Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Linked2Safety: A secure linked data medical information space for semantically-interconnecting EHRs advancing patients' safety in medical research

Athos Antoniades*, Christos Georgousopoulos†, Nikolaus Forgo‡, Aristos Aristodimou*, Federica Tozzi §, Panagiotis Hasapis †, Konstantinos Perakis ¶, Thanasis Bouras ¶, Dimitris Alexandrou, ¶ Eleni Kamateri‖, Eleni Panopoulou‖, Konstantinos Tarabanis‖ and Constantinos Pattichis*

*University of Cyprus, Nicosia, Cyprus Email: athos@cs.ucy.ac.cy

†INTRASOFT International, Brussels, Luxembourg, Email: christos.georgousopoulos@intrasoft-intl.com

‡Gottfried Wilhelm Leibniz Universitat Hannover, Hannover, Germany, Email: nikolaus.forgo@iri.uni-hannover.de

§School of Medicine, University of North Carolina, Chapel Hill, NC, USA. Email: federica.tozzi1@tin.it

¶Ubitech Ltd., Athens, Greece, Email: kperakis@ubitech.eu

‖The Centre for Research and Technology Hellas, Thessaloniki, Greece, Email: epanopou@iti.gr

***Abstract*—Electronic Health Records (EHRs) contain an increasing wealth of medical information. They have the potential to help significantly in advancing medical research, as well as improve health policies, providing society with additional benefits. However, the European healthcare information space is fragmented due to the lack of legal and technical standards, cost effective platforms, and sustainable business models. The vision of Linked2Safety is to advance clinical practice and accelerate medical research, by providing pharmaceutical companies, healthcare professionals and patients with an innovative secure semantic interoperability framework facilitating the efficient and homogenized access to anonymised distributed EHRs in an aggregate form that enables merging multiple data sources into a single analyses. In this paper a first public introduction to the project is provided along with a clear definition of the problems, and proposed architecture. Three usage scenarios are used to demonstrate the potential impact of the outcomes of the project.**

***Index Terms*—Semantic Interoperability, Electronic Health Records, Personal Data Protection, Adverse Event Prediction, Genetic Analysis**

I. Introduction

In this paper the Linked2Safety project [1] is presented, a project funded under the FP7 scheme of the European Commission focused on some key challenges and potential solutions that have been identified through the work performed so far in the project. The vision of the project is to advance clinical practice and accelerate medical research, by providing pharmaceutical companies, healthcare professionals and patients with an innovative semantic interoperability framework facilitating the efficient and homogenized access to distributed Electronic Health Records (EHRs).

EHRs contain an increasing wealth of medical information and they have the potential to contribute significantly and advance medical research, as well as improve health policies, providing society with additional benefits [2]–[5]. However, the European healthcare information space is fragmented due to the lack of legal and technical standards, cost effective platforms, and sustainable business models. The key factors that define the work done in this and any other project that attempts to merge or share medical data between multiple sources are legal and ethical issues, patient privacy protection and statistical power.

Linked2Safety, is attempting to develop a platform for analysing EHRs from multiple distributed institutions, while strictly adhering to the legal and ethical requirements as defined by each data provider at EU level. The primarily objective of analysing EHRs from multiple institutions, arises from the need to increase the total number of subjects that are included in each analyses, thus increasing the statistical power of detecting true positive effects. This may in turn enable the identification of key biomarkers, which may lead to new drugs, the identification of adverse events that may be associated with a specific drug or family of drugs, and the reduction of costs associated with the setting up of clinical trials [6]–[9]. The costs of clinical trials can be reduced, since it will make the process of identifying sites and the number of subjects to be included in a clinical trial from each site easier.

The structure of the paper is as follows. Section II covers the legal and ethical issues of the project, whereas Section III and Section IV present the Linked2Safety platform and architecture respectively. Section V presents the three showcases that will be used for demonstrating the use of the platform when it is completed. Finally in SectionVI the concluding remarks are provided.

II. Legal and Ethical issues

Within the Linked2Safety project, patients' personal health data will be processed, which prerequisites the study of the corresponding national and European legislative framework, and the establishment of specific legal and ethical requirements for the platform that will be developed and operating on these data.

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

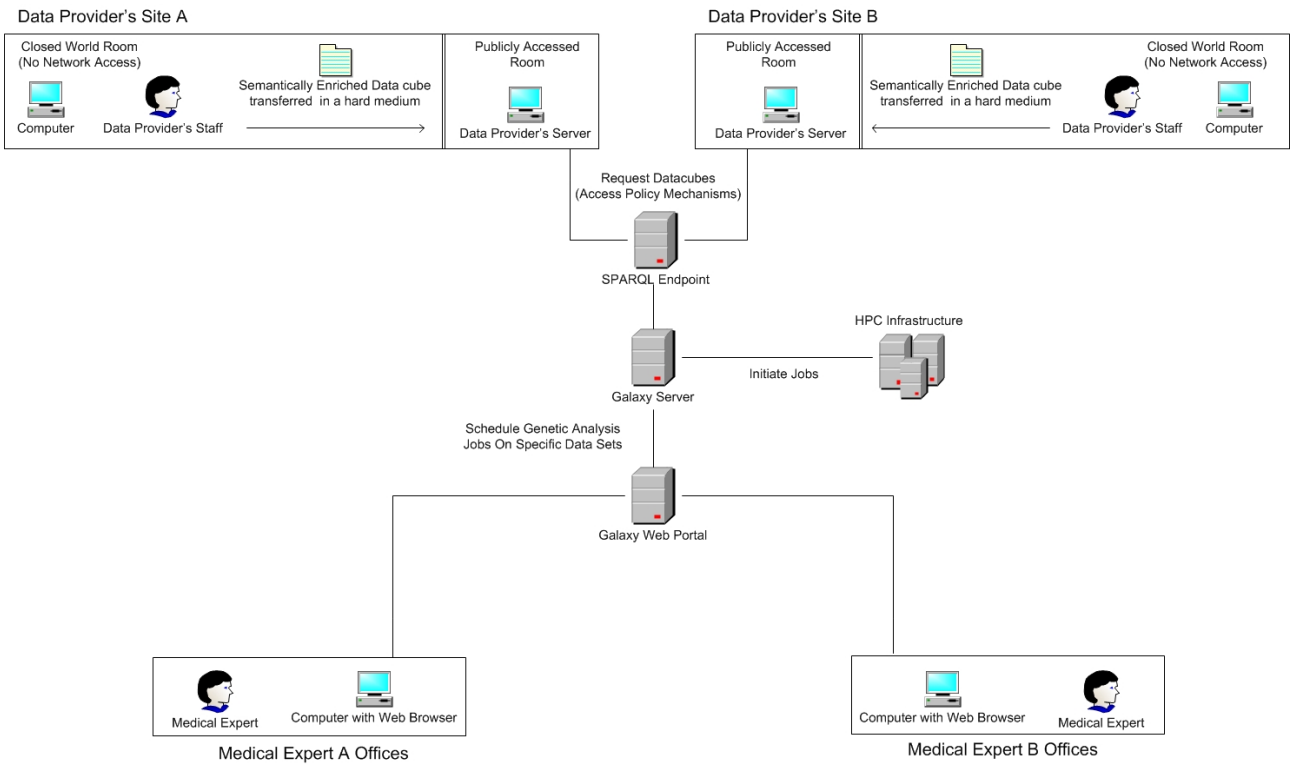

Fig. 1. The Linked2Safety platform. The data are initially anonymised and semantically enriched before making them available for access by a SPARQL endpoint. The medical expert connects to a Galaxy web portal, which schedules any analysis that the user wants to perform. The analysis is then performed on an HPC infrastructure and the results are returned to the user through the Galaxy web portal.

