "we used to do things *to* patients, then we did things *for* patients, and finally we want to do things *with* patients". This is not to be read as some sort of naive history of progress in healthcare; but it does

reflect the way in which rapid progress in the medical sciences, the rising share of gross domestic product dedicated to health care, and the "consumer revolution" have shaped the current landscape.

It is as well to remember how each form of care was conceived and justified: in the first place, the more objectified the patient, the less personal the engagement, the more collected and unbiased the physician would be—so we thought. Engagement would mean the physician was over-invested and would lose "his" (more often than not) objectivity.

The consumer revolution resulted in patients looking upon healthcare and the maintenance of their wellbeing as a consumer good. *I pay for my healthcare in much the same way as I pay for many other things, so I expect to be provided with a good service. And by the way, my health records are mine and I can take my business elsewhere whenever I choose.* And this resulted in many taking an arms-length stance towards their own health – there will always be medications to keep any condition, from indigestion and reflux to diabetes and hypertension, under control, as direct-to-consumer advertising still suggests.

As the cost to patients and health systems rises, there is a clear need to engage patients in their own healthcare. We want patients to be more engaged and more invested in their own health. There is a sense in which patients are more invested than ever, but there is also some evidence that positive rather than negative motivation is likely to be the more effective. We do expect patients to take responsibility for themselves, even in the face of poorly understood non-compliant behaviour (e.g. we deal with asthma sufferers who smoke and are not persuaded to stop by enumerating the dangers or even the self-evident discomfort they endure). We assert a new imperative, that of partnership with the patient, because learning provides the only way forward. This paper asserts the need to extend this relationship to support for research.

Beyond "Seventeen Years": From evidence to practice

It is sometimes asserted that "17 years elapse before a new element of validated clinical knowledge finds its way into routine clinical practice in the United States" [8, 11]. Although this is now widely seen as an oversimplification, it is still the case that new drugs typically require in excess of ten years before reaching the patient, while some medical devices may only require seven or fewer years to come through. These are still surprisingly large numbers, when one considers the focused effort that goes into the development and testing of a drug or device. [12]

Conversely, there have been several successful programs of data gathering from practice, including monitoring of drugs post-marketing, demonstrating a value in pragmatic studies of outcomes and of comparative effectiveness. For example, the FDA's Sentinel Initiative page asserts [13]

> *Sentinel enhances the FDA's ability to proactively monitor the safety of medical products after they have reached the market and complements the Agency's existing Adverse Event Reporting System. Through Sentinel, the FDA can rapidly and securely access information from large amounts of electronic healthcare data, such as electronic health records (EHR), insurance claims data and registries, from a diverse group of data partners. Sentinel uses a distributed data approach which allows the FDA to monitor the safety of regulated medical products, while securing and safeguarding patient privacy.*

The PCORnet approach has demonstrated a further value in a similar approach: involving stakeholders in creating a vision for a study, participation in expressing research questions and setting goals, in the collection of data, in reviewing technical analyses, in considering how findings may be translated into practice, in evaluating how effective an intervention is proving to be – for patients, caregivers and providers – and in maintaining a bond between a community and a health system. This approach to participation also provides a transparent means to address gender and racial/ethnic diversity, to acknowledge patients and other stakeholders in discourse, and to reflect the value of a "skilled PCOR community". [10]

The PCORI community has already demonstrated the value of its participatory approach, both through its large-scale infra-structure projects (e.g. the thirteen Clinical Data Research Networks), the demonstrator projects undertaken over this infrastructure (e.g. studying the effects of antibiotics in infancy on subsequent growth patterns), and individual grants to specific, locally devised projects, such as a study of women with depression in obstetrics and gynecology practices [14] and the CHICAGO study of racial disparities in asthma-related visits to the ED, where reportedly there is a *five- to seven-fold higher rate of visits to the ED for uncontrolled asthma in communities with a high proportion of African–American and Latino children compared with other communities.* [15] In both these last cases, engagement of stakeholders in the design process proved crucial to success.

