Grid-enabled sentinel network for cancer surveillance

Paul de Vlieger, Jean-Yves Boire, Vincent Breton, Yannick Legre, Jérôme Revillard, D. Sarramia, L. Maigne

Grid-enabled sentinel network for cancer surveillance

Paul De Vlieger[1,2], Jean-Yves Boire[2], Vincent Breton[1], Yannick Legré[3], David Manset[3], Jérôme Revillard[3], David Sarramia[1] and Lydia Maigne[1]

June 12, 2009

[1]LPC Clermont-Ferrand, Blaise Pascal University, CNRS-IN2P3, 63177 Aubière Cedex, France
[2]ERIM, Faculty of Medicine, P.O. Box 38, 63001 Clermont-Ferrand Cedex, France
[3]Maat-G, 74070 Archamps, France

Abstract

Recent developments of grid services for secured distributed data management open new perspectives for disease surveillance. In this paper, we report on our initiative to develop a surveillance network for cancer in the Auvergne region. The network gathers cytopathology laboratories, structures in charge of cancer screening and institutes in charge of cancer epidemiology. Data stored in laboratories are queried through the grid for the purpose of second diagnosis and to produce statistical indicators. The paper describes the network goal and design and discusses specific issues related to patient identification and security.

Contents

1 Introduction

Cancer is becoming the first cause of mortality in developed countries. In recent years, the number of patients treated for cancer has been constantly growing while mortality has started to decrease, thanks to the progresses accomplished in the treatment of this disease and to the development of cancer screening programs [2]. These programs allow an early detection of the malignant tumours which improves significantly the medical prognosis.

In order to evaluate the public health policies, reliable statistical indicators are needed. In France, several structures have been set up to collect epidemiological data on cancer such as CRISAPs *(Centre de Regroupement Informatique et Statistique en Anatomie et cytologie Pathologiques)* which are like regional data warehouses collecting anonymous data from anatomical pathology laboratories or from the healthcare structures involved in cancer treatment. The extraction of data from laboratories encounters reluctance from the healthcare professionals because of cost and also because they lose some control over the data they have produced.

Several projects in Europe have studied or are currently exploring the grid added value for addressing cancer: the pioneer projects focused on breast cancer, particularly computer-aided diagnosis of mammograms (e-Diamond [7] and MammoGrid [5],[6] projects). These projects have produced most of the middleware bricks being used to build our cancer surveillance network: MDM (Medical Data Manager) [10] and Globus Medicus [11] are some of them; and more recently, the Pandora Gateway designed for the Health-e-Child project [4].

In this paper, we propose a very innovative approach to both cancer screening and epidemiology based on grid technology. We describe how a 'collaboration' grid federating the cytopathology laboratories together with the screening associations and the institutes in charge of cancer epidemiology would manage easily the patient data in a secure and reliable way.

2 Objectives of a grid-enabled surveillance network

Context

Most EU countries have launched a national program for breast cancer screening [2]. In France, breast cancer screening is achieved through inviting women above 50 to have mammograms every 3 years. When a woman is positively diagnosed with a risk of tumour, cancer structures are in charge of providing a second diagnosis on the mammograms and have to follow-up on the anatomical pathology (or cytopathological) data about the tumour which are stored by the laboratories. Presently, the patient data are faxed on request or carried physically by the patient to the associations where they are recorded again. This process is costly and errors prone as data have to be typed and reinterpreted twice.

The cytopathological data are also extremely important for epidemiological analysis. The INVS (Institut National de Veille Sanitaire = Sanitary Surveillance Institute), French equivalent to (E)CDC[1] for the (EU)USA, is in charge of publishing indicators about global health and particularly about cancer. To produce its indicators, INVS relies on regional cancer registries (CRISAPs) set up to collect relevant information to support statistical and epidemiological studies about cancer incidence, mortality, prevalence and screening.

However, regional cancer registries have several drawbacks:

1. In any healthcare system, physicians are responsible of patient information. The pathologists refuse to trust these systems as data is exported outside their databases, so patient identification

[1] (E)CDC: (European) Centre for Disease Prevention and Control

criterias cannot be attached to the file in order to disengage responsibility. In this way, neither disambiguation nor patient linkage is possible, so statistics are biased. Effectively, only medical information is carried out without patient identification. As a patient can undergo several biopsies in different laboratories, the information about his cancer would be present twice in the register without linkage.