According to Art.17 para. 1 of the European Data Protection Directive [10], the data controller as well as the data processor must implement appropriate technical and organizational measures to protect the data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access and against all other unlawful forms of processing. At the same time, the data subject has several rights that must be respected, such as the right to be informed, the right of access, the right of rectification, erasure or blocking and the right to object.

From the ethical perspective, the processing of health data at EU level requires the consent of the data subject [11], though in certain cases and under specific conditions this might not be needed. For instance, the national laws of the clinical pilot partners of Linked2Safety in Greece [12], Cyprus [13] and Switzerland [14] declare that the use and processing of health data without the consent of the data subject is possible, if and only if, it is for research purposes, the data are properly anonymised, and analysis is done in aggregated level.

III. Linked2Safety Platform

The Linked2Safety architecture needs to address legal and ethical issues while enabling the analysis of electronic health data across multiple data sites with different owners and with potentially diverse legal and ethical requirements.

With those two facets of the problem, a novel idea for addressing both legal and ethical problems (at a commercial and not just a research level), as well as the need to merge the data from multiple sites simultaneously is introduced. All processing and analysis on medical data should be done on aggregated data from each site. This is nothing new, since aggregated results from the analysis of electronic health records are published in the majority of medical research publications that involve medical studies. In Linked2Safety however a system is needed that will enable future analysis of the aggregated data using methodologies or ways of analysis that may have not been considered at the time of aggregating the data. The proposed method that enables this is the use of data cubes [15]. This method aggregates values across many dimensions together, resulting in multidimensional data cubes. Getting a contingency table from any subset of the variables within the data cube can easily be achieved by adding the values across the excluded dimensions.

For tackling the issue of accessing electronic health records that contain potentially identifiable personal medical data, the concept of a "closed-world" room is introduced (a room located within a data provider's premises, featuring the required hardware infrastructure to process EHRs isolated from any kind of network connections).

The physical access to this machinery within the room is allowed only to specific personnel of the corresponding data provider and it is off line to the outside world. The data provider's staff will execute a program on the computers located in the closed room, that will aggregate the data gener-

ating the data cubes. The program will offer the option to the data provider to limit the way that the data will be aggregated so that any legal and ethical issues that may relate to the type of analyses that may be performed on the data can be addressed. Quality control will also be performed on the data by the program, so that the likelihood of reverse engineering of the data of a single subject or a group of subjects is minimized. As illustrated in Fig. 1, only the aggregated data (data cubes) will be physically carried outside the closed-world room to a server that will be accessible by the rest of the Linked2Safety infrastructure. The data cubes will be in RDF format and will be accessed through a SPARQL endpoint, whereas all of the tools for analysing the aggregated data will be on a dedicated Galaxy server [16], [17] for the needs of Linked2Safety. The analysis of the data will be available only to Linked2Safety users through a Galaxy web portal.

The operational procedures for the creation and semantic enrichment of the data cubes are as follows:

- The data provider's staff after reviewing the legal and ethical requirements for their data, make a decision on what data to include in Linked2Safety and what parameters they need to define for the creation of the aggregated data.
- A member of the staff of the data provider enters the "closed-world" room, where the data are maintained and performs the aggregation of the data, which will create the data cubes. This step includes the quality assurance and filtering of the data, based on the predefined settings of the previous step.
- The produced data cubes are then stored in the RDF format and are verified that they do not contain any personal medical records.
- The final data cubes are transferred to a server that is accessible by the Linked2Safety platform, outside the "closed-world" room.
- The data cubes are semantically enriched and made available to the rest of the Linked2Safety platform.

The above steps will be carried out by all sites that maintain clinical data, so as to include in the Linked2Safety platform aggregated data from multiple independent studies. In cases where the definitions of the variables overlap among sites, they will be recorded in a way that will enable them to be analysed and merged so as to increase the statistical power of detecting true positive results. These data will then be available to researchers with access to the Linked2Safety platform for analysing them.

IV. Linked2Safety Architecture

Fig. 2 portraits the Linked2Safety's architecture. As can be seen, the platform consists of four main spaces, which are described below.

A. Data Cube Generation Space

The *data cube generation space* provides the tools for transforming the proprietary EHR and Electronic Data Capture (EDC) data to the data cube structure. Specifically, the locally stored EHR and EDC data are aligned, mapped and transformed to a reference EHR schema. Then, this commonly referred EHR data are used to create the data cubes.

B. Interoperable EHR Data Space

The *interoperable EHR data space* provides a tool set to make the transition of data cubes from the "closed-world" of each clinical data provider to an open data environment accessible to all partners based on policies that enforce strong data security, privacy and anonymity. Thus, its responsibility is to transform the data cube information to a common referenced data cube format by means of a semantic EHR model, named common data cube reference EHR ontology. Moreover, the interoperable EHR data space provides the mechanisms for the semantic enrichment of the standardised data cubes with the use of appropriate, globally available healthcare and medical taxonomies and ontologies, enabling the delivery of machine interpretable information regarding their structure and content.

C. Linked Medical Data Space

The *linked medical data space* implements a secure knowledge base of semantically interconnected data cube related information resources. It also provides the mechanisms and tools required for publishing and interlinking the common referenced data cubes from different medical data providers. Access to this data is governed by adaptable access policies and mechanisms. In this way, the clinical research community is going to have homogenised access to the available anonymised patient related information needed, to perform complex data mining operations.

D. Genetic Data Analysis Space

The *genetic data analysis space* provides a scalable infrastructure for medical data mining, empowered by a set of algorithms and models. These methods are going to be applied in the semantically-interlinked data cubes containing anonymised patient's health records in order to analyse the associations among the genetic, environmental and phenotypical data related to identified and reported Adverse Events (AEs). Thus, clinical researchers and healthcare professionals are provided with an advanced genetic analysis statistical and data mining toolset, focusing on advancing patients' safety through the analysis of bio-markers associated to identified AEs and the proactive exclusion of specific patients' profiles from the wide patients' selection process.

V. Usage Scenarios

Although the Linked2Safety platform will enable many diverse types of analysis that may help in multiple facets of research in medicine, three have been identified and will act as showcases to demonstrate the use of the platform when it's completed.