A parallel activity has taken place in oncology, apparently arising out of the needs of that community and with little correspondence with the PCORI world. There we find a head-to-head comparison of Phase I-III trials with comparative effectiveness studies. [16]

Health Record Banking

Fuller engagement of patients both in their own healthcare and in research requires reasonable access to the medical record. We have learned from the PCORnet experience, if we did not already know, that stakeholders can be powerful allies when they are authentically integrated into the research process. Patients, advocates and caregivers bring a wealth of views and viewpoints to bear on the issues. While reimbursement models inhibit rather than support research engagement and physicians are too pressed to go the extra mile to support research, it is odd that we may forget patients as a resource. This is all the more paradoxical in view of our oft-asserted "patient-centered"-ness. We consistently neglect the patient, and more broadly the stakeholder, as a resource in research. (It has been observed that this may replicate itself even in PCORnet, where recently presented sustainability plans appear significantly more focused on the CDRNs than the PPRNs.)

Is there another way to look at this? Perhaps "imagineering" is what is needed here (in the sense it had before Disney trademarked it): *What we can imagine, we can attempt to engineer.* What we can imagine is a world in which patients play a much bigger part in the *management* of their health records, setting aside the issues of *ownership* and *custody*. Through "management" we capture the two senses of *maintenance* and *control of distribution* beyond immediate healthcare needs. This much we can at least imagine.

We may begin with the current of thought—not yet a movement—to have patients maintain their records. There has always been an argument, with good cause on both sides, for and against patients' "ownership" of their records. While the rhetoric has always been that the patient owns his or her medical record, the custodian of the record has always had to be a provider, whether an individual practice or a large institution. The idea of health record banking has its roots at least as far back as the 1990's when the idea sprang up both in the US and in the UK, and has more recently been the subject of several papers and a focal point in both Dr. Patricia Brennan's and more particularly Dr. Amnon Shabo's keynote lectures at MedInfo 2015.

We may not be quite there yet, but as patients and citizens in general collect increasing amounts of data about their health ("quantified self", wearables, fitness devices, etc.) the uptake of personal health banking is likely to increase. This presents an opportunity for researchers, but it will require a different attitude to consenting than the current standard. We need a consenting policy that allows patients and researchers to maintain a relationship to a degree independent of that with their health care provider, but without excluding the provider.

It should be said here parenthetically that Dr. Bill Dodds [17], the Scottish GP who first proposed a "Health Information Bank"—an institution made up of two non-profit and one for-profit corporations—foresaw many of the issues that would have to be addressed and devised a clever structure to address most, if not all of them. Of particular note are the primary data bank whose job it is to hold the data and manage it on a kind of mutual banking basis for the benefit of all clients; an academy which would bear the burden of ethical management and regulatory compliance, and be the vehicle for research. The for-profit corporation would deal, within the constraints of the other two, with the commercial exploitation of the data. The questions we wish to pose are:

- Can the data bank concept be extended to become a consent management system?
  - the patient gets to choose what studies may be of interest to him or her;
  - the patient also has the choice of how deeply to engage in a study; e.g. may provide specialist support if the patient happens to have the necessary skills, say statistics, or may work on the dissemination plan.
- Can the benefits gained from banking be less focused on the financial and more on the additional information that can accrue to the interested patient? E.g.
  - links to relevant articles at the level of the patient's choice (from links to the day's newspaper to PubMed references);
  - health messages and alerts – links to m-Health.

Patient Data as Commodity

Thus we can also anticipate a world in which patients have a means of subscribing to research—in general—as part of the management of their record. What would this mean? It would mean having options to receive bulletins about projects and results, information about upcoming studies, with the opportunity to choose studies in which to participate, perhaps subject to screening by a physician.

But there is dystopian vision also. Much of the discussion of "commodification" of patient, and more generally, personal data revolves around the use that those who gather data on a large scale make of this wealth of information, or more precisely, about the ways they turn data, often of uneven quality but in large volumes, into information with intrinsic value. The popular examples, such as the off-the-Wal-Mart "beer and diapers" story or the uncanny accuracy of Target's pregnancy predictor, emphasize the value these corporations respectively extracted from the data by using analytics to turn it into worthwhile information. It has, of course, been said that some of these stories have been overhyped, and more recent examples, such as Google's apparent ability to recognize influenza epidemics on the basis of search terms entered by users have also been questioned. Nevertheless, the terms "big data", sometimes capitalized or depicted as a massive monolith, and "analytics" are ubiquitous in the popular technical literature; one suspects that this degree of excitement and volume of investment must be a reflection of excellent results.

