Journal of Biomedical Informatics 62 (2016) 32–47

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The INTEGRATE project: Delivering solutions for efficient multi-centric clinical research and trials

Haridimos Kondylakis [a,*], Brecht Claerhout [b], Mehta Keyur [c], Lefteris Koumakis [a], Jasper van Leeuwen [f], Kostas Marias [a], David Perez-Rey [d], Kristof De Schepper [b], Manolis Tsiknakis [a,e], Anca Bucur [f]

[a] Computational BioMedicine Laboratory, FORTH-ICS, N. Plastira 100, Heraklion, Greece
[b] Custodix NV, Kortrijksesteenweg 214b3, Sint-Martens-Latem, Belgium
[c] German Breast Group, GBG Forschungs GmbH, Geschaeftsfuehrer: Prof. Dr. med. Gunter von Minckwitz, Handelsregister: Amtsgericht Offenbach, HRB 40477 Sitz der Gesellschaft ist Neu-Isenburg, Germany
[d] Biomedical Informatics Group, DLSIIS & DIA, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo S/N, 28660 Boadilla del Monte, Madrid, Spain
[e] Department of Informatics Engineering, Technological Educational Institute of Crete, Estavromenos 71004, Hearklion, Crete, Greece
[f] PHILIPS Research Europe, High Tech Campus 34, Eindhoven, Netherlands

ARTICLE INFO

Article history:
Received 14 December 2015
Revised 5 May 2016
Accepted 17 May 2016
Available online 17 May 2016

Keywords:
Innovative biomedical applications
Semantic integration
Post-genomic clinical trials
Collaborative environment
Cohort selection
Predictive models
Quantitative analysis

ABSTRACT

The objective of the INTEGRATE project (http://www.fp7-integrate.eu/) that has recently concluded successfully was the development of innovative biomedical applications focused on streamlining the execution of clinical research, on enabling multidisciplinary collaboration, on management and large-scale sharing of multi-level heterogeneous datasets, and on the development of new methodologies and of predictive multi-scale models in cancer.

In this paper, we present the way the INTEGRATE consortium has approached important challenges such as the integration of multi-scale biomedical data in the context of post-genomic clinical trials, the development of predictive models and the implementation of tools to facilitate the efficient execution of postgenomic multi-centric clinical trials in breast cancer.

Furthermore, we provide a number of key "lessons learned" during the process and give directions for further future research and development.

1. Introduction

There is a strong need in medical research, especially in complex, heterogeneous diseases such as cancer, to achieve an all-comprising harmonization of efforts across disciplines, organizations and industries. Wide-scale utilization of clinical data for research is often hindered by fragmentation of methodologies [1], by limited interoperability and adoption of standards [2], and by lack of efficient and easy to use solutions to support collaboration [3]. This results in high financial burden for health consumers and slow transfer of new knowledge and technology into care.

The INTEGRATE project focused on overcoming several key obstacles in oncology research by delivering an environment and solutions to facilitate efficient clinical research and clinical trials in particular in a multi-centric setting, building on experiences of prior efforts and research results in the domain [4]. The project was guided during the development by scenarios and requirements stemming from breast cancer research, provided by the clinical partners in the project. We developed innovative infrastructures to enable data and knowledge sharing, to foster large-scale collaboration, to bring together heterogeneous multi-scale biomedical data generated through standard and novel technologies within post-genomic clinical trials and seamlessly link to

Abbreviations: AC, Access Control; AP, Analysis Platform; BC, Breast Cancer; BRIDG, Biomedical Research Integrated Domain Group; CDM, Common Data Model; CDP, Centre for Data Protection; CIM, Common Information Model; CNS, Central Nervous System; CRP, Central Review for Pathology; eCRF, electronic Clinical Report Form; EHR, Electronic Health Record; ER, Estrogen Receptor; FUH, Frankfurt University Hospital; GBG, German Breast Group; HGNC, database of human gene names; HL7, health level 7; INTEGRATE, project acronym for the project with the full title "Driving Excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures"; IJB, Institute Jules Bordet; ISO, International Organization for Standardization; IT, Information Technology; IdPs, Identity Providers; LOINC, Logical Observation Identifiers Names and Codes; MEDDra, Medical Dictionary for Regulatory Activities; PDP, Policy Decision Point; RIM, Reference Information Model; SIL, Semantic Integration Layer; SUS, System Usability Scale; SSO, Single Sign On; STS, Security Token Service; TTP, Trusted Third Party; XACML, eXtensible Access Control Markup Language.

* Corresponding author.
E-mail address: kondylak@ics.forth.gr (H. Kondylakis).

http://dx.doi.org/10.1016/j.jbi.2016.05.006
1532-0464/

existing research and clinical infrastructures (clinical trial systems, eCRFs, hospital EHRs) to enable a range of innovative applications with high added-value. Essential to the INTEGRATE environment, our semantic interoperability solution leverages existing and widely-adopted standards. The semantic interoperability layer has been implemented based on the HL7 v3 standard and on existing medical ontologies/terminologies: SNOMED-CT, MedDRA and LOINC. The BRIDG standard has been used to represent the clinical trial information in our environment. To be able to reuse previous efforts in data sharing, modeling and knowledge generation and to access relevant external sources of data and knowledge it is beneficial to adhere whenever possible to existing standards. The use of standards will also support the adoption of our solutions.