2. Some pathologists refuses to export since this method requires losing control of the data produced in their own laboratories which is like the fruits of their labor.

3. The current data gathering system needs physicians to export manually data from their software, format it according to the Crisap specification, connect to the central repository and send their file. This method is costly in time without any compensation.

Now, their usage is being questioned as the global quality of these repositories decreased and less and less laboratories contribute to the registers.

Solution proposed

Our alternative is for the clients to query anatomical pathology laboratories databases directly on site. A collaborative data grid, federating the laboratories, (see **Figure 1**) would provide a secured framework enabling the screening associations to query databases and fill their local patient file. No action is required by physicians to push their data on the network. Thanks to the Grid Security Infrastructure (GSI) [8], the pathologists are able to define and modify the access rights of the users querying their data.

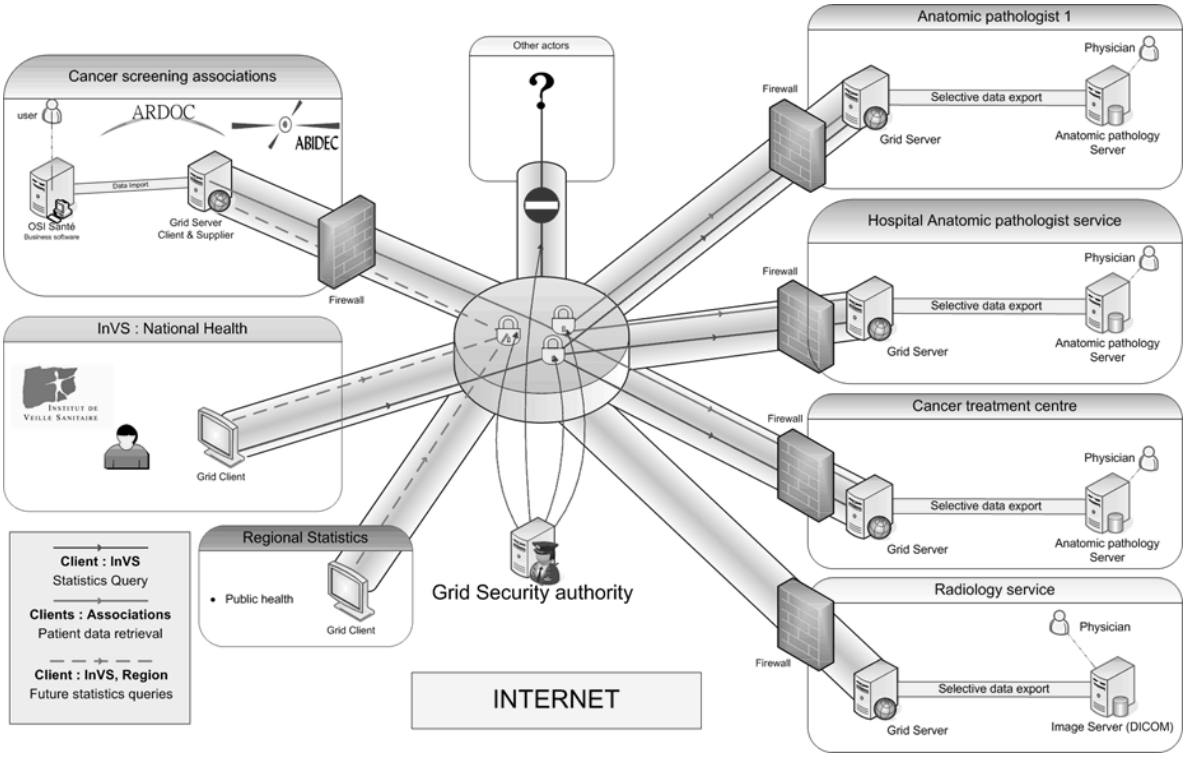

Figure 1 Cancer surveillance network using Grid technologies.

If a sentinel network is able to federate anatomical pathology databases, it can be used by the epidemiological services of the INVS and the regional epidemiological observatory to build epidemiologic studies.

3 Material and methods

List of requirements

As shown in **Figure 1**, the main medical data providers are the cytopathology laboratories. These different laboratories host different software systems and local databases for medical data management. Radiology services are additional data providers for mammograms in a step futher. The data requesters are the cancer screening associations and the different epidemiological structures in charge of producing statistics on specific cancers at regional and national level. Contrary to cancer screening associations who need to obtain the entire medical patient sheet, epidemiological structures need only anonymous medical data to produce statistics but with disambiguation to avoid double counting of cancer patients.

