# Cloud4health – On effective ways to deal with sensitive patient data in a secure Cloud environment

Steffen Claus[1], Horst Schwichtenberg[1], Julian Laufer[2], Florian Berger[2]

[1]Fraunhofer SCAI, Schloss Birlinghoven, 53754 Sankt Augustin
[2]RHÖN-KLINIKUM AG, Schlossplatz 1, 97616 Bad Neustadt a. d. Saale

{steffen.claus, horst.schwichtenberg}@scai.fraunhofer.de
{julian.laufer, florian.berger}@rhoen-klinikum-ag.com

## Abstract

The cloud4health project researches secondary analysis of clinical patient data, such as surgery- and discharge-reports in a secure and trusted Cloud infrastructure. Given the data's sensitive nature, a main emphasis rests on guaranteeing its confidentiality during the course of the analysis. The paper outlines infrastructure developments of the first year of the cloud4health project and highlights requirements towards a secure Cloud environment. The first solution architecture is sketched and the lifecycle of data processing is presented.

## 1 Introduction

During the course of a patient's treatment in clinics, several documents emerge along the way; from initial diagnostics to surgery- and discharge reports. Partly, this documentation consists of unstructured information, i.e. free text, whose style and semantic clearness varies widely between the respective responsible doctors. Utilizing such patient data for so-called secondary usage scenarios, e.g. for retrospective studies on drugs' side effects or effectiveness of certain surgery methods, exhibits a great potential but requires sophisticated text mining tools. Depending on the size of the study, respectively the number of relevant documents, increasing requirements towards processing power and memory cannot be met by single clinic infrastructures. Utilizing Cloud computing resources for these scenarios offers a promising approach but guaranteeing the patient data's confidentiality throughout the whole lifecycle of data processing represents a challenging task.

### 1.1 The cloud4health project

The cloud4health (c4h) project is funded by the German Federal Ministry of Economics and Technology in the funding program "Trusted Cloud" (FKZ 01MD11009) and researches secure and trusted secondary use of clinical patient data in a Cloud infrastructure. The project demonstrates practical relevance by concentrating on real-world use cases. For example, cloud4health analyzes narrative surgery reports in implantations to extract information about the type of prosthesis and the kind of operation technique. Thus, it allows statistical analysis on the effectiveness and durability of implants.

To support such secondary usage scenarios, the project develops several Cloud services. The main service – the text mining - applies Natural Language Processing (NLP) techniques for analyzing clinical patient data, more specifically the contained free text. This service utilizes study-specific medical terminologies provided by a terminology management service. Analyzed - i.e. semantically annotated - patient data can be further processed through data mining services. These Cloud services can be offered as Software-as-a-Service (SaaS). Depending on the considered secondary usage scenarios or specific clinic requirements, different deployment scenarios are envisaged and supported.

In this context, the project proposes three solution architectures. In the first solution, patient data is pre-processed and anonymized before it is sent to the text mining services. This pre-processing step removes personal-identifiable attributes from the patient data, thereby transforming it into so-called de facto anonymized data. The second solution applies clinic-internal pseudonymization which allows tracing patient's documents after the analysis in the Cloud. The third solution enables across-clinics pseudonymization. Figure 1 shows an overview of the implementation of the first solution and sketches its three central building blocks: the hospital infrastructure, the text mining Cloud and the study portal. A more detailed description of the individual components is found in section 3.

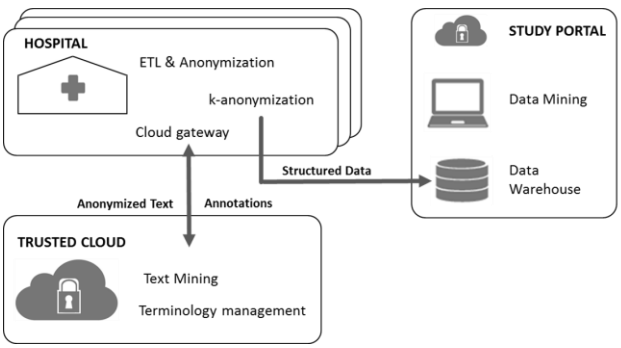


**Figure 1** cloud4health (c4h) architecture

## 1.2 Related Work

There are several projects in the area of secondary use of clinical data, some of them making use of NLP techniques [1], [2], [3]. Some undertakings are focused on specific use cases [4]; others support a wide range of potential research scenarios [5]. Secondary use of clinical data within the German regulatory framework is researched by [6]; special focus on privacy enhancing tools is laid by [7].