In healthcare, analytics has been applied to service improvement in hospitals and other provider organizations. The large-scale distributed research data repositories currently envisioned by such projects as the PCORI Clinical Data Research Networks, are expected to bring value to healthcare delivery through comparative effectiveness research (CER) and patient-reported outcomes. In due course, it is anticipated that industry, including pharma, will be able to mine these to identify optimal care pathways, to accelerate drug development, to rationalize services and to manage public health.

A more intimate example lies in the concept of Microsoft's HealthVault, which, at least at one time, had ambitions to join up all the commonly collected health-related information about a person (under the individual's control, so it was said) from the content of their shopping basket (courtesy of their supermarket loyalty card), to their daily exercise levels (through wearables or gym machines), to their relationship with their healthcare provider (numbers and kinds of visits, prescriptions, etc.). From here, it is easy to imagine one's mobile phone might soon be delivering messages about the inadvisability of chocolate, given one's BMI, just as one was reaching into the shelf in the store. But there are more benign examples: HealthHeritage has been established to help anyone construct and link up their health family tree, a social network focused on conditions that may have a genetic component and whose prevalence in one's family is worth knowing about.

A variant of the personal health record idea took the form of a scheme to support the global poor in countries where, in any case, pharmaceutical companies are already conducting lightly regulated clinical trials. The increasing adoption of EHRs in developing countries opens up an opportunity to conduct research based on data at the same time as supporting "the global poor" with payment for use of their data. The unpublished paper by Dzenowagis and Eyal [17] discusses the ethical, social and economic issues that arise: as well as the immediate issues of consent, confidentiality and privacy protection, the paper explores the form and distribution of benefits, who may be counted among "the global poor" and should be allowed to pay to access their data, whether such payments may constitute "undue inducement" and even an incentive to corruption. The authors acknowledge the issues to be addressed, but are favorably inclined none the less.

In the context of such wide-ranging uses and possible abuses of personal data, electronic informed consent and the conditions for regulatory compliance themselves become rather obscure.

Informed and Active Consent

There is a need for vigilance in what one allows one's data to be used for: this much has been clear in the quotidian world of internet and mobile apps, social media, cloud storage, unencrypted email, and much else. Informed choices in health-care are certainly advocated by all organizations, but are often observed peremptorily, as when the task is delegated to reception staff rather than a physician or other clinician who can address questions.

Defining an "adequate consent process", the Federal Drug Administration's *Information Sheet – A Guide to Informed Consent* [21] asserts:

> *Thus, rather than an endpoint, the consent document should be the basis for a meaningful exchange between the investigator and the subject.*

This expresses the true requirement as clearly and succinctly as it can be put. Informatics offers a genuine opportunity to create an informative, up to date, intelligent consent process that allows patients to choose, e.g., whether they wish to participate in any given study, and how, how they would wish their data to be used, and in what form (fully de-identified, or in limited data set form?), what information to receive back, whether lay or technical communications or both, and whether to offer a greater degree of involvement, such as undertaking a task in a project or co-presenting to other patients in the context of a self-management program.

Current discussion around e-Consent is relevant here, since the quality of information that can be provided through an electronic consent process is itself under scrutiny. An online presentation by Quorum Review IRB [5] highlights certain advantages of e-Consent: information and possible choices can be presented in a variety of ways, so that the user can choose one that chimes with their personal cognitive style: one may prefer to see a video, or a presentation with voiceover, an animation, text and tabulation, or a graphical explanation. While direct interaction with another person has many advantages, the explainer's performance will almost certainly vary as they get tired or bored, faced with a keen listener or someone who is also tired or bored. There is also the potential to refer to other places for further explanation, to materials presented by other patients, by other research teams with a different point of view, and so on. However, among the issues that would need to be addressed in such an ecumenical approach to knowledge sharing is the poor quality of much of the information available on the internet. [19] A research team could be the best guide to what to view. Universal standards, such as those put forward by the Health on the Net Foundation, [20] have gained limited traction, so perhaps a "think global, act local" attitude is most effective—where local means in one's area of expertise, in one's specialty, or even in one's community. We note in passing that the recent OHRP/FDA "guidance", meaning non-binding advice on good practice, asserts:

> *Although both OHRP and FDA affirm that the informed consent process begins with subject recruitment,4 recommendations on using electronic media and processes for subject recruitment are outside the scope of this guidance.*

This acknowledges a difficulty that would be faced by anyone seeking to provide a broad spectrum of commentary and yet avoid a free for all leading to confusion and misinformation.

Education, Citizen Science and Quantified Self

Meaningful stakeholder engagement with the research process requires, on one hand, an understanding of what research involves, how it achieves results, how the significance of such results may be assessed, and on the other, an appreciation of the professional, social and economic issues that constrain decisions on what research to pursue. Knowledge of these forces enables stakeholders to intervene at an appropriate point and on an appropriate scale, whether to influence the direction of a research program or to nudge a small project to include some aspect of particular interest. The Colorado Boot Camp Translation project offers many possibilities. [6] The value of the "boot camp" has certainly been demonstrated in many specific conditions, some of which have resulted in publication. [21]

Also in the realm of patient self-care, the remarkable work begun by Kate Lorig at Stanford, which has now spread internationally has led to a book of essays centering mainly on its British incarnation. [7] The essays in the book are drawn from different points of view, but focus on the notion of an "expert patient" and home in particularly on self-management education in the UK and especially on the NHS Expert Patient Programme.

Education of patients, now specifically with the goal of enabling them to become co-investigators, can derive many lessons from the Citizen Science movement [22] and its open door philosophy. The virtual organization Zooniverse [23] showcases many projects and lists an extraordinary number of publications, especially in Astronomy and Space Science, where data analysis is a prominent activity, but remarkably includes no fewer than 22 publications in the field of "Meta Studies" exploring, e.g., *Science Learning via Participation in Online Citizen Science*, *Playing with Science: Aspects of Gamification Found on the Online Citizen Science Project - Zooniverse*, and *Exploring the Motivations of Citizen Science Volunteers* (the full list at [22]). Indeed, all this is in addition to all the contributions of patients and other stakeholders who happen to have, from their own professional life, a set of necessary skills that a project may exploit—project management, statistics, education, and others.

The Quantified Self movement [23] represents a different aspect of Citizen Science. There are numerous remarkable examples of individuals putting their own "numbers" under scrutiny, as in Erica Forzani's exploration of her own pregnancy, unusual pattern of resting metabolic rate, apparent gestational diabetes and weight gain. [24]. While many women may monitor their pregnancy closely, there is a particularly scientific spirit of quantification in this and other exemplars of 'quantified self".

Conclusions

We have discussed, in turn, the values and the "learning" aspect of the Learning Health System. We have explored the perceived need for Comparative Effectiveness and Outcomes Research to accelerate the translational process from scientific evidence to practice. We have proposed forms of Health Record Banking that would be supportive of research. We have touched on the "commodification" of patient data and the possibility of patients taking control. We considered informed consent and electronic consent (e-Consent) policies and their potential to keep an open line between patient, stakeholder-researcher, provider, and research team. Finally we touched on self-education and research for potential stakeholder-scientists

through such approaches as expert patient programs, boot camps, and the Quantified Self movement.

We have demonstrated that the elements are at least available, either because they already exist and can be used, or can be implemented if the right regulatory framework were in place. Integration of these would be sufficient to realize the vision of stakeholder scientists, participating in the formulation of research questions, being subjects in studies, providing and sharing their own data, and providing unique insights into chronic conditions for researchers and fellow patients alike.

Acknowledgements and Disclaimer

The author wishes to acknowledge colleagues in the AMIA ELSI and CRI Working Groups. All opinions are solely those of the author.

References

[1] *What is a Learning Health System?* Web page at http://www.learninghealth.org (accessed 18/12/16)

[2] *Patient-Centered Outcomes Research Institute.* http://www.pcori.org (accessed 18/12/16)

[3] Amnon Shabo (Shvo). It's Time for Health Record Banking! *Methods of Information in Medicine* February 2014 [doi: 10.3414/ME13-02-0048]

[4] Use of Electronic Informed Consent: Questions and Answers. Guidance for Institutional Review Boards, Investigators, and Sponsors. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm436811.pdf Jill Wechsler. Central vs. Local: Rethinking IRBs: Regulators and sponsors encourage alternative review models to fit a growing research enterprise. Applied Clinical Trials (Feb 01, 2007) http://www.appliedclinicaltrialsonline.com/central-vs-local-rethinking-irbs

[5] QUORUM Review IRB Using Electronic Consent and Technologies in Research: Regulatory and IRB Considerations http://www.quorumreview.com/wp-content/uploads/2011/12/Electronic-Consent-webinar_11.30.11.pdf

[6] JM Westfall, et al. Reinventing the Wheel of Medical Evidence: How the Boot Camp Translation Process is Making Progress. *Health Affairs* 35, no.4 (2016):613-618

[7] F. Roy Jones (Ed.) Self-Management Courses. Oxford University Press, 2010

[8] Achieving a Nationwide Learning Health System. C.P. Friedman, A.K. Wong and D. Blumenthal. *Science Translational Medicine* **2** (57), 57cm29. (2010) [doi: 10.1126/scitranslmed.3001456]

[9] *LHS Core Values* Web page at http://www.learninghealth.org/corevalues/ (accessed 18/12/16)

[10] Laura Forsythe *Evaluation Update* presentation (October 2016) http://www.pcori.org/sites/default/files/PCORI-Patient-Engagement-Advisory-Panel-Fall-2016-Meeting-Presentation-Slides-102116.pdf

[11] E A Balas and S A Boren - Managing Clinical Knowledge for Health Care Improvement. *Yearbook of Medical Informatics*, 2000.

[12] Hanney *et al*. How long does biomedical research take? Studying the time taken between biomedical and health research and its translation into products, policy, and practice. *Health Research Policy and Systems* 2015, 13:1

[13] *FDA's Sentinel Initiative* http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm

[14] Ellen Poleshuck. Using patient engagement in the design and rationale of a trial for women with depression in obstetrics and gynecology practices. *Contemporary Clinical Trials* 43 (2015) 83–92

[15] K Erwin *et al*. Engaging stakeholders to design a comparative effectiveness trial in children with uncontrolled asthma. *J. Comp. Eff. Res.* (2016) 5(1), 17–30

[16] T G Roberts, Jr. Preparing for Success With Comparative Effectiveness Research**.** *The Oncologist* 2013;18:655-657; S D Ramsey *et al*. Oncology Comparative Effectiveness Research: A Multistakeholder Perspective on Principles for Conduct and Reporting. *The Oncologist* 2013;18:760–767.

[17] David Creasey, *An independent 'Health Information Bank' could solve data security issues*, an interview with Dr. Bill Dodds. The British Journal of Healthcare Computing & Information Management, Vol. 14, No. 8, October 1997

[18] Joan H Dzenowagis and Nir Eyal, Another Way for Electronic Health Records Research to Benefit the Global Poor (unpublished manuscript, 2014)

[19] Anthony E. Solomonides and Tim Ken Mackey. Emerging Ethical Issues in Digital Health Information: ICANN, Health Information, and the Dot-Health Top-Level Domain. *Cambridge Quarterly of Healthcare Ethics* 24(3):311-322 (2015)

[20] Health on the Net Foundation. When the quality of health information matters: Health on the Net is the Quality Standard for Information You can Trust. http://www.hon.ch/Global/pdf/TrustworthyOct2006.pdf

[21] Richard E. Deichmann *et al*. Long-Term Effects of a Diabetes Boot Camp on Measures of Diabetic Care. The Ochsner Journal 15:13–18, 2015

[22] Citizen Science Philosophy https://www.citizensciencealliance.org/philosophy.html ; What is the Zooniverse? https://www.zooniverse.org/about

[23] Technori. The Beginner's Guide to Quantified Self (Plus, a List of the Best Personal Data Tools Out There) http://www.technori.com/2013/04/4281-the-beginners-guide-to-quantified-self-plus-a-list-of-the-best-personal-data-tools-out-there/

[24] Erica Forzani. Tracking Pregnancy and Baby Growth. http://quantifiedself.com/projects/1027

Address for correspondence

Anthony Solomonides
NorthShore University HealthSystem
Research Institute
1001 University Place
Evanston IL 60201
United States

tony.solomonides@gmail.com