In a near future, the network should be able to grant access to medical images like mammograms to the cancer screening associations in order to ease the second diagnosis. The infrastructure design should offer a good flexibility to ease the entrance of new actors in the network.

The security infrastructure of the network needs to comply with French regulation on medical data transfer and exchange. The pathologists need to control the access rights to their own data. The network users must authenticate themselves using recognized accreditation tools like healthcare professionals cards. The individual certificates used have to be delivered by a certification authority (CA) recognized officially by the ministry of health and compatible with the grid infrastructure deployed.

Sentinel network infrastructure

The proposed dedicated grid architecture is built upon a central set of servers hosting security features and core grid services as illustrated in **Figure 2**:

- VOMS is an authorization manager, which implements a PKI-based authentication with certificates delivered by trusted authorities (CA) [3]. The usage of VOMS in this project is almost mandatory as VOMS is part of the gLite [13] middleware and guarantees a robust access control to the grid. Each user must own a certificate in order to log in. During the network development phase, a local certificate authority will be created, but once the network is operational, only official certificates for duly authorized healthcare professionals will be used to log in.

- The Pandora GateWay is a set of software designed as a Service Oriented Architecture (SOA). It was developed by the Maat-G society for the Health-e-Child project [4]. The Pandora Gateway is used to address medical data accessibility, exchange and processing while guaranteeing a high level of security for sensitive data. The main added value of the GateWay, compared to a classic SOA platform, is the high-level security. The GateWay Authentication Service is based on several security checkpoints required for log in. The access point is a Two Factor Authentication with user certificates and pin code followed by an authentication process using a VOMS Grid Proxy creation.

- The AMGA (ARDA Metadata Grid Application) [1] server, which provides a way to access and store metadata. Beside its high performance, the main advantages of AMGA are the full implementation of the grid security infrastructure (GSI) as well as the integration VOMS. When dealing with medical images, the use of metadata is mandatory and MDM (Medical Data Manager) allows bridging DICOM servers and the gLite middleware through AMGA [10]. AMGA is a very attractive software to fulfill the strong security requirements and access right management of medical data in a grid infrastructure.

- GridFtp server for data transfer [9].

- LFC server, for data management, included in the gLite middleware [13]

