Secure Secondary Use of Clinical Data with Cloud-based NLP Services
Towards a Highly Scalable Research Infrastructure

J. Christoph1; L. Griebel1; I. Leb1; I. Engel1; F. Köpcke1; D. Toddenroth1; H.-U. Prokosch1; J. Laufer2; K. Marquardt2; M. Sedlmayr1
1Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; 2Rhön-Klinikum AG, Bad Neustadt/Saale, Germany

Keywords
Cloud-computing, secondary use, text-mining, privacy, natural language processing, software design

Summary
Objectives: The secondary use of clinical data provides large opportunities for clinical and translational research as well as quality assurance projects. For such purposes, it is necessary to provide a flexible and scalable infrastructure that is compliant with privacy requirements. The major goals of the cloud4health project are to define such an architecture, to implement a technical prototype that fulfills these requirements and to evaluate it with three use cases.

Methods: The architecture provides components for multiple data provider sites such as hospitals to extract free text as well as structured data from local sources and de-identify such data for further anonymous or pseudonymous processing. Free text documentation is analyzed and transformed into structured information by text-mining services, which are provided within a cloud-computing environment. Thus, newly gained annotations can be integrated along with the already available structured data items and the resulting data sets can be uploaded to a central study portal for further analysis.

Results: Based on the architecture design, a prototype has been implemented and is under evaluation in three clinical use cases. Data from several hundred patients provided by a University Hospital and a private hospital chain have already been processed.

Conclusions: Cloud4health has shown how existing components for secondary use of structured data can be complemented with text-mining in a privacy compliant manner. The cloud-computing paradigm allows a flexible and dynamically adaptable service provision that facilitates the adoption of services by data providers without own investments in respective hardware resources and software tools.

Correspondence to:
Dr. Martin Sedlmayr
Lehrstuhl für Medizinische Informatik
Friedrich-Alexander-Universität Erlangen-Nürnberg
Wetterkreuz 13
91058 Erlangen
Germany
E-mail: martin.sedlmayr@fau.de

Methods Inf Med 2015; 54: 276–282
http://dx.doi.org/10.3414/ME13-01-0133
received: December 1, 2013
accepted: October 8, 2014
epub ahead of print: November 7, 2014

1. Introduction

The growing amount of data in operational electronic health records (EHRs) presents unprecedented opportunities for its reuse for many tasks, including quality assurance, benchmarking and clinical/translational research [1]. Thus, many research projects have been initiated in recent years to reuse such data for example for IT-supported patient recruitment [2] or to create institution specific as well as cross-institutional research data warehouses (DWHs) [3–5], which are often used for patient cohort identification. Large European projects such as TRANSFORM [6], EULAR ADR [7] and EHR4CR [8] aim at establishing cross-institutional research data platforms and repositories to support various scenarios within clinical trials’ processes (e.g. the feasibility phase, the patient recruitment, clinical trial data acquisition or adverse drug event detection).

However, reusing EHR data for secondary purposes is not without challenges and problems. Just recently, for example, Hersh et al. have published seven caveats for the use of operational EHR data in comparative effectiveness research [1]. One of those caveats is that data captured in narrative texts (e.g. clinical notes, discharge letters) may not be directly recoverable as structured data elements for further analysis in research projects. In order to also unlock such data and make them accessible for further research, thus “realizing the full potential of electronic health records” natural language processing (NLP) approaches have been proposed by several researchers [9, 10]. Nevertheless, to the best of our knowledge, no framework has yet been established that combines both: the direct retrieval of structured and coded health record data as well as the NLP algorithms, which take unstructured narrative texts from an EHR, extract coded information associated with a particular research question and thus enhances the already existing set of structured data items.

Thus, the primary objective of the cloud4health project was to design an architecture and to define the building blocks necessary within a framework of research tools which should enable the creation of comprehensive research data repositories built upon structured as well as...
unstructured EHR data. Further, since NLP is intensive in terms of computation [11] or maintenance [12] and thus can require huge computing resources, a secondary objective was to explore the promises, but also the challenges of applying highly scalable and flexible computing resources in form of cloud-based services [13]. Finally, since personal medical data are highly sensitive and therefore subject to strong national and state specific data protection regulations [14–16], one of the specific challenges when considering cloud-services is to de-identify clinical data before sending them to the cloud while at the same time retaining its scientific value [17].

The research project cloud4health meets these challenges with a comprehensive set of research tools integrated within the cloud4health architecture. This set of tools supports the stepwise extraction of structured and unstructured clinical data from routine EHR databases, the de-identification of data which shall be further processed by text analyses tools provided as cloud-services and finally reintegrating such NLP results data with the previously extracted structured data within a study specific data repository.

Below, we first present the architecture and its components and then discuss the issues of data protection and scalability/flexibility.

2. Architecture

The cloud4health architecture connects scientific users (customers) of medical data with the providers of such data. A consumer may be an investigator of a retrospective clinical study or a vendor of medical technology while typical data providers are hospitals. Hospitals in particular want to use routinely generated data from their information systems to conduct studies, thus being both: customer and provider. Cloud4health provides the infrastructure to extract structured and free text raw data from one or several data providers, to transform them into the format required by the scientific user and to deploy such data into a central study portal from which the result can be accessed by the customer. In order to comply with privacy regulations, the data have to be de-identified before being provided to the scientific user; k-anonymization and l-diversity [18, 19] may be applied to foster privacy.

We differentiate three privacy models:
- **Model I – anonymization**: All data are anonymized locally by the data provider before any further processing takes place. This means that all potential patient identifying data are removed or replaced “so that the individuals who are the subjects of the data cannot be identified” [19].
- **Model II – local pseudonymization**: Like Model I, only k-anonymous and l-diversified data are provided to the central study portal. However, the data provider uses self-generated pseudonyms to replace patient identifiers before invoking the cloud services. Hence, patients can be re-identified by the data-provider, which makes this model especially suited for in-house projects.
- **Model III – global pseudonymization**: A central trust center provides pseudonyms and thus enables the record-linkage of data sets between several data providers. This is the only model that allows for follow-ups on patients who switch providers.

Model III – and, depending on local regulations, even model II – requires the patient’s informed consent to the processing of the data which makes – from a legal point of view – the implementation of a study more complicated. This is why model I (and partially model II) constitutes the primary focus of the cloud4health project.

The system architecture supports single institution scenarios as well as multi-centric scenarios. A single institution may want to supply a particular subset of its clinical EHR data to either a local research group or to any external partner, whereas in a multi-centric scenario EHR data from multiple data providers shall be integrated and be supplied for further analysis in a central study portal.

2.1 Overview

The architecture is divided into three main areas of concern (Figure 1): the data providers, the text-mining cloud and the central study portal:
- Each data provider extracts both structured and unstructured patient data from the original source (e.g. EHR). The local cloud4health services map these data to a common terminology and de-identify the data. The data provider invokes the text-mining cloud and prepares the resulting data for upload into the study portal.
- Upon request by the data provider, a text-mining cloud processes the de-identified free texts according to a study-specific scheme. The result is returned in the form of structured data. The text-mining cloud in cloud4health is provided by a third party in a protected cloud environment. The text-mining is multi-tenant, i.e. separate instances of the text-mining pipelines will be automatically launched for each request.
- Each data provider uploads data to the central study portal. The customer’s request can then be fulfilled either by the export of the raw de-identified data (e.g. as comma separated values) or by additional analytical services (e.g. graphical reports or data mining results).

The flow of data is requested by the customer and controlled by the data provider. Once the data have been provided to the study portal, the system is reset, i.e. data created temporarily during processing are deleted because of privacy reasons.

2.2 Components

Below, we briefly explain the data flow within the system and the task of each component. For illustration purposes, a hypothetical retrospective study with defined eligibility criteria and a fix target data schema is assumed.
- **First, the Patient-Selector** queries one of the source systems such as the clinical data warehouse or the EHR database for the list of all patient identifiers which are in accordance with the predefined inclusion and exclusion criteria that can easily be gained by structured data filters (e.g. age, sex, diagnoses in ICD-code). Note, that this might not be the final patient cohort, because additional...
exclusion criteria might be available after pursuing the free text analysis.

- The Data-Collector then extracts all relevant data of these patients and loads them into the Staging-Database.
- Since the Staging-Database shall contain all data in a common terminology, the Local-Mapper transforms and maps the local data to a standardized terminology (e.g., laboratory values to LOINC (20)) as a step in the ETL process pursued to load the data from their primary source system into the Staging Database.

When all raw data have been stored in the Staging-Database, they will be de-identified.

- For unstructured data, the protected health information tagger (PHI-Tagger) marks identifying text passages such as names, dates and addresses.
- Once the identifying fields in structured and unstructured data are defined, the content can be substituted: This step is conducted by the IDAT-Translator, which uses a set of study-specific rules to replace the value of the identifying attributes. This allows the Non-structured Replacer (for free text such as clinical notes), the Structured-Replacer (for structured data such as dates) and the Metadata-Replacer (e.g., for HL7 message envelopes) to perform a uniform change of the respective data sets (e.g., a consistent time shift of dates or coarsening of ZIP codes).
- The de-identified (i.e., anonymous or pseudonymous) data are then loaded into the Transfer-Database.

Next, the still unstructured data have to be transformed into structured information by text analysis and text-mining processes based on NLP algorithms.

- The Tempifier connects to the Text-mining Cloud and announces the number and type of documents, which allows for the cloud to instantiate specific Text-mining Pipelines. Before the transfer of the documents, temporary identifiers are assigned to each docu-
ment to prevent record linkage in the cloud.

- A **Text-mining Pipeline** is optimized for the processing of a certain type of document and extracts a structured data set from the free text and returns the result in a structured format to the **Tempifier**, which inserts it into the **Transfer-Database**. The structured data set is predefined by means of a terminology provided by the researcher based on his study endpoints.

At this stage, all data to be provided to a customer are available in a structured format. Before finally exporting this data into the central **Study Portal**, the **Database-Cleaner** prepares the data in three steps:

- Both, structured data and the results from the text-mining process, are merged by removing duplicates; and within future enhancement releases phenotyping components can be included in order to resolve contradictions between data sources [21].
- Now that all patient data are available in a structured form, the study criteria are applied again to further remove non-eligible patients from the result set.
- According to the requirements of the respective scenario, k-anonymization and l-diversity [19] are applied before the provision of the data to the central study portal.

At the end, all data are uploaded into the central **Study Portal**, which aggregates the data from multiple data providers and provides them to the customer.

The local extraction of raw data into the **Staging-Database** has to be implemented according to the individual source systems of each data provider. All other components can be standardized so that a data provider is able to adopt them with minor efforts of customization. The replacement rules of the **IDAT-Translator** and the text-mining target schema are study-specific and can be adapted to the particular data reuse scenario’s requirements. For customization purposes the **IDAT-Translator** provides a dedicated component for parameterizing the replacement rules.

### 2.3 Implementation

The implementation of the architecture is based on existing components by the consortium members as well as on open-source tools to reduce the cost of adoption for new data providers. For the management of the data flow Talend Open Studio [22] has been used to model the ETL (extraction-transformation-loading) process and to orchestrate the local components such as the **Local-Mapper** based on LexEVS [23]. Within the cloud4health project two pilot hospitals (university hospital Erlangen and RHÖN-KLINIKUM AG) extracted their data from their local data warehouse or directly from HL7 messages.

![Figure 2](image.png)

**Figure 2** Process Queue to transfer unstructured text into a structured presentation within a Study Portal: First personal health information is tagged and eliminated, secondly structured data elements are extracted by NLP and exported in an ODM dataset, and finally the extracted data is presented to the end user and can provided for further analysis methods.
They are processed by the above-described components, which are mostly implemented in Java as SOAP-webservices. The adoption of the i2b2 database scheme [24] for the staging and the transfer databases allows for the re-use of tools which have been created for the import/export of data and which were developed by other projects such as IDRT [25].

The trusted cloud is based on OpenNebula [26]. On demand it automatically instantiates Apache-UIMA text-mining pipelines [27] that have been customized to the different kinds of texts which were provided by the data providers. The XML-based Operational Data Model (ODM) [28] is a recognized standard for the exchange of patient data in clinical trials and hence has been chosen to return the structured values of the text-mining process to the Transfer-Database (Figure 2).

Once the data are collected, they are locally prepared by the use of k-anonymization and l-diversity algorithms implemented in the ANON-Tool [29, 30] before they are uploaded to the study portal, which is currently also based on an i2b2-instance but could also be exchanged for tranSMART later since the database scheme is the same.

Three different use cases have guided the cloud4health development:

- A retrospective study aims at analyzing the durability of hip replacements (endoprostheses) by the examination of the type of implant and the surgical procedure of the hip replacement. Until now more than five hundred surgery reports from five different hospitals have been mined to extract parameters such as the use of cement to fix the implant or the existence of antibiotics in the cement.

- The extraction of tumor classification data from pathology reports: Half a million narrative reports on tumor tissues are to be processed to extract tumor classifications (e.g. ICD-O codes and TNM code) and further staging information. This enables the pathologists to use the data for further scientific purposes. Approximately 4,000 already annotated reports serve as a gold standard for the cloud-based text-mining algorithms.

- The analysis of discharge letters to detect medication data and check those against contraindications: as medication – at least in Germany – is often not yet documented in a structured way, the extraction of respective data elements (e.g. the drug agent and its dosage) from free text is a first step to implement pharmaceutical/pharmacovigilance use cases. 200 discharge letters from the Psychiatric Clinic of Erlangen University Hospital have been annotated by medical students and serve as a gold standard for the cloud-based text analysis. The aim is to deliver 800 additional documents by text-mining in less than the time was needed for the manual annotation.

3. Discussion

The secondary use of medical data is a growing field of interest and it is under investigation by many projects. Each of these projects addresses different topics, such as the (secure) research access to medical data and algorithms [31, 32], the use of cloud technologies [11], the identification of patient groups based on already structured data [3, 33], the application of NLP methods for cohort identification [34, 35], or the access of study data by web-based front-ends [36]. However, none of the projects combines all such components within a single comprehensive architecture like cloud4health: the added value of the cloud4health approach lies in the provision of managed tools from end to end – from source system to end user. Cloud4health supports medical studies, which are based on structured and unstructured data with an infrastructure that considers privacy, scalability and flexibility at the same time.

3.1 Data Protection

Many approaches for secondary use have been developed in the United States and hence focus on the requirements of the US health care system and the obligations of the Health Insurance Portability and Accountability Act (HIPAA) [11]. In Europe, privacy requirements depend on federal and state governments, so that many different regulations have to be taken into account [37], which limits the direct transfer of such project results until existing harmonization efforts come into effect. For example, upfront (“prophylactic”) data export into a data warehouse like in [31, 32] is not allowed in many federal states; so each study must be approved before the minimal required dataset can be exported into the study database. Cloud4health provides an infrastructure to be established once while extracting and anonymizing data per study on the fly.

Cloud4health originates from the German health care system and meets its laws and regulations. The design of cloud4health with regards to data protection – such as the de-identification of free texts and the data processing within the cloud – has been discussed both with data protection officers from three hospitals on a local and regional level as well as with a state commissioner on a national level. It was also presented at the working committee for data protection in healthcare of the TMF e.V. [38]. So far, all parties have approved the architecture.

Still one issue remains unresolved: The elimination of identifying elements in free texts cannot be performed to a full extent in an automated process. For each new document type a manual training of the PHI-Tagger is necessary in order to improve the quality of the detection algorithms until approvable results can be attained [39, 40]. However, tests have shown that the difference in quality between the tagging of all identifying elements by hand and an automated implementation by the PHI-tagger is very small. The responsible officers for data protection of three hospitals confirmed that they do not expect a 100% elimination (which in any case cannot be achieved, regardless of whether a manual or an automated annotation is used). This is why we are confident to continue this automated approach.

3.2 Flexibility

Flexibility refers to many facets of the infrastructure. First, it is easy to integrate new data providers: only the process of data provision within the Staging-Database is dependent on the local IT-infrastructure.
and has to be implemented individually. All other components are provider independent or can easily be configured to local settings. Second, the text-mining cloud is not bound to a trusted third party: it might also be hosted within a private cloud of a computing center of a large hospital. The main advantage is that data can be fully processed in-house and no anonymization is required. This is an example for the paradigm of cloud computing to provide services in a flexible and scalable way [41]. Third, to set up a new study, it is only necessary to identify the clinical source documents and to specify the terminology in order to configure the local mapping (e.g. [23]) and the NLP-processing pipelines [42].