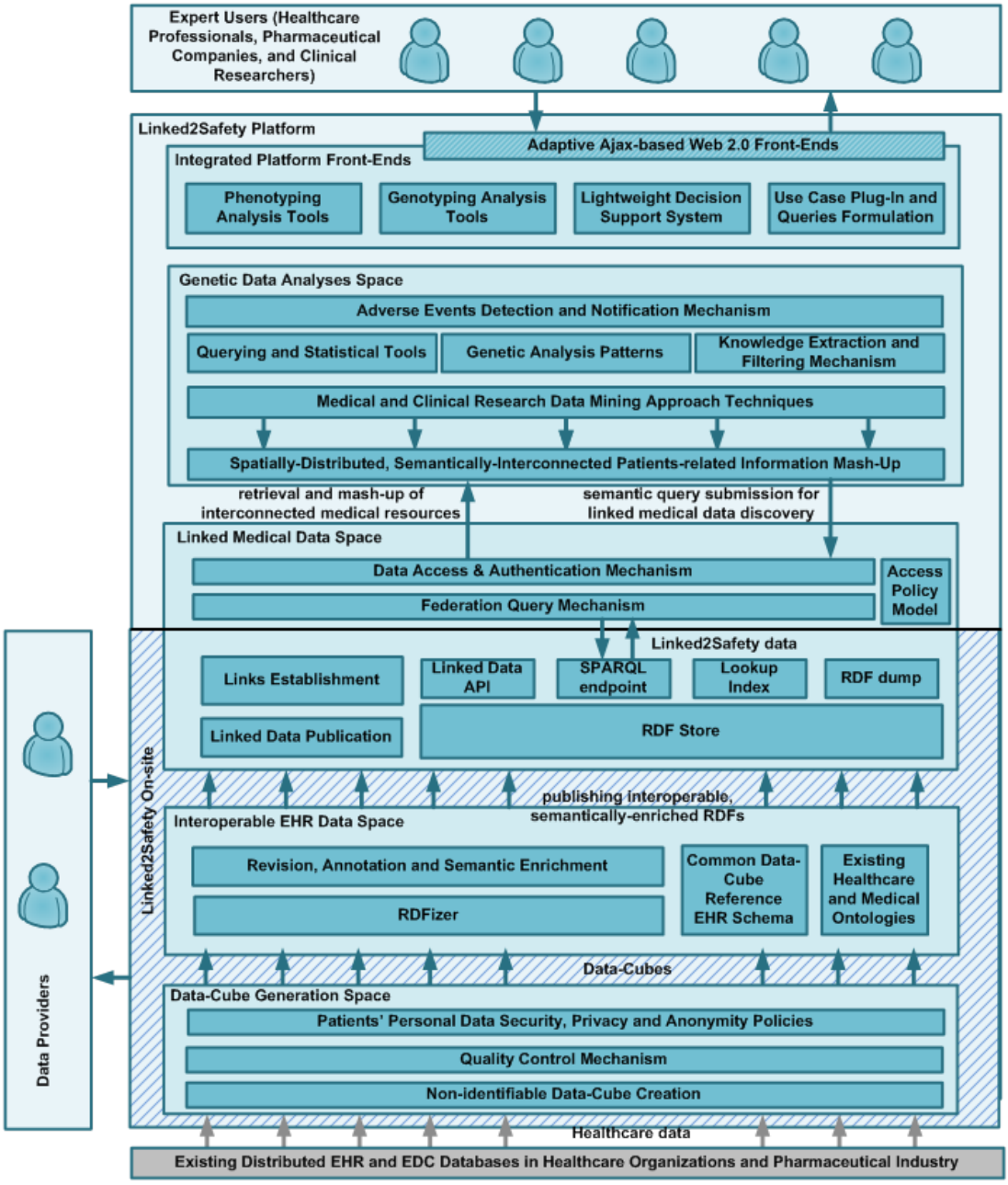

Fig. 2. The Linked2Safety architecture consists of four spaces. The *data cube generation space* creates the data cubes for anonymising the patients personal data. The *interoperable EHR data space* provides the tools for transforming the data cube information to a common referenced data cube format. The *linked medical data space* provides a secure knowledge base of semantically interconnected data cube related information resources. The fourth space, *genetic analysis space*, provides the tools for data mining and statistical analysis of the medical data.

A. Subject Selection for a Phase III Clinical Trial

Recruitment of investigators, clinical sites, and study participants is a constant concern during the course of any clinical trial. Recruiting participants for clinical trials is a time-consuming, intensive process which constitutes the critical first step for a study's ultimate success. Once investigators and clinical sites are selected, participants recruitment and retention is performed. Clinical researchers then determine whether clinical studies adhere to time-lines. Depending upon the size of the trial, the condition or disease being investigated, and the participant population available, subject recruitment often proves to be a time-consuming step for many clinical trials.

Clinical studies are no longer exclusively centralised in university and research hospitals. Many clinical trials have been relocated to private practice or small clinical research centres, often spread across the continent in pan-European trials. Investigators and academic research organisations are also implementing global clinical trials and mining databases

in search of physicians who treat patients eligible for trials. Sponsors and clinical research organisations usually rely on a network of known investigators and sites, tending to go to the sites that they have been working with in the past. To succeed in the identification of superior sites in terms of quality and subject recruitment, they maintain active investigator databases for Europe and the United States. Information typically included in such databases is: contact details, area of research interest, investigator's experience with clinical trials (phase of trials, number of patients enrolled at site per trials, classes of drug/devices tested), site infrastructure and capacity (equipment, internal logistics, patient database), etc.