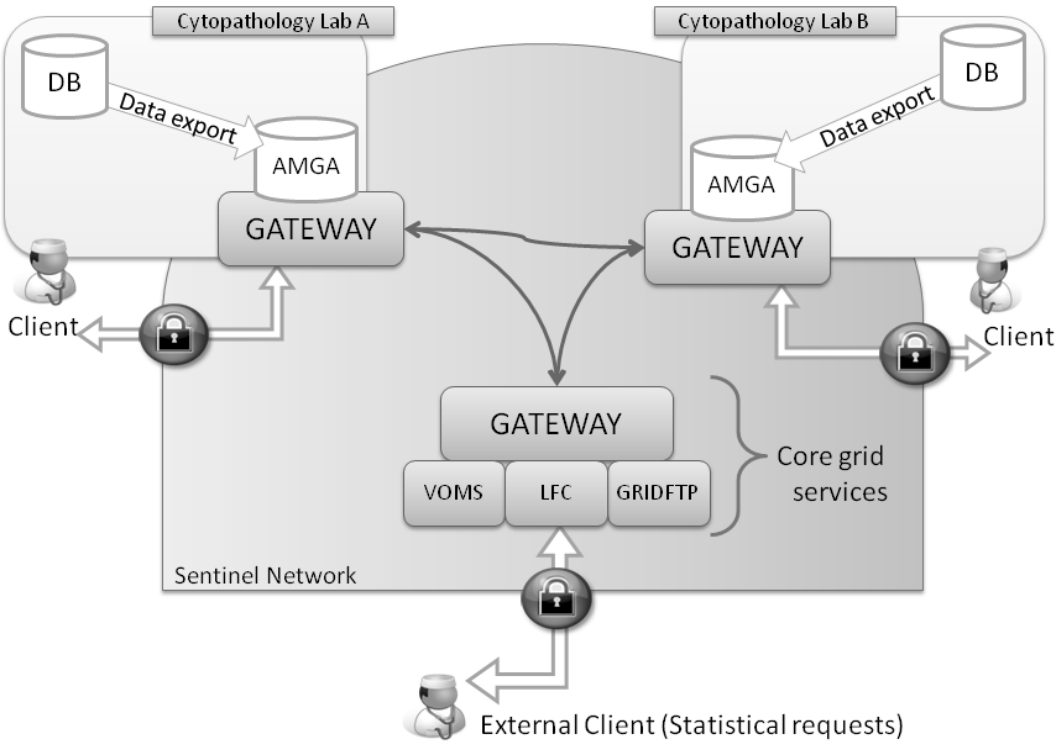

Figure 2 Service Oriented Architecture of the sentinel network.

If a client wants to use the system, he has to login through his local GateWay which authenticates himself on the sentinel network. When he launches a request, his local GateWay calls the different available AMGA servers on the network aggregate the responses and finally delivers the result to the client. In this way, no information is exchanged out of the different GateWays, guarantying a high level of security and easing data tracking. . Thanks to the Grid Security Infrastructure, authentication and credentials are available in the whole Grid, through GateWays and AMGA server.

4 Specific issues for prototype implementation

Data specification and retrieval system

For a better readability of cytopathological data, data sheets standardization is used to simplify the addition of medical records in the database without interfering with other data. Care must be taken at this step not to lose data coherence and the ownership of medical diagnosis. The customer part of the application hosts a grid server (under a firewall) to link patients' data in the network. For public health centers, a computer with an Internet access is just needed to launch epidemiologic requests.

Patient identification and data linkage

Patient identification is one of the major issues of modern healthcare systems. How can a healthcare system provide a way to identify surely his citizens while respecting their privacy?

As disambiguation and patient linkage is one of the central part of this project, patient identification is at the core of the data linkage problem. Currently, due to the lack of global identifier, each patient is linked to two different identification numbers (medical folder numbers) which are used inside each medical structure (see **Figure 3**).

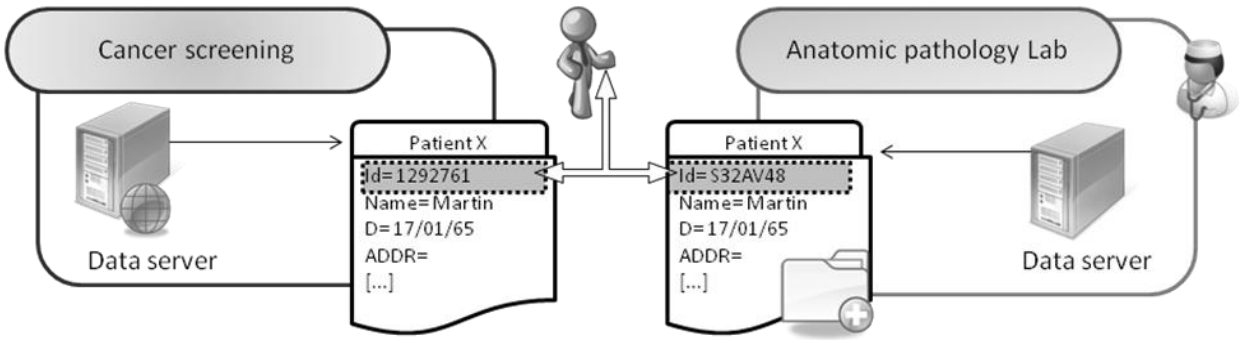

Figure 3 Patient identification problem.

Our solution requires an additional identifier for the sentinel network. This identifier consists in a random number generated (uuid type as defined in the RFC 4122 [12]) for each patient. This identifier would be created only for data linkage and would always be encrypted using different keys in each database to protect patient privacy.

When a data provider downloads some new data from his local data server to the local grid server, the Pandora Gateway is in charge of searching the patient in all the local databases respect to information on the patient. It will produce a unique identification number corresponding to the medical data if two identifiers are correlated to the same patient. With this procedure, we build a consistent network with unambiguous usage as we provide statistical requests free of doubles.