Furthermore, the dynamic resources of the cloud allow for an increase in the amount of data that can feasibly be processed. While individual data providers might only have a few hundred documents for some studies, the number of both studies and data providers is not limited by technical constraints; in our pathology use-case, for example, hundreds of thousands of documents are to be processed. This also means that the text-mining algorithms can be chosen according to quality and not only according to performance. We use different algorithms on the same text in parallel and we combine the results to compensate uncertainties in the findings and to improve the outcome.

In addition, the distinction of text-mining from the de-identification of free text [40], which computational requirements are less demanding, allows for outsourcing the resource-intensive task of extracting the structured information from the text to third-party clouds if desired. Although the text-mining pipelines could be installed and maintained locally, the use of appliances reduces the burden on the IT department [43]. As software as a service (SaaS) [44] the deployment of the components is managed centrally and thus provides immediate benefit to all users. It also reliefs data providers from substantial upfront investments and thus facilitates the adoption of services. This kind of flexibility provided by the cloud-computing paradigm therefore yields advantages beyond scalability. Especially for smaller data providers with barely any IT staff managed and virtual appliances and SaaS offers can lower the threshold for participation in secondary use projects. This data provision process is often not described or cared for by secondary use projects.

3.3 Limitations and Future Research

The development and the evaluation of the prototype are still going on. Currently, the output from the text-mining is evaluated against the gold standards as presented in the section on the implementation. We also have implemented an additional private cloud at the site of one of the data providers to demonstrate the flexibility of the technical approach. It turned out, that the transfer of the data from local resources into a cloud – whether local or remote – can be neglected in comparison with instantiation and processing time needed for text documents given the typical speed of the internet. A systematic and thorough benchmarking of the end-to-end processing on a large scale has yet to be performed. However, we see the main benefit of our cloud approach in the reduced effort of data providers to participate in secondary use projects due to managed and provided appliances.

Although the described components have already been transferred among the participating hospitals, they still have to be revised to allow for a ready-to-deploy solution (e.g. as an appliance). We see this as a necessary step in order to acquire additional data providers and to ensure low investment and low maintenance costs.

Future visions for development are the extension of the study portal with tools for data mining by establishing a platform like tranSMART as well as the possibility to connect the local components at the data provider with the text-mining pipelines within the cloud by an i2b2 hive.

Two issues further hinder the easy adoption of data services like cloud4health. First, there is no standard interface to acquire data from clinical source systems [45]. Thus, if a data provider has not yet established a clinical data warehouse, new processes for data extraction, transformation and loading into the Staging Database have to be established for each research project. This is why cloud4health appliances include generic ETL-components. Second, transinstitutional data integration still requires cumbersome terminology mappings since no standardized common terminologies do yet exist [3, 46].

At present, the cloud4health prototype does not allow for the amendment of data once a study set has been exported. This function would require the pseudonymization of patient data to link the records, which is more complex from an organizational point of view because of the necessary informed consent of the patient (see above) and hence not yet fully implemented. Furthermore, the kind of data processing is not suited for pro-active or continuous data transfer from the data provider to the study portal that would be required for time critical observations such as serious adverse events.

4. Summary

The secondary use of clinical data is an important field in terms of scientific aspects and economic efficiency. As a consequence, it is necessary to provide a flexible and a scalable infrastructure that is compliant with privacy. Cloud4health has designed an architecture and has implemented a prototype that fulfills these requirements. It provides a researcher with structured information that has been extracted from structured and unstructured data according to the privacy model I (anonymization). The implementation is based on established standards and proven open-source tools. It can be provided as virtual appliances and manages services (SaaS) to lower the burden for participation of smaller data providers. The function and the interaction of all components have been successfully tested with hundreds of documents in three use cases and were acknowledged by responsible data protection officers.

Acknowledgment

The research has been co-funded by the German Federal Ministry of Economics and Technology within the Trusted Cloud initiative (FKZ 01MD11009).
The present work was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg.

References


28. J. Christoph et al.: Secure Secondary Use of Clinical Data with Cloud-based NLP Services