Moreover, another way to identify sites and investigators is the number of publications released in the specific field. Placing a study in the best recruiting centre, where the appropriate subject population is available, and in a competition-free environment are also critical factors for a successful clinical trial.

The objective of this scenario is to demonstrate how sponsors could identify recruiting sites for participants into clinical trials in the most time- and cost-efficient way. For them, any delay in approval for a successful drug can potentially cost millions in sales, in addition to preventing promising novel therapies from reaching future patients.

B. Phase IV Post Marketing Surveillance trial

A Phase IV study is a clinical trial, a quasi-experimental study, or an observational study to gather specific information about an approved drug, a biological product, a device, or a procedure. Post-approval research is typically initiated to better understand product use in real-world situations, to obtain evidence for higher reimbursement or submission for expanded labelling, to fulfil a specific requirement of regulatory authorities, or to monitor safety of a drug or device in a larger, non-clinical trial setting.

The need for post-approval surveillance studies stems from inherent limitations in the clinical trial process used for regulatory approval. While studies of safety and efficacy in optimal populations are required to bring a drug or a device to market, it is only through post-approval research and studies of patient outcomes from product use in real-world settings that strong evidence of the effectiveness and safety of a new product emerges. Studies in large populations, with real-world dosing, longer duration of exposure, long-term follow-up, and comparison data based on current physicians' practices, are the most informative approaches to monitoring device and drug safety.

Recent regulatory initiatives have confirmed the significant value and appeal of information generated from registries that are well-designed and appropriately conducted, analysed, and reported [18]. Observational studies constitute a subcategory of post marketing studies. Observational studies are studies in which the researcher observes and does not alter the participant's experience in any way. In an observational study, treatment decisions, visit schedules, and any tests/measurements are generally left to the discretion of the provider. Observational studies can provide the health community with invaluable data on safety and effectiveness of a product and/or information about the natural history of a disease under standard care practices. These studies provide real-world data that can benefit patients, healthcare providers, pharmaceutical companies, sponsors, and regulatory agencies.

The objective of this scenario is to provide an analytical framework to analyse existing databases' content with information including (but not limited to): demographics, medication used, AEs and medical history. The aim is to detect safety signals as soon as possible, and identify association with factors that may act as causative or predisposing to the AE.

C. Identification of Relations between Molecular Fragments and Specific Adverse Side Effect Categories (Chemoinformatics)

This scenario focuses on the identification of structural features in drugs that may be related to AEs in population sub-groups. The inspiration stems from the application of similar chemical structure-based techniques for the prediction of biological properties including potency, solubility, toxicity, etc. in the drug discovery field.

To succeed in this, we need detailed information on patients' records including drugs administered and AEs. The availability of this information will allow the application of chemoinformatics methods that suggest potential chemical structure-adverse effect relationships. A subset of such potential relationships will be subjected to further analysis at a molecular level to elucidate the mechanism of action of the associated molecular sub-structure and the characteristics of the sub-population affected. The success of this process relies heavily on the quantity and quality of information on patients and their characteristics, as well as, on AEs.

The objective of this scenario is to demonstrate the usefulness of Linked2Safety framework to search and identify association between molecular features (e.g. chemical structure of drugs), and the occurrence of AEs.

VI. Conclusion

The Linked2Safety platform aims at exploring today's technology into ways that it can be applied to merged electronic health data from multiple distributed sites, for analyses that will ultimately have a positive impact on patients safety and outcomes. In order to do so, a maze of legal and ethical issues need to be addressed that are not necessarily consistent across different sites, especially when the sites are in different countries. The use of data cubes as proposed by Linked2Safety, allows the analysis and merging of data from multiple distributed sites, while conforming to all legal and ethical issues relevant to each dataset.

