

Integrated Image Data and Medical Record Management for Rare Disease Registries. A General Framework and its Instantiation to the German Calciphylaxis Registry

**Thomas M. Deserno, Daniel Haak,
Vincent Brandenburg, Verena Deserno,
Christoph Classen & Paula Specht**

Journal of Digital Imaging

The Journal of the Society for Computer Applications in Radiology

ISSN 0897-1889

J Digit Imaging

DOI 10.1007/s10278-014-9698-8


ONLINE FIRST


Journal of Digital Imaging


Innovating Imaging Informatics

The Official Journal of the Society for
Imaging Informatics in Medicine

VOLUME 27 NUMBER 3
JUNE 2014

 Springer

 **SIIM**
SOCIETY FOR
IMAGING INFORMATICS IN MEDICINE
INNOVATING IMAGING INFORMATICS



*JDI Goes Mobile
Download the free app*

10278 • 27(3) 287–418 (2014) www.siiim.org

Your article is protected by copyright and all rights are held exclusively by Society for Imaging Informatics in Medicine. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

Integrated Image Data and Medical Record Management for Rare Disease Registries. A General Framework and its Instantiation to the German Calciphylaxis Registry

Thomas M. Deserno · Daniel Haak · Vincent Brandenburg · Verena Deserno · Christoph Classen · Paula Specht

© Society for Imaging Informatics in Medicine 2014

Abstract Especially for investigator-initiated research at universities and academic institutions, Internet-based rare disease registries (RDR) are required that integrate electronic data capture (EDC) with automatic image analysis or manual image annotation. We propose a modular framework merging alpha-numerical and binary data capture. In concordance with the Office of Rare Diseases Research recommendations, a requirement analysis was performed based on several RDR databases currently hosted at Uniklinik RWTH Aachen, Germany. With respect to the study management tool that is already successfully operating at the Clinical Trial Center Aachen, the Google Web Toolkit was chosen with Hibernate and Gilead connecting a MySQL database management system. Image and signal data integration and processing is supported by Apache Commons FileUpload-Library and ImageJ-based Java code, respectively. As a proof of concept, the framework is instantiated to the German Calciphylaxis Registry. The framework is composed of five mandatory core modules: (1) Data Core, (2) EDC, (3) Access Control, (4) Audit Trail, and (5) Terminology as well as six optional modules: (6) Binary Large Object (BLOB), (7) BLOB Analysis, (8) Standard Operation Procedure, (9) Communication, (10) Pseudonymization, and (11) Biorepository. Modules 1–7 are implemented in the German Calciphylaxis Registry. The

proposed RDR framework is easily instantiated and directly integrates image management and analysis. As open source software, it may assist improved data collection and analysis of rare diseases in near future.

Keywords Clinical trial · Rare disease registry · Electronic data capture · Data management · Image management · Image processing · Image annotation

Introduction

Although sound data is lacking, it is currently stated that there are over 7,000 rare diseases identified and reported, which affect approximately 6–8 % of the world population [1, 2] and an estimated 25 million to 30 million Americans.¹ Hence, investigators in clinical research foster the establishment of rare disease registries (RDRs), which are seen as an essential tool to improve knowledge and monitor interventions for rare diseases [3]. In France for instance, the Ministry of Health has initiated a national plan for rare diseases, which involves 132 reference centers for a specific disease or a group of diseases [4]. Among their missions, these centers are involved in the epidemiological monitoring of pathologies, based on RDRs. Similarly in Korea, a national initiative for rare disease management has been recently established [5]. In Germany, starting in 1999, the Competence Networks in Medicine and Clinical Trial Centers were funded by the German Federal Ministry of Education and Research with an annual budget of about 2.5 million Euro [6]. However, the funded activities have been focused on investigator initiated trials, pseudonymization

T. M. Deserno (✉) · D. Haak · C. Classen
Department of Medical Informatics, Uniklinik RWTH Aachen,
Pauwelsstr. 30, 52057 Aachen, Germany
e-mail: deserno@ieee.org

V. Brandenburg · P. Specht
Medical Clinic I, Cardiology Department, Uniklinik RWTH Aachen,
Aachen, Germany

V. Deserno
Clinical Trial Center Aachen, Uniklinik RWTH Aachen, Aachen,
Germany

¹ <http://www.ncats.nih.gov/about/faq/rare/rare-faq.html#How%20many%20rare%20diseases%20are%20there?>

and randomization tools rather than on information technology (IT) for rare disease management, which is still certified to be based on proprietary university-driven solutions [7].

RDR registries possess a diverse range of functionality, operate in different and often incompatible software environments, and serve various and sometimes incongruous purposes [2]. Consequently, the United States (US) National Institute of Health (NIH) Office of Rare Diseases Research (ORDR)—as manifested in its 2010 workshop report [1]—urgently recommends the development of a minimal common registry model, that should be open source and broadly available. In particular, ORDR has developed specific recommendations for

- (i) *standardized vocabulary, terminology, codes and diagnoses*: which aim at finding commonalities across all rare diseases and at developing a minimal common registry model;
- (ii) *technology and informatics*, which aim at developing an open-source software/hosted registry solution;
- (iii) *biorepositories and biospecimens*, which aim at establishing disease biospecimen repositories using patient registries as sources for donors;
- (iv) *clinical research, patient care and disease management*, which aim at developing centralized registries;
- (v) *patient participation, outreach activities and patient advocacy*, which aim at writing a “Registry-building for Dummies” handbook; and
- (vi) *bioethical and legal issues*, which aim at bringing in ethical as well as regulatory expertise.

Still, however, these goals have rarely been reached. In Europe, 95.3 % of $n=514$ registries are sponsored by or hosted by academic institutions, 3.1 % by companies, and 1.6 % by patient organizations [8]. Regarding the academia majority, neither a standard on database design, nor on functionality, nor on user or data interfaces has been established yet [6, 7]. Even worse, there is no solution to secure sustained funding for rare disease registries, and registries that have been maintained with governmental support for even a decade or more are being terminated [9].

Disregarding the 3.1 % systems that are company driven (and hence eventually well-funded), academic research lacks sufficient financial resources. There is need for an open source framework that can be easily instantiated to RDRs and straightforwardly set up by the IT departments of university hospitals and other academic institutions.

Electronic data capture (EDC) has been essentially researched for controlled clinical trials (CCT) resulting in research-based frameworks (e.g., OpenClinca, REDCap) [10]. Furthermore, commercial systems such as SecuTrial (interActive Systems, Berlin, Germany) exist. However, RDR differ from controlled clinical trials (CCT) because data

is collected rather incidentally. Subjects included in a clinical registry do not undergo well-determined and scheduled examinations but data quantity, data quality, and creation times vary from subject to subject. Therefore, EDC systems for CCT are not applicable to medical registries, suitable software is unavailable generally, and, hence, developed individually.

For instance, Natter et al. have developed a self-scaling chronic disease registry for the Arthritis and Rheumatism Research Alliance [11]. This software is based on the Informatics for Integrating Biology and the Bedside (i2b2) framework [12]. Interfacing medical record entries, their work is restricted to alphanumerical data. Wang et al. have addressed the lack of sufficient integration of binary data such as images and biomedical signals, when developing a user customizable system for rehabilitation clinical trials [13]. However, this work is based on proprietary interfaces and concepts.

The lack of adequate IT infrastructure supporting imaging-based clinical trials has also been addressed by Langer and Bartholmai [14]. In particular, tools are required for multisite imaging collaboration and data mining. More specifically, Erickson, Pan, and Marcus recently modeled the general workflow of an imaging infrastructure for research [15]; among others, (i) development and distribution of the imaging protocol, (ii) insertion of research identifiers, (iii) image transfer, (iv) automated quality control checks, and (v) integration with clinical information were addressed. Here, quality control requires instantaneous and automatic processing of the image bitmap, which is best performed directly after acquisition and transfer of the binary data. Furthermore, the need of tools for manual image annotation and markup has been stressed [16].

In this paper, we address the particular needs of investigators initiating RDRs. The key requirement of modular software supporting EDC of both, binary data and medical records is focused in “[Requirement Analysis](#)”. Accordingly, we implement a suitable framework based on open source components (“[Implementation and Graphical User Interfaces](#)”) and show how this framework is instantiated to support a certain rare disease (“[Example Application](#)” and “[The German Calciphylaxis Register](#)”). The resulting general database model is described in detail in “[RDR Core Modules](#)”. In “[Discussion](#)”, we critically reflect on our approach with respect to other’s work and provide a conclusion.

Materials and Methods

Based on existing RDRs that currently are hosted at the tertiary care academia institution Uniklinik RWTH Aachen, Germany, we perform a requirement analysis and concordantly suggest an implementation design for database and

graphical user interfaces (GUIs). As a proof of concept, we are instantiating the framework to a certain rare disease, i.e., the calciphylaxis, which also is briefly introduced.

Requirement Analysis

According to the methods of Prins and Abu-Hanna [17], we conducted semi-structured depth interviews. The interviews were lasting half an hour till 1 hour. Different experts (three RDR's principle investigators, one chief executive officer of a clinical trial center) and end-users (two experienced study nurses, two rather novice students) were interviewed in order to obtain a final checklist of information services categorized in static and functional services. In particular, the requirement analysis is based on two RDRs, the German Calciphylaxis Registry (ORPHA number 280062 [8]), the German Myeloproliferative Neoplasia Registry on Chronic Diseases (ORPHA number 98274 [8]), and a third clinical registry, where cases of various neuromuscular diseases are collected. The list items were ranked as (i) important, (ii) neutral, or (iii) not important. We ended up with the following itemize:

- *General requirements*: for rare diseases, we can expect individual providers to encounter only very few patients with particular diagnoses, so the case numbers at each institution will often be insufficient for research purposes. Whenever multiple users from different hospitals need to access the same data, internet (web)-based applications are preferable, because they can be accessed easily from anywhere.
- *Study subjects and users*: central to any medical registry is the study subject or patient, whose data is included in the database. However, these subjects do not have access to the database. Nonetheless, they need to be identified uniquely, since the data entry persons need to reidentify subjects to modify or add follow-up data to their clinical records. Furthermore, the registry is used by personal having access to the system for data entry, retrieval, and monitoring.
- *Access control on role level*: referring to the previous item, not all persons shall have access to the registry and not all data that is hosted in the registry shall be accessible by any user. Typically, the physicians may be allowed to see and revise the data records of their "own" patients, where "own" means hosted at their institution, while a monitor is allowed to view all of the entries. Furthermore, it is useful to provide certain rights, for instance, the user is allowed to integrate binary data such as images in the database. Both, access levels and special rights are mandatory functionality for any registry.
- *Electronic data capture (EDC)*: the core purpose of any registry is to collect medical data on subjects electronically. All of such data must be given a data type, which can be either numerical, date/time, or one or more items selected from a predefined list (terminology, cf. next item). Regarding an appropriate statistical assessment, unstructured text is disadvantageous and shall be avoided. Numerical items must have a unit and a reference interval for instantaneous plausibility checks. Of course, EDC is an inherent component of any registry.
- *Terminology*: ontology provides the basic categories of being and their relations. It deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences. To avoid unstructured text in a registry, any such question in an electronic case report form (eCRF) shall be referring to structured terms, where the phrases offered for selection shall not overlap in meaning and completely cover the entire semantic definition range. Hence, terminology is also a required component of any registry. However, such structured lists may need extensions or modifications over the lifetime of a registry.
- *Audit trail*: from its definition, an audit trail or audit log is a security-relevant chronological set of records that provides documentary evidence of the sequence of activities that have affected a data field at any time. A RDR must support the reconstruction of the entire life cycle of any data, starting from its creation or receipt over its use or maintenance to its disposition or erasure. The audit trail is part of any registry.
- *Binary large objects (BLOB)*: In addition, a registry may require collecting binary data such as electrocardiography (ECG) recordings, photographs, or any other medical images that have been acquired from the study subject. Such accompanying binary large objects (BLOBs) have a certain type, which is of particular importance for further automatic processing. For instance, annotated reports may be scanned, added to the database, and linked to the subject. The same holds for photographic sequence protocols, which may further need automatic analysis. Accordingly, the BLOB module shall be designed as optional.
- *Standard operation procedures (SOP)*: it might also be useful to host BLOBs, which are rather general, i.e., not linked to a study subject. For example, the PDF versions of study protocols, descriptions on how to record the ECG, and other instructions, which usually are called standard operation procedures (SOPs). Hence, this module shall be optional, and, furthermore, only applicable if the BLOB module has been installed for a certain registry, as an instance of the framework.

- *BLOB analysis*: images that are captured for the registry may be further analyzed and automatically processed yielding quantitative measurements. Furthermore, manual annotation may be required, e.g., for marking regions of interest (ROIs) in the images. For automatic image analysis, the procedure calls shall be invoked instantaneously after completion of data transfer. Note that image or signal processing is not yet integrated in any registry that has been published so far. Hence, this module is regarded again as optional and conditional.
- *Communication and messaging (Com)*: a valuable but optional module of a medical registry, in particular if applied for rare diseases, is supporting communication and case-based interaction between physicians. Emails between the system users, individually or grouped, sent manually or automatically, may therefore enrich an RDR as an optional component.
- *Pseudonymization (PID)*: privacy concerns or data protection requirements necessitate that any de-identification is used in the first place. In contrast, practical research often requires that individual records can be complemented with subsequent findings, and that duplicate entries of patient should be avoidable. A pseudonym is an identifying tag that a person is given for a particular purpose, such as participating in a medical register. It differs from his or her original or true name (orthonym). Like masks to hide the face, pseudonyms are adopted to hide an individual's real identity but support re-identification. In the register, the pseudonym is used to address data, and the pseudonymization service is used to reproducibly compute a pseudonym, which is not allowing reconstruction of the orthonym. In several registries, the personal identifier (PID) is given manually based on lists and randomization protocols. Anyway, some RDRs instances are requiring access to the medical records of study subjects by identifying data. Then, an optional PID module may ensure that the medical records are disconnected from any identifying data.
- *Biorepository (BioRep)*: alongside clinical data and laboratory values derived from the routine and recorded into the RDR, it is valuable to store body fluids or tissue for later analysis. The samples are normally aliquoted and stored in -80°C freezers. The aliquots are to be labeled—preferably using a barcode label—in such a manner, that pseudonymization is ensured on the one hand and the samples can be related to their patient's registered clinical data on the other hand, at all times. It is also required to give an overview in the database, how many aliquots of which material (e.g., serum, citrate plasma, EDTA plasma, buffy coat) are available and how many have been retrieved already. Furthermore, it is also necessary to be able to insert the results gathered from later analyses into the

RDR [18]. The BioRep module is considered optional and requires the PID module.

Implementation and Graphical User Interfaces

Usability, visual integration, and performance were regarded most important selecting the IT software platform for the registry. With respect to the study management tool that has been developed according to these requirements and that is already successfully operating at the Clinical Trial Center Aachen (CTC-A) [19], the registry core has been implemented using the Google Web Toolkit (GWT),² Sencha GXT,³ Gilead⁴, and Hibernate.⁵ GWT allows development of web applications based on a client-server architecture with Java. During compilation, GWT client-side components are translated into light-weighted JavaScript code, which is visualized in the user's browser. Additional integration of the Sencha GXT library offers GUI widgets for data visualization such as grids, supporting paging, and filtering of large data sets. On server-side, Hibernate and the application programming interface (API) Java Persistence API (JPA)⁶ are used for storage and retrieval of GWT data objects in a MySQL database.⁷ For this, Gilead (formerly Hibernate4gwt) connects GWT with Hibernate and supports exchange of data object between both technologies. All libraries are available under open source licenses (GWT, Gilead: Apache License 2.0⁸; Hibernate: LGPL v2.1⁹; Sencha GXT: GPLv3¹⁰).

Furthermore, the team-based software development is supported by the Redmine flexible project management web application,¹¹ which offers issue tracking for maximizing the team's ability to deliver quickly and respond to emerging requirements. In particular, the Scrumbler plugin¹² for Redmine supports Scrum-based agile software development processes. Furthermore, Apache subversion (SVN) control is used.¹³ Unit and GUI tests are designed using the programmer-oriented JUnit testing framework for Java¹⁴ whereby implementation of interface tests are supported by the browser automation tool SeleniumHQ.¹⁵

² <http://www.gwtproject.org/>

³ <http://www.sencha.com/products/gxt/>

⁴ <http://sourceforge.net/projects/gilead/>

⁵ <http://www.hibernate.org/>

⁶ <http://www.oracle.com/technetwork/java/javaee/tech/persistence-jsp-140049.html>

⁷ <http://www.mysql.de/>

⁸ <http://www.apache.org/licenses/LICENSE-2.0.html>

⁹ <http://www.gnu.org/licenses/lgpl-2.1.html>

¹⁰ <http://www.gnu.org/licenses/gpl.html>

¹¹ <http://www.redmine.org/>

¹² <http://www.redmine.org/plugins/scrumbler>

¹³ <http://subversion.apache.org/>

¹⁴ <http://junit.org/>

¹⁵ <http://www.seleniumhq.org/>

Example Application

Calciphylaxis (calcific uremic arteriolopathy) is still an incompletely understood rare disease, which most often affects patients on haemodialysis [20]. It is a devastating condition associated with high morbidity and a mortality rate >80 % after 2 years. It is characterized by painful, ischemic, partly necrotic skin ulcerations. Pathomorphologically, media calcification of skin arterioles is the hallmark of the disease. The complete clinical picture may include large areas of skin ulceration predisposing to infection.

A national Internet-based registry has been established in Germany in November 2006 to allow online notification for all cases of established or suspected calciphylaxis. The principle investigator (Vincent Brandenburg) obtained approval of the Ethics Committee at the Medical Faculty of RWTH Aachen University (Refs: EK 023/06; amended EK 082/12) before performing the study in accordance with the ICH-GCP standards.

A comprehensive database including various medical parameters concerning patient characteristics, laboratory data, clinical background, and presentation as well as therapeutic strategies was established using paper-based CRFs. So far (November 2013), 213 patients with calciphylaxis have been documented in 7 years: 62 % females; 86 % dialysis (peritoneal dialysis and hemodialysis) patients, median age 67 years (21 years–88 years) [21]. For most subjects, photographic documentation of the skin lesions has been performed, which are currently not stored systematically in the database.

Recently, a continuous photographic monitoring of the disease has been suggested [22], where the number of images per patient is increased significantly regarding the locations on the body and the times of acquisition. Furthermore, such documentation illustrates the demand for RDR-integrated image analysis and management. Hence, the German Calciphylaxis Registry has been selected as test bed for the RDR framework, co-instantly preparing it as a European Calciphylaxis Registry, i.e., designing a multiuser, multicenter, multination registry.

Results

Based on the requirement analysis, some components are seen as mandatory for any RDR, while others are rather optional. Therefore, required components are integrated coring the framework, while optional functionality is provided by additional modules, which can be instantiated only on demand and conditionally. For instance, the “BLOB analysis” module can be added if—and only if—the “BLOB” module has already been added (Fig. 1, left).

RDR Core Modules

As parts of the core functionality of any RDR, we suggest five modules: Data Core, Access Control, Audit Trail, EDC, and Term. All data is stored in a relational database, but—according to the special needs of any instantiation of such a registry—the data tables, their fields, and the respective web rendering may vary.

In the *Data Core*, we assume the study subject (patient) as the main data element, who may be related to (multiple—for instance, in case of movement) study centers (departments), and the persons (users), which are always linked to only one department. Hence, departments are considered as third core table.

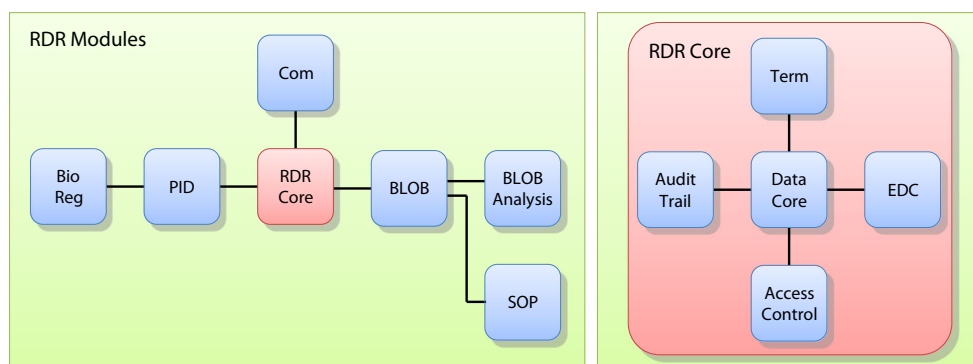
The database model of *Access Control* is shown in Fig. 2. Login is based on user ID (email address) and password. Access is derived from the origin of users, since any person is strictly associated to a single department. These departments have locations (affiliation, city, state, country), which form instant levels on merged access. In addition, a regional cooperation across departments can be defined using the “TypeOfRegion” table. Furthermore, individual rights can be granted to each user by means of Boolean flags and its “TypeOfFlag” relation. To speed up system interaction, all access-relevant information is collected in the central table “Access”, which specifies the internal person model that is used in the code implementation of the framework. This also allows single sign on to several registries, which are defined by the “TypeOfSystem” table. External authentication services that allow users to reuse existing credentials may be interfaced to the “Access” table, too. Both, access level and special rights (modeled as binary flags) are implemented using different views on the data in the database.

The *EDC* component hosts all medical information that is collected for a subject. It is separated from identifying information. The pseudonym is stored in the RDR data core, while identifying information to retrieve or recalculate the pseudonym is hosted in the PID optional module (see “RDR Optional Modules”).

In general, *Terminology (Term)* is modeled via the “TypeOf...” relation tables. For instance, the role of a person follows the terminology provided in the table “TypeOfPerson”. All such database tables hold fields for the name, the short name, and the description of that part of the terminology. Hence, the terminology is simply extensible by adding further lines to the defining table without the need to recompile the application. All tables of that type are automatically included in the help pages, where—after a specific table has been selected by the user—names and descriptions are displayed.

Audit Trails implement a complete logging of any database transaction. Disregarding the data that is changed, all changes

Fig. 1 Modular framework for RDR registries. *Left* composition of mandatory core and six optional modules. *Right* detailed view on the RDR core composed of five modules



are logged in the same table build from the columns (i) User, (ii) Timestamp, (iii) Action, (iv) Revision, (v) Entity, (iv) Property, (v) DataType, (vi) OldValue and (vii) NewValue. The Action identifies whether new data has been created or existing has been modified (i.e., insert or update) and Revision is a counter that is incremented with the transaction. Hence, modifications in the database resulting from the same user action are labeled with the same revision number and can be

easily joined. Entity and Property refer to the database table that has been modified and the according field, respectively.

RDR Optional Modules

So far, we have defined six optional modules, but the modular structure is easily extensible to fit on other needs. Out of those, five have already been implemented. The *BioRep* module has

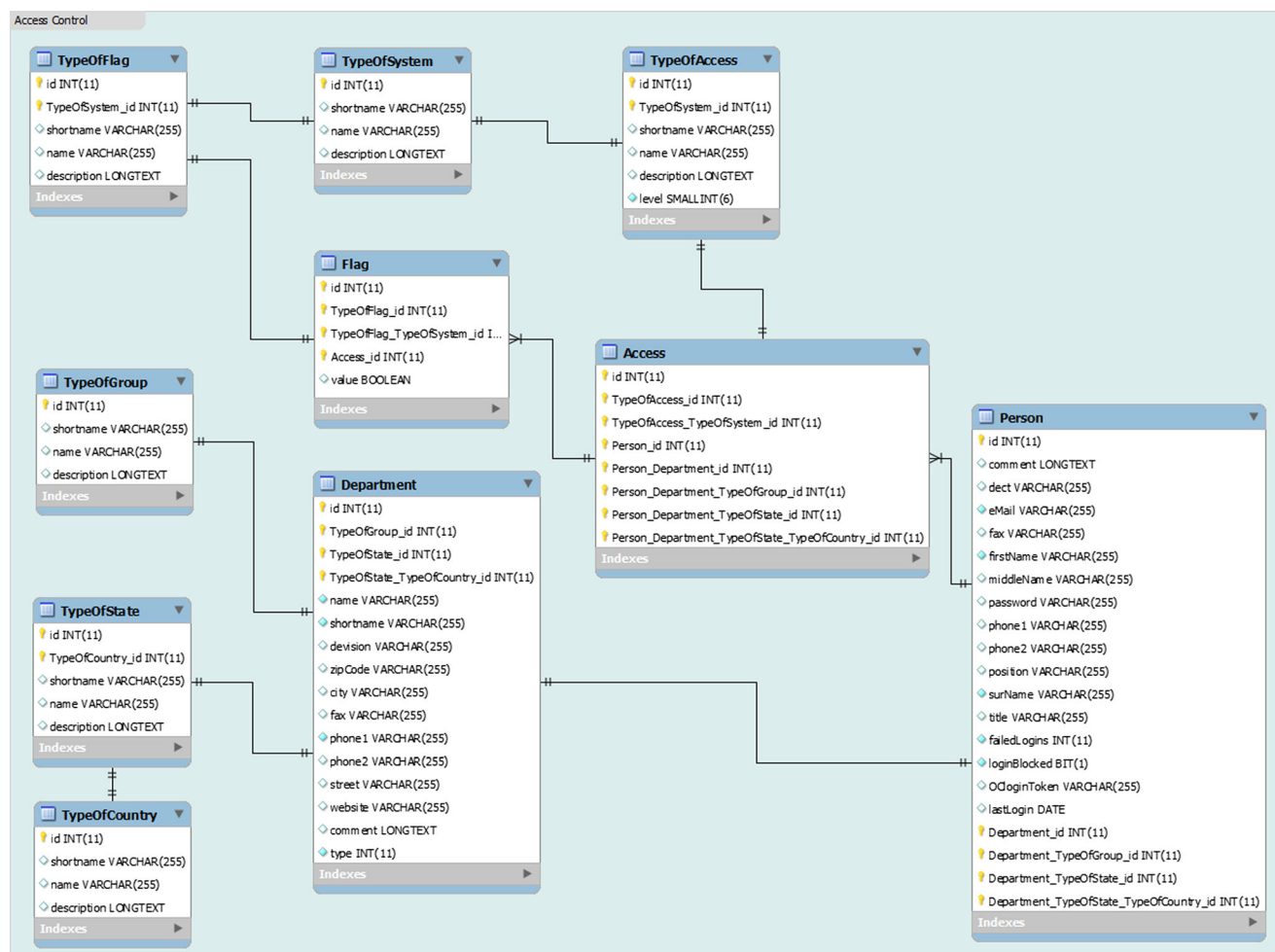


Fig. 2 Access module of RDR framework

not yet been implemented since our university operates a centralized biomaterial bank [23], which is well integrated in the network of German biobanks [24].

Basically, the *BLOB* module ensures that large data files are safely transferred via the internet and attached to the subject's identifier providing the date of filing. The date of transfer and the transferring person's identifier are logged in the audit trail module. Versioning of data is possible, since the module defines a document identifier (DID) that is not unique, and a Boolean flag "latest" indicating the latest version of the respective DID. Terminology is used to classify the types of data (e.g., photograph, ECG recording, scanned diagnostic letter) as well as the according file endings such as the portable document format (PDF), portable network graphics (PNG), or DCM for digital imaging and communication in medicine (DICOM) files. Technically, the *BLOB* module parses, extracts, and handles *BLOB* data received from a hypertext transfer protocol (HTTP) request object. The request object is built by the hypertext markup language (HTML) file upload object, according to the specification of the Internet Engineering Task Force (IETF) [25].

The *BLOB Analysis* module supports manual and automatic data processing. Manual annotations are provided to the user using the GWT Graphics Lib (Apache 2.0 License).¹⁶ Automatic image analysis is integration on server side. The programs are called by remote procedure calls (RPCs). A proprietary Extensible markup language (XML)-based interface is designed to interchange input and output of automatic analysis. Optionally, jar-based algorithms can be outsourced into other GWT applications and called by the web service server component, which is based on JAX-WS¹⁷ Hence, any Java *.jar file is executable, and input as well as output can be controlled via web services. This technique has also been used to integrate image analysis to OpenClinica eCRFs in controlled clinical trials [26].

The *SOP* module is simply establishing a system inherent special patient identifier (pseudonym), where all general data files are linked to. Then, it calls the *BLOB* module with the versioning option enabled and sets the appropriate *TypeOfBlob* terminology (including the file endings that shall be accepted). The required GUI components are also part of that module.

The *Communication (Com)* module refers to different types of electronic communication support such as blackboards, newsgroups, emails, and links to social media. Due to privacy reasons, internal communication is preferable. The disadvantages are, however, that such messages are only accessible if the user logs into the system. In the study management tool, we have implemented an email service sending messages to individuals and groups, which again are coupled to the

intrinsic access model that is defined by the user's affiliation. Such a feature has not yet been linked to our test bed. However, it is worth mentioning that the modular concept of the RDR framework does not only allow adding novel modules but also supports update or exchange of existing modules.

Initiated by the Federal Ministry for Education and Science (BMBF), the German Technology and Method Platform for Networked Medical Research (TMF)¹⁸ is providing a software component for the creation and error-tolerant matching of first-order pseudonyms on the basis of identifying patient data, the so called TMF PID generator [27]. The PID generator uses static data such as first names, date and place of birth, and other not changeable attributes associated to that person to calculate the identifying tag. It is robust to some phonetics and spelling errors, but not capable to handle name changes due to marriages, for instance. Based on the TMF engine, the *PID module* is integrated into the RDR framework.

The German Calciphylaxis Register

So far, the German Calciphylaxis Registry is composed of the RDR core and two of the optional modules, the *BLOB* module and the *BLOB analysis* module (Fig. 3). It supports quantitative color, size, and shape analysis of a sequence of photographs that have been taken from the skin ulcerations [22].

Instantiation of the Core Module

The core strictly follows the architecture of the RDR core of our framework. The access model defines three levels for department, country (already facing the next level of development: European nations), and all. Therefore, the Calciphylaxis Registry is designed for hosting the national registries of all participating countries, too, although so far, only German scientists have contributed data.

Pseudonymization is performed with the centers, and identifying data is not captured in the registry. We have defined according terminologies for all parameters requested in the data fields of the CRFs (Fig. 3).

The terminology *TypeOfExamination* specifies the general type of records, which allows for one history to assess the anamnesis of the subject, one record describing the diagnosis, a medication template that holds all drugs applied including the dose, application scheme, start and end dates, and as much as required follow-up sheets to document the development of the subjects' health over the time.

Figure 4 shows some screen shots taken from the application. On the left hand side, the list of subjects is seen as displayed after successful login. Due to the access model, only the according view on the entire data is given. In the figure, the list is shaded gray because of the popup window that is placed

¹⁶ <http://code.google.com/p/gwt-graphics/wiki/Manual>

¹⁷ <https://jax-ws.java.net/>

¹⁸ <http://www.tmf-ev.de/>

Instantiation of the BLOB Module

In the German Calciphylaxis Registry, the RDR BLOB module allows web-based image integration. Photographs that have been taken on patient's bed site are uploaded and linked to the subject ID in the registry. All images are described by their recording date and the body region that is visualized in the image. For precise localization, the body part terminology has been defined according to the image retrieval in medical applications (IRMA) code for medical images, i.e., a monohierarchical multiaxial classification scheme [28].

Figure 5 (left) visualizes the BLOB module integrated into the German Calciphylaxis Registry. An overview of images is displayed, which can be magnified on the user's selection. The list can be accessed by patient ID, recording date, body region, detailed position, left or right hand side, and any combination of those (filter bars on top).

Instantiation of the BLOB Analysis Module

The BLOB analysis module provides both, manual and automatic image manipulations. For manual annotations, the BLOB analysis module links terminology with images, enabling the user to instantly specify the body region when uploading images by selecting the according parts with the mouse on the sketch of the human body (Fig. 5, right). If the mouse is moved on top of a certain region, it is changed in color, and when the user clicks on that region, the color is not released on displacement of the mouse.

If images are captured with a reference color plate, automatic calibration of color and geometry is applied [22]. However, the original photograph stays accessible due to the versioning of the BLOB module. The processed image is not substituting the captured original, but regarded as updated version, allowing to seek "older" versions using the BLOB modules core functionality.

Discussion

Rare disease registries (RDRs) are an essential tool to improve knowledge, summarize expertise, and monitor interventions for rare diseases [2]. They have to be designed with the agility to evolve and efficiently interoperate in an ever changing rare disease landscape as well as information and communication technology (ICT). The need of efficient RDR frameworks has often been claimed [1, 2], but—to the best of our knowledge—a RDR framework that instantaneous manages and analyzes binary data such as medical images has not yet been published or deployed.

The i2b2-based framework of Natter et al. does not cope at all with images or signals as part of the registry [11]. The recent work of Wang et al. addresses this lack [13]. Both,

image and signal data is linked to the eCRF, but neither image nor signal analysis is integrated. The same holds for the variety of commercial software solutions for EDC. SecuTrial, for instance, has started with version 4.3 to manage image data in its eCRFs, but processing and automated analysis is not supported. Contrarily, our framework allows integrating and processing of image data. Instantiated to the German Calciphylaxis Registry, for instance, we apply geometric and color normalization of all photographs due to a 24-field reference card, that is placed next to the lesion and automatically extracted using the scale invariant feature transform (SIFT) [21].

Bellgard et al. have specified an RDR development checklist for (1) technology choices, (2) software development, (3) interoperability, (4) system design, (5) security, (6) sustainability, and (7) open source [2]. Our modular RDR framework is composed of the RDR core with five mandatory modules and (so far) six additional modules, which may be extended on demand. In line with Bellgard et al. [2], the framework is web-based (GWT) and connects to a relational database (MySQL). Java and JavaScript programming languages are used on a physical IT infrastructure. All libraries in use are open source.

Interoperability is ensured by supporting the definition of terminology without needing to recompile the application. Export of data in comma separated values (CSV) files as well as specific reporting using JasperReports Library¹⁹ is integrated. As part of the core framework, web services are supported and can be used to transfer data from and into any instantiated application. The system design supports specific diseases and clinical registries rather than patient registries. Its extensible modular design ensures that a new terminology, new data elements, and new features or modules can be added at any time. Two-factor authentication and multilevel user access further ensures security of data in the framework-based registries. Following the key criteria list of Bellgard et al. [2], working groups can be established in the access control module using the TypeOfRegion terminology. Encryption and de-identification process is supported accordingly by the RDR PID module. The modular concept further addresses sustainability. Introducing appropriate levels of documentation allows open source distribution of the framework. In summary, almost all of the key criteria that have been cataloged by Bellgard and coworkers are completely fulfilled or at least sufficiently concerned in the proposed RDR framework. Hence, our framework yields robust and sustainable RDR implementations, which has been verified exemplarily by its instantiation to the German Calciphylaxis Registry.

Image and signal analysis is seen essential in any RDR. Hence, two of the optional modules are targeting binary data support. So far, the versioning in the BLOB module follows a simple set-oriented approach, but not really establishes an

¹⁹ <http://community.jaspersoft.com/project/jasperreports-library>

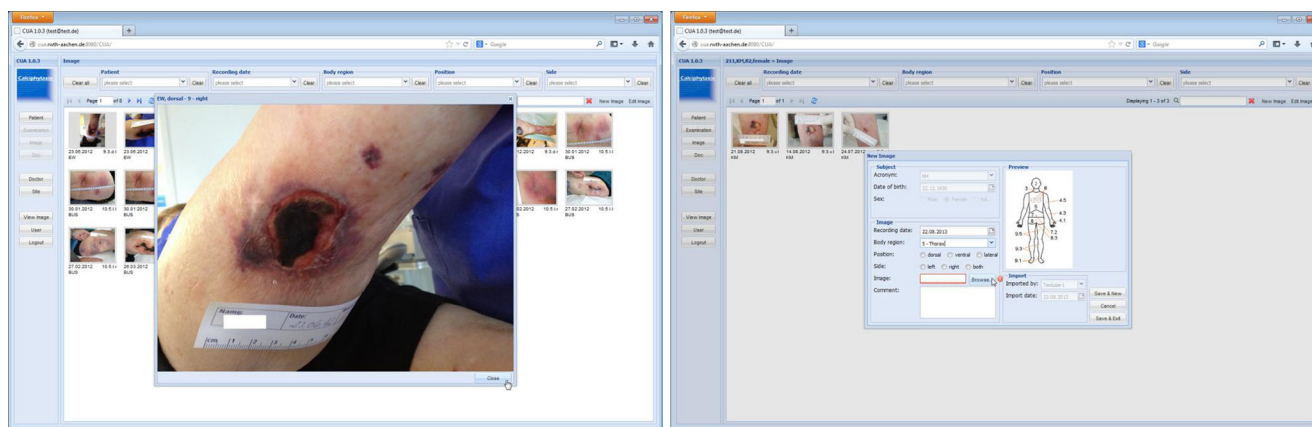


Fig. 5 Framework-based RDR for calciphylaxis. *Left* BLOB module that allows web-based image integration. *Right* detailed view on images

ordered list of versions. We only emphasize the latest BLOB in a set of BLOBs with identical DID. Another limitation is that following our approach, installation of new or modification of existing algorithms currently needs a recompilation of the web application. Although, for instance in 2011, Application Hosting became Part 19 of the DICOM standard [29], and several other initiatives have been proposed [30], plug&play of medical image processing still is lacking in clinical routine and medical research. As mentioned before, we support automatic image normalization for any image uploaded to the registry. Both, instantaneous automatic data processing (as required for quality checks [15]) and manual image annotations and markups (as required for image-based clinical trials [16, 31]) are provided.

The BLOB approach, however, is not limited to signal or image data. Paper-based documents can also be integrated. This might be useful to store medical documents with the patient data. Here, however, the system cannot ensure privacy. There is no mechanism to identify patient names or other identifying data in scans, and it remains in the user's responsibility to ensure that such attachments uploaded to the subject's EDC record is not weakening her de-identification. This, of course, also holds for image and signal data. For example, the aperture in photography should be adequately hiding the identity of the person.

The RDR framework's intrinsic support of web services can be used to interface the RDR to data-delivering instances of a medical information system, such as a laboratory information system. Here, the clinical data acquisition standards harmonization (CDASH) of the clinical data interchange standards consortium (CDISC) are followed and extensible markup language (XML)-based data structure is used for system interconnection. As a next step, a general RDR CDISC module supporting automated data entry into the eCRFs will be connected directly to the RDR core. However, too many standards are existing [32] and harmonization is required before a reliable module is integrated. Hopefully, the

Biomedical Research Integrated Domain Group (BRIDG) model²⁰ is providing solutions here soon [33].

The need for biological tissue registries and aliquot handling is sufficiently recognized in our RDR framework proposal by the BioReg module, but it has not yet been implemented. Appropriate biomaterial handling requires supporting barcode readers as input devices [34]. Within the RDR framework, the sequence of cursor positioning within the eCRF data fields is fully controlled, which allows seamless integration of barcode readers. Taking advantage of the modular concept, the BioReg module may only provide minimal GUI and functionality, while focusing on interconnection of existing bio-bank management systems.

The terminology module allows easy extension or modification of the medical terminology used in the RDR framework. So far, the applied terminology is configured locally and specifically adopted to the registry instance. It has been shown that latest versions of Logical Observation Identifiers Names and Codes (LOINC) almost completely cover medical findings occurring in large hospitals [35, 36]. Bridging the terminology module and LOINC supports data aggregation across registries and, particularly, eases interfacing with other medical information systems. Furthermore, searching and filtering, so far limited to sets of specified database entities, can be extended into a flexible querying over database relations using Hibernate Search.²¹ Full text search over complete databases—combined with LOINC mapping web services²²—will result in a powerful EDC analysis module extending the RDR framework.

In conclusion, our framework represents an ORDR-compliant minimal common registry model that is particularly aligned to the RDR development checklist. The modules have been instantiated to the German Calciphylaxis Registry,

²⁰ <http://bridgmodel.nci.nih.gov/>

²¹ <http://hibernate.org/search/>

²² <http://rxnav.nlm.nih.gov/LoincAPI.html>

integrating EDC and image management. Beside image archiving, automatic image analysis and manual image annotation is supported by the RDR framework adding sustained value to medical research and understanding of rare diseases. The advances in knowledge derived from this study are seen in (i) the system architecture and structure, (ii) the identification and appropriate interfacing of the proposed modules, and (iii) the selection of technology and protocols, which are used for implementation and operation.

In future, expansion to a European Calciphylaxis Registry is planned, without needing to change code or concept, since the German Calciphylaxis Registry is already prepared for a multinational use. In cooperation with the Clinical Trial Center Aachen, further RDRs will be instantiated from the framework.

References

- Rubinstein YR, Groft SC, Bartek R, Brown K, Christensen RA, Collier E, Farber A, Farmer J, Ferguson JH, Forrest CB, Lockhart NC, McCurdy KR, Moore H, Pollen GB, Rangel Miller RV, Hullm S, Vaught J: Creating a global rare disease patient registry linked to a rare diseases biorepository database: Rare Disease-HUB (RD-HUB). *Contemp Clin Trials* 31:394–404, 2010
- Bellgard M, Beroud C, Parkinson K, Harris T, Ayme S, Baynam G, Weeramanthri T, Dawkins H, Hunter A: Dispelling myths about rare disease registry system development. *Source Code Biol Med* 8(1):21, 2013
- Bellgard MI, Macgregor A, Janon F, Harvey A, O'Leary P, Hunter A, Dawkins H: A modular approach to disease registry design: successful adoption of an internet-based rare disease registry. *Hum Mutat* 33(10):E2356–E2366, 2012
- Messiaen C, Le Mignot L, Rath A, Richard JB, Dufour E, Ben Said M, Jais JP, Verloes A, Le Merrer M, Bodemer C, Baujat G, Gerard-Blanluet M, Bourdon-Lanoy E, Salomon R, Ayme S, Landais P: CEMARA: a Web dynamic application within a N-tier architecture for rare diseases. *Stud Health Technol Inform* 136:51–56, 2008
- Seo H, Kim D, Chae JH, Kang HG, Lim BC, Cheong HI, Kim JH: Development of Korean rare disease knowledge base. *Health Inform Res* 18(4):272–278, 2012
- Stausberg J, Altmann A, Antony G, Drepper J, Sax U, Schütt A: Register for networked medical research in Germany. *Appl Clin Inform* 1(4):408–418, 2010
- Drepper J, Semler SC (ed). IT infrastructure in patient-centered research. State of the art and required actions (German). Technical report. IT-Reviewing Board of TMF. Akademische Verlagsgesellschaft AKA GmbH, Berlin, 2013 (ISBN 978-3-89838-690-6)
- Ayme S (ed). Disease Registries in Europe. Orphanet Report Series, Rare Diseases Collection. Orphanet Technical Report. Jan 2013. Available at: <http://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf>
- Souza MP, Miller VR: Significance of patient registries for dermatological disorders. *J Invest Dermatol* 132:1749–1752, 2012
- Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG: Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 42(2):377–381, 2009
- Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, Marsolo K, McMurry AJ, Sandborg CI, Schanberg LE, Wallace CA, Warren RW, Weber GM, Mandl KD: An i2b2-based, generalizable, open source, self-scaling chronic disease registry. *J Am Med Inform Assoc* 20:172–179, 2013
- Kohane IS, Churchill SE, Murphy SN: A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc* 19:181e5, 2012
- Wang X, Martinez C, Wang J, Liu Y, Liu BJ. Development of a user customizable imaging informatics-based intelligent workflow engine system to enhance rehabilitation clinical trials. *Proceedings SPIE* 2014; 9039, in press
- Langer S, Bartholmai B: Imaging informatics: challenges in multi-site imaging trials. *J Digit Imaging* 24(1):151–159, 2011. doi:10.1007/s10278-010-9282-9
- Erickson BJ, Pan T, Marcus DS: CTSA Imaging Informatics Project Group. Whitepapers on imaging infrastructure for research: Part 1: General workflow considerations. *J Digit Imaging* 25(4):449–453, 2012
- Marcus DS, Erickson BJ, Pan T: Imaging infrastructure for research. Part 2. Data management practices. *J Digit Imaging* 25(5):566–569, 2012
- Prins AH, Abu-Hanna A: Requirements analysis of information services for patients on a general practitioner's website—patient and general practitioner's perspectives. *Methods Inf Med* 46(6):629–635, 2007
- Prokosch HU, Beck A, Ganslandt T, Hummel M, Kiehintopf M, Sax U, Ückert F, Semler S: IT Infrastructure Components for Biobanking. *Appl Clin Inform* 1(4):419–429, 2010
- Deserno TM, Deserno V, Legewie V, Schafhausen J, Eisert A, Schmidt-Kotsas A, Kirstein S, Willems J, Spitzer K, Schulz JB. IT support for translational management of clinical trials based on the Google Web Toolkits. German Medical Science (GMS) Meeting Abstract 2011: 11gmms023. (German)
- Brandenburg VM, Kramann R, Specht P, Ketteler M: Calciphylaxis in CKD and beyond. *Nephrol Dial Transplant* 27(4):1314–1318, 2012
- Brandenburg V, Specht P, Floege J, Ketteler M. Seven Years of Experience with the German CUA Registry. Abstract Presentation. American Society of Nephrology Renal Week, 2013
- Deserno TM, Sarandi I, Jose A, Haak D, Jonas S, Specht P, Brandenburg V. Towards quantitative Calciphylaxis. *Proceedings SPIE* 2014; 9035: in press
- Jäkel J, Schmidt R, Leusmann P, Spreckelsen C, Knüchel R, Dahl E, Veeck J: Biospecimen quality management at the RWTH centralized biomaterial bank (RWTH cBMB). *Pathologie* 34(1):136, 2013
- Hirschberg I, Knüppel H, Streh D: Practice variation across consent templates for biobank research. a survey of German biobanks. *Front Genet* 4:240, 2013
- Nebel E, Masinter L. Form-based file upload in HTML. Request for Comments no. 1867. The Internet Engineering Task Force (IETF), Network Working Group. 1995. Available at <http://www.ietf.org/rfc/rfc1867.txt>
- Deserno TM, Haak D, Samsel C, Gehlen J, Kabino K. Integration image management and analysis into OpenClinica using web services. *Proc SPIE* 8674: 0F1-10, 2013
- Wagner M, Glock J, Sariyar M, Borg A. PID Generator. Technical Manual. Version 1.2. Dept. of Medical Informatics and Biometry. Johannes-Gutenberg-Universität Mainz, Germany, 2008
- Lehmann TM, Schubert H, Keyzers D, Kohnen M, Wein BB: The IRMA code for unique classification of medical images. *Proc SPIE* 5033:440–451, 2003

29. Digital Imaging and Communications in Medicine (DICOM). Part 19: Application Hosting. PS3.19-2011. National Electrical Manufacturers Association, Rosslyn, VA, USA, 2011
30. Prior FW, Erickson BJ, Tarbox L: Open source software projects of the caBIG In Vivo Imaging Workspace Software special interest group. *J Digit Imaging* 20(Suppl 1):94–100, 2007
31. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL: The caBIG annotation and image markup project. *J Digit Imaging* 23(2): 217–225, 2010
32. Ohmann C, Kuchinke W: Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. *Methods Inf Med* 48(1):45–54, 2009
33. Fridsma DB, Evans J, Hastak S, Mead CN: The BRIDG project: a technical report. *J Am Med Inform Assoc* 15(2):130–137, 2008
34. Hullsiek KH, George M, Brown SK: Designing and managing a flexible and dynamic biorepository system: a 15-year perspective from the CPCRA, ESPRIT, and INSIGHT clinical trial networks. *Curr Opin HIV AIDS* 5(6):538–544, 2010
35. Lin MC, Vreeman DJ, McDonald CJ, Huff SM: A characterization of local LOINC mapping for laboratory tests in three large institutions. *Methods Inf Med* 50(2):105–114, 2011
36. Dugas M, Thun S, Frankewitsch T, Heitmann KU: LOINC codes for hospital information systems documents: a case study. *J Am Med Inform Assoc* 16(3):400–403, 2009