In the area of collaborative use and management of Electronic Health Records (EHR) initiatives such as [8] try to facilitate patient treatment across clinics, while [9] concentrates on underlying security aspects of such processes. A clinical data management system with a strong emphasis on data protection and security measures is sketched in [10]. Challenges introduced by legal and ethical aspects - especially in multinational clinical data management – are discussed in [11].

Projects such as [12] and [13] research secure ecosystems for use cases within the health sector. Other undertakings concentrate on fundamental principles of securing clinical data in Cloud usage scenarios [14], [15]. To secure clinical patient data before delivering them to secondary analysis workflows in external infrastructures, one approach is to pre-process and remove patient identifying information [16]. Potential re-identification risk of such pre-processed data is researched by [17].

On the one hand, most of the aforementioned projects don't apply to the German data protection framework, including its diverse regional regulations and hospital laws. On the other hand, some of the initiatives focus on German regulations, but don't incorporate challenges that arise through the use of Cloud computing. Security measures are often only applied for a specific use case or only fit to a certain part of the whole ecosystem, including clinic infrastructures, public transport networks and clinic-external Cloud infrastructures. Summing up, there is no integrated solution for a secure and trusted secondary use of clinical data in Cloud infrastructures that is explicitly focused on the German regulatory framework and includes all relevant players – clinics, data protection officers, legal experts and Cloud providers. Cloud4health approaches to close this gap.

# 2 Secure Cloud environment

The compute- and memory-intensive processing of patient – i.e. the text mining - is handled by a Cloud testbed infrastructure provided by the Fraunhofer institute SCAI. This testbed is used by the partners and clinics within cloud4health and thus represents a community Cloud. It constitutes a potential deployment model for the Cloud services and can furthermore serve as a best practice implementation for processing patient data in a Cloud in a secure and trustworthy manner. Given the sensitive nature of patient data, the Cloud provider should meet high security requirements regarding the data's confidentiality. As clinics are the essential data providers and users of the Cloud services, security measures not only have to be aligned to applicable data protection regulations, such as the BDSG (German Federal Data Protection Act) or the diverse LDSGs (Regional Data Protection Acts), but also to hospital laws and clinics' specific regulations and guidelines for data processing in external infrastructures. On that basis and in close cooperation with the clinics and legal experts within cloud4health, fundamental requirements for secure and trusted patient data processing in an external Cloud have been gathered. These aspects – outlined in section 2.1 – are incorporated into the current operated testbed infrastructure.

## 2.1 Fundamental security requirements

Only in case of personal data the aforementioned legislative regulations – e.g. BDSG and LDSG - apply. But under certain conditions, also non-personal and de facto anonymized data can become personal data again, e.g. through inside knowledge (keyword: k-anonymity). Thus, the Cloud provider should implement appropriate security measures in order to protect the data's confidentiality and thereby reducing the risk of data re-identification.

As a first step, the Cloud provider should perform a detailed risk analysis of the respective infrastructure and its components. This analysis should specify and examine diverse worst-case scenarios regarding breaches of data confidentiality and has to group the infrastructure components into different categories related to their protection needs. Based on this classification, the provider can implement technical security measures accordingly. Guidelines from the working groups of the German federal data protection officers [18], [19], the procedures from the German Federal Office for Information Security (BSI) [20], [21], [22] or the recommendations from the European Network and Information Security Agency (ENISA) [23] are advised to be followed. In the course of the analysis, it is beneficial to directly construct an information security management system, e.g. by following the BSI standard 100-2 [24], which is closely related to the well-known ISO standards 27001 [25] and 27002 [26].

Furthermore, the concrete purpose of data processing not only has to be explicitly defined but also maintained appropriately. Thus, security measures must guarantee that personal patient data won't be used for any other purpose than the predetermined one. Therefore, unauthorized users – respectively their processes - mustn't access other users' data; multi-tenancy has to be supported and separation of data processing has to be ensured on all levels during the complete lifecycle of data handling - from transfer to storage over processing to deletion. Ideally, personal patient data should not be stored in the Cloud infrastructure at all. At least, it should be deleted after a reasonable or agreed upon period of time. Furthermore, an incident response process is required, preferably integrated in the aforementioned overarching concept for security management. Finally, the user of the Cloud infrastructure has to be able to check and verify the security measures, e.g. through a certification executed by trusted third party experts. Even though there are standards for

self-assessment and self-certification, an independent verification of security measures is always preferred.