The usability of the Linked2Safety platform will be demonstrated through three diverse usage scenaria that are outlined in this paper. These were designed to show the diverse impact that Linked2Safety aims to make across multiple sub-domains in medicine, from adverse event detection, to subject selection for clinical trials, chemoinformatics and genetic analyses.

ACKNOWLEDGMENT

The research leading to these results was conducted as part of the project *A next-generation secure linked data medical information space for semantically-interconnecting electronic health records and clinical trials systems advancing patients safety in clinical research (Linked2Safety)* that received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No 288328.

REFERENCES

[1] "A next-generation secure linked data medical information space for semanticallly-interconnecting electronic health records and clinical trials systems advancing patients safety in clinical research," http://www.linked2safety-project.eu/. [Online]. Available: http://www.linked2safety-project.eu/

[2] L. Kun, R. Beuscart, G. Coatrieux, and C. Quantin, "Improving outcomes with interoperable EHRs and secure global health information infrastructure," in *29th Annu. Int. Conf. of the IEEE Eng. in Med. Biol. Soc., 2007. EMBS 2007*, Aug. 2007, pp. 6158 –6159.

[3] "Inovative medicine initiatives." [Online]. Available: http://www.altaweb.it/documents/imi-call-topics-2009_en.pdf

[4] L. He, X. Li, and P. Huang, "Sharing of ehr clinical test results," *Zidonghuayu Yibiao/ Automation & Instrumentation*, vol. 25, no. 5, pp. 18–21, 2010.

[5] B. A. Stewart, S. Fernandes, E. Rodriguez-Huertas, and M. Landzberg, "A preliminary look at duplicate testing associated with lack of electronic health record interoperability for transferred patients," *J. of the Amer. Medical Informatics Assoc.*, vol. 17, no. 3, pp. 341–344, May 2010.

[6] O. Kilic and A. Dogac, "Achieving clinical statement interoperability using r-MIM and archetype-based semantic transformations," *IEEE Trans. on Inform. Technology in Biomedicine*, vol. 13, no. 4, pp. 467 –477, Jul. 2009.

[7] A. Stell, R. Sinnott, and J. Jiang, "A federated data collection application for the prediction of adverse hypotensive events," in *9th Int. Conf. on Inform. Technology and Applications in Biomedicine, 2009. ITAB 2009*, Nov. 2009, pp. 1 –4.

[8] Y. Xiao, T. Pham, X. Jia, X. Zhou, and H. Yan, "Correlation-based cluster-space transform for major adverse cardiac event prediction," in *2010 IEEE Int. Conf. on Sys. Man and Cybern. (SMC)*, Oct. 2010, pp. 2003–2007.

[9] N. Ramakrishnan, D. Hanauer, and B. Keller, "Mining electronic health records," *Comput.*, vol. 43, no. 10, p. 7781, 2010.

[10] E. 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," *Official J. of the European Communities*, vol. 281, pp. 31–50, 1995.

[11] R. Faden, T. Beauchamp, and N. King, *A history and theory of informed consent*. Oxford University Press, USA, 1986.

[12] "Privireal: Data protection - greece," Aug 2012. [Online]. Available: http://www.privireal.org/content/dp/greece.php

[13] "Office of the commissioner for personal data protection - home page," Aug 2012. [Online]. Available: http://www.dataprotection.gov.cy/dataprotection/dataprotection.nsf/d1813d5911e138bdc2256cbd00313d1c/f8e24ef90a27f34f c2256eb4002854e7

[14] "Federal act on data protection," Aug 2012. [Online]. Available: http://www.vud.ch/generaldocs/vud_revdsg/235.1_FADP_en.pdf

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, Jun. 2011.

[16] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: a web-based genome analysis tool for experimentalists," *Current Protocols in Molecular Biology*, vol. 19, no. 19.10, pp. 11–19, 2010.

[17] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, p. R86, Aug. 2010.

[18] S. Bateman, "Riskmaps: A route to drug safety," *Good Clinical Practice J.*, vol. 12, no. 9, p. 15, 2005.