To facilitate efficient execution of post-genomic multi-centric clinical trials, we focused on the study of issues related to the automatic identification of eligible patients for inclusion into a clinical study, which, based on available scientific evidence [5] remains a key problem in contemporary clinical trials. We proposed a tool, named DECIMA, to support recruitment through the automatic evaluation of the eligibility of patients for trials based on matching the characteristics of the patient population required by the trial to the patient data available for instance in the hospital EHR.

We also developed a tool, named Central Review for Pathology (CRP), focusing on efficient central review of pathology images in order to enhance the collaboration among expert pathologists and enable high-quality decision making.

In addition, to facilitate the use of the datasets available in the environment for future research, we have built NONA, an expressive and intuitive cohort selection application that enables users to define, select, share and collaborate on cohorts of patient datasets that suit their research questions.

Finally, to enable both statistical and prediction analysis within the INTEGRATE environment the Analysis Platform (AP) was implemented supporting multiple project analysis scenarios.

The INTEGRATE technologies can be used by a large and multidisciplinary biomedical community, ranging from basic, translational and clinical researchers to the pharmaceutical industry, to share data and knowledge, and to collaborate on creating new knowledge with the end goal of improving patient outcome. This paper initially describes our methods in Section 2. Then Section 3, presents the key solutions delivered by the INTEGRATE project: Reconfigurable infrastructure components; applications supporting collaboration and the efficient execution of clinical research; a semantic interoperability solution and a standards-based data model focused addressing the needs of breast cancer research and care. We also detail, in Section 4 the technical validation of our technologies and the clinical evaluation of the solutions carried out with the clinical organizations in the project who are vanguards in research and care in oncology. In Section 5 we present related work and finally Section 6 concludes this paper and presents directions for future research.

2. Methods

2.1. Legal requirements

The first objective that needs to be supported by the INTEGRATE computational platform is to enable efficient and effective data sharing across institutional boundaries. Both health and genetic data are collected and stored by the INTEGRATE platform and as such distributed collaboration and research should be enabled and promoted. Such secondary use of health data has a vital role in improving and advancing medical knowledge [6], but it remains essential that steps are taken to respect wishes of the patient regarding secondary usage, and to ensure the privacy of the patient during such use scenarios. Informed consent [7], together with depersonalization and its related concepts of anonymization, pseudonymization, and data minimization are key methods used to provide this protection [8].

However, in principle, the processing of health and genetic data is prohibited, except if the data subject (the patient) has given his/her explicit consent to such processing (Data Protection Directive [9], Article 8, §2, lit. a) or if one of the exemptions provided for in article 8, §§2–5 is met (not applicable to our case). The consent to the processing of sensitive personal data shall be given in writing and in addition, the data controller cannot legitimately allow the processing of sensitive personal data in the absence of explicit written consent of the patient.

The easiest way to run such a platform from a legal perspective would be to use only anonymous data, as the processing of anonymous data does not need a legal basis or an informed consent of the patient, as anonymous data is not personal data and would therefore not fall under the scope of the Data Protection Directive. Unfortunately, most of the data cannot be processed anonymously. In addition, as the identification of each patient has to be guaranteed in order to give the best therapy, most of the data needed for such a platform has to be processed in a pseudonymous way (meaning replacing a person's name and other identifying characteristics with a label, in order to preclude identification of the data subject or to render such identification substantially difficult). The INTEGRATE platform should carefully conform to the aforementioned requirements and as such an ethical framework should be constructed.

2.2. High-level architecture

The rationale for the technological developments of the INTEGRATE project lie in the actual IT needs of multicenter clinical trials. Setting up, managing and implementing such trials is an arduous and costly process [10]. There is a strong need to integrate the available data and knowledge in comprehensive models supported by interoperable infrastructures and tools, to standardize methodologies, and to achieve wide-scale data sharing and reuse, and multidisciplinary collaboration. It is therefore becoming critical that several IT solutions need to be in place, while novel concepts are also necessary for facilitating security, recruitment and execution of such trials including quality control [11]. The INTEGRATE project proposed and implemented a technological environment based on the following methodological principles:

- A security central authentication framework with Single-Sign-On (SSO) functionality allowing the user to log in once to all the services of INTEGRATE.
- Advanced tools and services for ensuring privacy protection of patients.
- A semantic interoperability solution for the homogenous representation and assessment of clinical data in order to reduce manual operations and render trial data seamlessly interoperable with the specialized tools of INTEGRATE.
- Advanced, specialized tools for automating patient screening processes, enhancing trial recruitment, facilitating the selection of patient cohorts, allowing the remote collaboration of pathologists for reviewing digital pathology data, perform quality control between participating pathology centers and analyzing data securely on the same platform for performing statistical analyses and predictive modeling.

In addition, it follows the principles of the View – Viewpoint model, as formalized in ANSI/IEEE 1471-2000, ISO/IEC 42010:2007 [12] and leverages several design principles: