Improving Epidemiology Research with Patient Registries Based on Advanced Web Technology

Tim A. Majchrzak
Department of Information Systems
University of Münster, Germany
tima@ercis.de

Oliver Schmitt
Department of Information Systems
University of Münster, Germany
research@schmitt.io

ABSTRACT
To store patients’ medical histories and to exchange them between physicians, patient registries are used. Registries contain detailed data on patients and their treatments, and may comprise additional documents. This makes them very valuable for epidemiological research due to the amount of information contained. Providing data for research requires anonymization and pseudonymization to address privacy laws and security concerns. To be able to give feedback to physicians e.g. about discovered treatment possibilities, advanced pseudonymization has to be used. We present progress in the development of patient registries and the required functionality to support research. Usability, explicability, and performance requirements are addressed by incorporating Web technology. Our findings include ways to develop patient registries as well as the description of mechanisms built into them. In particular, we show applications to and implications for crisis management.

Keywords
Epidemiology research, patient registry, registry, electronic data capturing, electronic health record

INTRODUCTION
Taking care of patients requires a large amount of information. If a physician has access to a detailed medical history, treatment effectiveness can be increased (Baumer, Earp and Payton, 2000), the quality of life of patients improved, and complications avoided. Access has to be efficient to not overburden physicians; in cases of crises (think e.g. of mass incidents) having only vital information at hand is essential. Patient registries are used to keep relevant data. They provide an invaluable source for medical research (cf. Koeling, Tate and Carroll, 2011) by providing details about patients’ demographics, treatment plans, therapeutic effects, and treatment success.

Providing patient-oriented data for purposes of research is no straightforward task. Most countries have imposed data-protection and privacy laws (cf. e.g. Baumer, Earp and Payton, 2000) that tend to be particularly strict for medical records. Therefore, means are needed to anonymize data from patient registries in a way in that it can be used by researchers. Simply deleting names does not do the trick. Consider a sample of only ten patients: height would be an insightful parameter if there is a particularly tall person. However, deleting all potentially revealing data is impossible since some studies might e.g. require body height. To make things worse, feedback and the above described synergies require a way to get back to a patient. Thus, pseudonymization should be used.

To support epidemiological research with as much relevant data as possible and to enable fast feedback and help for patients, colleagues from the field of medical informatics first checked the applicable laws (see below). We then specified and build a patient registry using modern Web technology. As the next step, we considered implications for the field of crisis management. This paper is structured as follows. The next section presents the background and related work. We then describe requirements, specification, and implementation of the system. We particularly provide details on how to technologically address these requirements. Based on our experiences, we present a short evaluation as well as relations to emergency preparedness. Eventually, we draw a conclusion.

PATIENT REGISTRIES IN EPIDEMIOLOGICAL RESEARCH
This paper is based on work by our colleagues from the field of medical informatics, who have designed and developed DocFrame, a Web-based documentation framework for electronic data capture for use in clinical care.
Patient registries in epidemiology collect data for long-term observations from a patient population. The main purpose is to induce new knowledge and to test hypothesis with statistical methods over a large set of data. Submission of data can be voluntary or is forced by law for some diseases. Commonly, patient registries for epidemiological research still use paper-based questionnaires in order to retrieve data. However, Web-based systems are on the rise as they decrease costs of data handling and collection. As every medical data collection contains sensitive personal information, privacy requirements of patient registries are high. In general, the entire patient data are available to physicians who submit records to researcher and are in regular personal contact with their patients. The researchers who are often connected to the operators of the registry should access pseudonymized data. The barrier of pseudonymization is often controlled by a supervisory committee. It examines and controls the design and conduction of research to guarantee adherence to ethical standards. External researchers may access data after approval by registry operators and supervisory instances.

Some related approached could be identified. Garson and Adams (2008) propose a mobile triage system that provides decision support based on data drawn from patient registries. While we do not include data analysis and they do not need pseudonymization, there are architectural similarities. Agrawal, Grandison, Johnson and Kiernan (2007) sum up the “U.S. government’s vision of the health care information infrastructure”. Another closely related system is HICCUPS (Molina, Salajegheh and Fu, 2009). The authors propose homomorphic encryption to enable privacy-preserving access by researchers. Christen (2008) presents a system that emphasizes record linkage for improved usage of records in research. We expect more systems to exist than can be described. Some of them are vendor-distributed and most have not been assessed scientifically.

PATIENT REGISTRY REALIZATION

The development process of the registry is described in the following sections.

Requirements Engineering

Requirements engineering was combined with a sophisticated analysis. Firstly, use cases needed for a generic patient registry architecture were determined. Secondly, its functionality needed to connect patient care module to research module was identified. Thirdly, demands derived from guidelines and laws were integrated. Fourthly, requirements were checked with medical researchers. Refinements of the requirements were based on guidelines, especially the generic data protection concept by TMF (Deter and Elsner, 2010). In particular, a pseudonymization algorithm has to be developed. Using it is to be controlled by a supervisory committee.

Functional requirements – i.e. use cases for the registry – were compiled by interviewing hospital physicians and medical researchers. The goal was to understand typical workflows that include patient data. The nine core use cases are patient inclusion, patient exclusion, patient assignment, change of medical center, input of master data, input of static medical data, input of patient histories, quality assurance in the patient database, and contact between researchers and physicians. For reasons of brevity and scope, details are omitted.

Patient care and research module connection is established via three components. An EDC (Electronic Data Capturing) system located in the patient care module collects input data entered by physicians. A chain of software tools is used to transfer data to the research module.

From a general viewpoint, the system has to provide a flexible mechanism for handling changes in the appearance and functionality of input forms. This mechanism has to take into account existing data that has been entered before. It must also be able to handle different revisions of a document that are needed for realizing the audit trail, which also includes a full log of all actions that occurred. Additional requirements arise from the secrecy of medical data, namely restricted access, encrypted data transmission, and data separation.

From a non-functional standpoint, performance and usability are important. Design of the user interface needs to honor that the system should seamlessly integrate into patient care and research workflows.

Design and Specification

To enforce pseudonymization of data, two separated information systems (IS) for physicians (patient care module) and researchers (research module) were designed. Physicians use the electronic data capturing system. Both systems are connected through the pseudonymization module, which provides one-way data conversion en-
forced by cryptography. The main mechanisms to prevent complete leakage of medical and personal data is to separate identifying data (IDATV) from the medical datasets (MDATV). This is done on the servers; the client combines the datasets on access with a temporary access key. An attacker gaining access to one server could only retrieve medical data or personal data. The connection to both servers is encrypted with up-to-date methods and secure settings. In order to generate a temporary access-key for the EDC system to combine separated datasets, a dedicated connection between both servers has to be established.

Pseudonymization works in two steps. Firstly, a cryptographic function is used to derive a new identifier for the patient from the existing identifier (PDI) of the identifying data (IDATV) from the medical datasets (MDATV). The derivation under the usage of a symmetric encryption function works as follows:

$$\text{PSN} = \text{symmetric\_encryption\_function}(\text{PDI} + \text{salt}, \text{key})$$

The salt is a random value that is stored on the IDAT server. It hardens the PSN against attacks that exploit the fact that PIDs are created in a sequential order. Moreover, a copy of the medical data is created and the existing identifier (PID) is replaced with the new identifier (PSN). This new dataset (MDATF) is free of identifying data and contains the identifier that can only be resolved to the real identifier by using cryptography. MDATF is the pseudonymized data set that can be used by researchers. Pseudonymization is done utilizing a tamper-resistant crypto-hardware device, namely a smartcard. To enforce the power of the supervisory committee that has the ethical responsibility of the research, the smartcard is controlled by the committee. Hence, acquiring the secret keys that allow an inversion of the pseudonymization and therefore an identification of the patient by the researcher (with help of the IDAT database) is only possible with approval of the supervisory committee.

**Implementation**

The EDC system is based on DocFrame, a Web-based framework designed for patient registries and patient documentation (eKernPäP) (Brüntrup, Hartz, Lablans, Baumann-Köhler, Zernikow, Müller and Ückert, 2010; Hartz, Verst and Ueckert, 2009). Modern HTML5-enabled Web browsers are used as the client platform. The client logic is implemented in HTML5 and JavaScript. The offline mode with integrated synchronization and conflict resolution enabled by HTML5’s application cache allows capturing data while no network connection is available. The EDC client runs on almost all major PC operating systems as well as on mobile platforms. Although no explicit requirement, this implementation would also suffice crisis response scenarios.

PHP is used for the server logic. The IDAT server is one of the two application servers used in the EDC system. It implements the business logic that is needed to process IDAT data and administrative information, besides managing client authentication and session management. Furthermore, the IDAT server contains the identifying data. The MDAT server offers the business logic that is needed to process medical data. As the IDAT server is responsible for client authentication, the MDAT server is notified by the IDAT server about authentications and sessions. The server offers a RESTful JSON API for the client software that communicates asynchronously with both MDAT server and IDAT server. To ease accessing two servers simultaneously, a reference implementation of the data separation concept of MDAT and IDAT by the TMF for PHP-based Web applications was used (Brüntrup, Hartz, Lablans, Baumann-Köhler, Zernikow, Müller and Ückert, 2009). For storing data CouchDB – a document-based database – is used (Brüntrup, Lablans and Ückert, 2011b).

In order to realize the needed functionality for data transport and processing from the patient care module to the research module, different programs and services are needed (Figure 1). These services have to communicate using HTTP for retrieving data from the EDC system’s programming interface and for serving data to client programs. Data that is stored in the EDC system as JSON-formatted information objects is transported through the entire export stack as JSON objects. Python, which is well-known for rapid prototyping capabilities, was used as the programming language because the inclusion of much advanced functionality makes it easy to use.

The DQS (data quality and consolidation service) is the first stage of exporting data from the patient care module to the research module. This service extracts data from the IDAT and MDAT servers and is the first stage of the export stack. Its task is to find duplicates in the patient datasets by comparing the master data of a patient with an edit-distance-based algorithm. Following duplicate detection, a unique identifier is added to the datasets (UID) to make them identifiable for later usage. After detection is finished, master data are removed from the export data. In the second stage, the identity protection service takes over and fetches the data from the DQS by using its JSON API and secured HTTP for transport. The IPS (identity protection service) provides pseudonymization. Its task is to replace the UID by the PSN. This replacement is done by using the smartcard.

Researchers can view the pseudonymized data with a Web-based viewer. This Web-based data viewer is a lightweight version of the EDC system and allows researchers to inspect the medical data within a Web browser. In
comparison to the EDC system it only uses one application server that accesses the research database. A second server is not required as there is no need for data separation on the researcher side.

![Figure 1. The Export Stack](https://example.com/figure1.png)

**EVALUATION AND DISCUSSION**

The suggested technology of this paper is planned for some future project. Partial aspects have been already integrated into running registry projects at the Department for Medical Informatics (IMBEI) at the University of Mainz. The first system is in productive use at the University Hospital of Münster since February 2011. Nephreg is a Web-based documentation system for long term observation of patients suffering of the cystic kidney disease nephronophthisis (König, Omran, Kurlemann, Schmitt, Lablans, Ückert and Konrad, 2012). To date it covers more than 60 patients from twelve different research centers in Germany. Nephreg consists of the EDC system without further incorporation of the export stack. The identity of the patients is only known to the physicians that enter the patient data and not to the researchers.

The presented work is not limited to its core usage as a patient registry. In fact, it can be seen as a step towards being better prepared for emergencies (also cf. Turoff, Hiltz, White, Plotnick, Hendela, and Xiang, 2009) by incrementally improving the understanding of medical data. This allows better responding to emergencies. Moreover, it enables physicians and medical researchers in systematically avoiding mistakes and in revealing inadvertences in treatment. *Muddling through* (Lindblom, 1959) may seem little scientific at first but it has been found to be well suited for reaching high (even highest in some cases) reliability.

In our case the data in the patient registry is used for studies instead or in combination with data gained in field studies. By having researchers *muddle* through the datasets drawn from the registries for distinct research projects, they not only gain insights with regard to their studies’ aims. In fact, they are likely to reveal hidden (i.e. not obvious) information that might help to improve treatment of patients or to draw results that were not intended at first. At the same time, it is conceptually and technically ensured that arbitrary access to datasets is impossible. In fact, researchers are limited to access data that is non-revealing and in a way that complies with ethical guidelines – yet they are able to learn much more than they could in an ordinary study.

It is hard (if not impossible) to quantify the benefit of these findings or to estimate, what might be revealed in *which* study – in case anything will be found at all. Nevertheless, this fuzziness and the “additional benefit by chance” come at almost no price; moreover, it contributes to the value medical research has both for individuals and for progress in treatment possibilities. Finally, it improves the overall reliability of healthcare.

Due to the novelty of the research, discussion is essential. First of all, it has to be noted that the above described concept and implementation of a patient registry is feasible. Exporting to the research module has not been exhaustively evaluated with large real-world datasets but first results are promising and the concept should be a foundation that future work can rely on. Developing the registry and evaluating its functionality showed that data quality can become a problem – even in closed environments. In settings of multi-national collaborations of a variety of hospitals, physicians, and researchers, data heterogeneity, consistency, and quality are likely to require attention. The same applies to crisis response scenarios that in general have a very dynamic nature. Moreover, a few technical challenges remain. Since multiple servers have to be set up for data separation, export stack, and data drain in the research area, maintenance and support are costly. Furthermore, pseudonymization of binary data is very complicated. Additionally, duplicate detection is a tricky task. Ideally, it should be improved by using a phonetic-based algorithm.

Regarding functional issues, it has to be noted that export of data from the EDC-system to the research database is time consuming. Real-time processing would aid researchers in their work. Finally, an organizational boundary has to be mentioned. Although, the supervisory committee has the power over encryption keys, their members often do not have sufficient knowledge to enforce their power. This might be addressed with coaching.
CONCLUSIONS AND FUTURE WORK

We presented work on electronic patient registries. We highlighted the background of registries and then described requirements engineering, specification, and implementation of a registry in detail. Eventually, we discussed implications. Despite the holistic nature of our approach, open questions remain. Firstly, it is uncertain whether the concept we relied on is perfect. Besides scenario-based testing, measurements during actual operation should be done. At the same time, experiences of operating a patient registry have to be gained. Open questions also concern the data. Moreover, the application to crisis management has to be elaborated in much more detail and tested in a scenario. Future work will likely comprise the iterative refinement and evaluation of both concept and implementation of the registry.

ACKNOWLEDGMENTS

The second author of the paper wrote his master thesis in cooperation with the department of medical informatics at the University of Münster, Germany, in the group of Frank Ückert. During his work he developed important parts of Nephreg. Most developments described in this paper are based on work of the department of medical informatics. We would like to thank Jens König for keeping us up-to-date regarding Nephreg.

REFERENCES