Security issues: strategy

One important step in the analysis is to clarify all security aspects in the sentinel network. The usage of SSL (Secure Socket Layer) and GSI will create a trusted network where data exchange and authorization procedures will fit the security requirements. However, and in order to protect patient privacy, the network access has to be restricted to authorized physicians or related staff only. The usage of Electronic Health Cards seems to be the best solution to settle this issue. These cards exist in France[2] and also in other European countries thanks to the EU framework[3]. Basically, these cards are smartcards containing certificates on a chip and supports strong authentication, electronic signature and data encryption. As bundled certificates are X509 formatted, they are intrinsically compatible with authentication on a gLite-powered Grid. These cards are also a response for data holders (i.e. pathologists) to accept the sharing of their data. As the sentinel network will be reachable only by physicians, the different queries will be launched under the responsibility of the physician who launches the query. The pathologists are now free to make available safely their data without any risk in case of wrong use.

Validation

Once the prototype operational and upstream of the real exploitation phase, the following step will consist in a validation of epidemiologic queries in order to certify the cohenrency and consistancy of the results obtained. The comparision will be possible with the previous conclusion of epidemiologic surveys locally obtained with manual methods by the regional public health service.

5 Conclusion

The article describes the goals and design of a surveillance network for breast cancer in the Auvergne region. The network will allow federating, in a fully secured way, cytopathology databases with cancer associations. The cancer network will be used to improve cancer screening programs and to produce reliable and theorically exhaustive cancer epidemiological indicators. Implementation of the sentinel network has started. The aim is to deploy a first prototype by the summer 2009 between one of the 2 cancer screening associations in Auvergne and a cytopathology laboratory. The beta version of the network will be operational by the end of 2009.

The integration of additional laboratories as well as the development of query interfaces for epidemiological structures will be addressed in 2010. Extension of the network to other cancer types and medical images sharing are foreseen within the framework of a new project to be funded by French Research Funding Agency.

Acknowlegements

The authors wish to acknowledge numerous discussions with P.Bouchet, A.Gaillot, L.Gerbaud, M-A.Grondin, A.Lautier, P.Lonchambon and C.Mestre.

[2] www.gip-cps.fr
[3] www.hprocard.eu

The work described in this article was partly supported by grants from the European Commission (EGEE, Embrace), the French Ministry of Research (GWENDIA) and the regional authorities (Conseil Régional d'Auvergne, Conseil Général du Puy-de-Dôme, Conseil Général de l'Allier). The Enabling Grids for E-sciencE (EGEE) project is co-funded by the European Commission under contract INFSO-RI-031688. The EMBRACE project is co-funded by the European Commission under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092. Auvergrid is a project funded by the Conseil Regional d'Auvergne. The GWENDIA project is supported by the French ministry of Research.

Reference

[1] B. Koblitz et al, The AMGA Metadata Service, *Journal of Grid Computing* **6** (2008), 61-76.

[2] Cancer Screening in the European Union; *Report on the implementation of the Council Recommendation on cancer screening* (2007).

[3] R. Alfieri, R. Cecchini, et al, From gridmap-file to VOMS: managing authorization in a Grid environment, *Future Generation Computer Systems* **21** (4) (2005): 549-558.

[4] The Health-e-Child project: http://www.health-e-child.org/

[5] R Warren et al, A Prototype Distributed Mammographic Database for Europe, Clinical Radiology 62.11 pp 1044-51 (Elsevier) (2007)

[6] 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, Clinical Radiology Vol 62.11 pp 1052-60 (Elsevier) (2007)

[7] M. Brady et al, eDiamond: a grid-enabled federated database of annotated mammograms, *Grid Computing: Making the Global Infrastructure a Reality F Berman*, G Fox and T Hey (eds), Wiley, (2003).

[8] V. Welch et al, Security for Grid services, *High Performance Distributed Computing,* 12th IEEE International Symposium on (2003), 48-57.

[9] V. Allcock et al, *The Globus Striped gridFTP Framework and Server,* ACM/IEEE conference on Supercomputing (2005), 54-64.

[10] J. Montagnat et al, Bridging Clinical information systems and grid middleware: a Medical Data Manager, *Studies in health technologies and informatics* **120** (2006), 14-24.

[11] SG. Erberich et al, Globus MEDICUS - federation of DICOM medical imaging devices into healthcare Grids. *Studies in health technologies and informatics* ***126*** (2007), 269-78.

[12] P. Leach, M. Mealling, and R. Salz. A Universally Unique IDentifier (UUID) URN Namespace. IETF RFC 4122 (2005), http://www.ietf.org/rfc/rfc4122.txt

[13] gLite Middleware : http://glite.web.cern.ch/glite