## 2.2 Technical security measures in the Cloud

The project performed an in-depth risk analysis of all components of the Cloud testbed following abovementioned state-of-the-art techniques [18], [19], [20], [21], [22], and [23]. This analysis was oriented at the procedures as described by the BSI in its standard "100-2: IT-Grundschutz-Vorgehensweise". Given its broad view on potential risks and accompanied suggestions for mitigating measures, data confidentiality is ensured throughout the whole workflow of data processing in the Cloud.

The following list shows an excerpt of the fundamental security measures applied in the Cloud testbed:

- **Secure data transfer over public networks:** Confidentiality of the data transfer from the clinic to the virtual machines hosting the text mining services is guaranteed by encryption (through OpenVPN, [27]). The selection of cipher suites and the definition of key lengths closely follow the respective BSI guidelines [21], [22].
- **Exclusive text mining services for each user:** The virtual machines hosting the text mining services are directly initialized by the user. Only he maintains control over the virtual machine, other users aren't authorized to access any of the contained services. Exclusive allocation of Cloud nodes is supported as well to further strengthen the separation of different Cloud users.
- **Limited lifetime of virtual machines:** Virtual machines are immediately terminated once the data processing has finished. Thus, patient data only resides in the Cloud as long as it is needed for the respective text mining processes.
- **No persistence of personal patient data:** All data generated during the document processing, such as log files and temporary files, is discarded during the shutdown of the respective virtual machine. No patient data or any data related to the input documents is stored in the image of the virtual machine or anywhere else in the Cloud infrastructure, e.g. in an object storage or a shared file system.
- **Secure virtual machine image storage:** The basic images for the virtual machines are kept in a secure central storage. Once an image is requested for start-up, it is copied to the respective execution node via an encrypted channel (SCP). Depending on the users' preferences, encryption techniques and key lengths can be adjusted accordingly. The basic images serve as templates and don't contain any sensitive patient data or other information related to the respective user.
- **Separation of Cloud-internal communication:** The text mining service is distributed and scalable: multiple virtual machines containing the text mining services can be started, the services themselves support multithreading. Thus, patient data processing can potentially be spread over multiple Cloud nodes. The required communication between those nodes, respectively the virtual machines, is supported by dynamically set-up VLANs. Each user gets its own VLAN, which maintains a separation of the Cloud-internal data transfer.

With these security measures integrated in the overarching security concept, cloud4health addresses the requirements and challenges gathered in section 2.1. Even though the Cloud testbed hasn't been certified yet, the carried out steps support the preparation for certification processes based on e.g. EuroPriSe [28] or [24].

# 3 Life cycle of data processing

Patient data processing starts within the clinic's infrastructure (see overview in Figure 1). Data from clinic-internal systems such as hospital information systems or data warehouses is extracted and harmonized. Subsequently, this patient data is anonymized. This step includes the replacement of certain identifying attributes, such as names, dates of birth and post codes in unstructured text as well as in the document's meta-data. Preprocessed documents are then stored in a so-called transfer database. The transfer database enables clinical data protection officers or another personal to control and verify the data set that is intended for analysis in the Cloud. All these abovementioned steps – extraction, harmonization, anonymization and storing - are integrated into an ETL workflow (Extract, Transform, Load). A more detailed description of this workflow can be found in [29].

Figure 2 shows the sequence of data processing in the Cloud. The clinic's gateway to the Cloud is embodied by a component that maintains an encrypted connection to the text mining services in the Cloud. As well, this gateway temporarily assigns random IDs to the documents before sending them to the text mining services. Thus, documents inside the Cloud cannot be traced back to the underlying patient. Before the data processing can be started, the Cloud gateway initializes and configures the required virtual machines and their integrated text mining services via a Cloud management interface. The initialization is controllable by parameters such as the number of documents or the document type, resulting in an adjusted number of started virtual machines or different text mining services. Within the Cloud, documents are processed and annotated based on study-specific terminologies. For example, documents within a hip surgery follow-up study would be searched for occurrences of types of hip joint prostheses, applied surgery methods or other medical terminology e.g. describing the initial bone structure of the patient. These terminologies are dynamically provided by the terminology management service.
Once all documents have been processed, annotated and transferred back to the hospital, the Cloud gateway termi-

nates all virtual machines. No temporal or any other patient-related data is persisted in the Cloud. Before the processed documents are uploaded to the study portal for further data mining and study-specific processes, the data's confidentiality can be further improved by computing k-anonymous sets of the input documents.

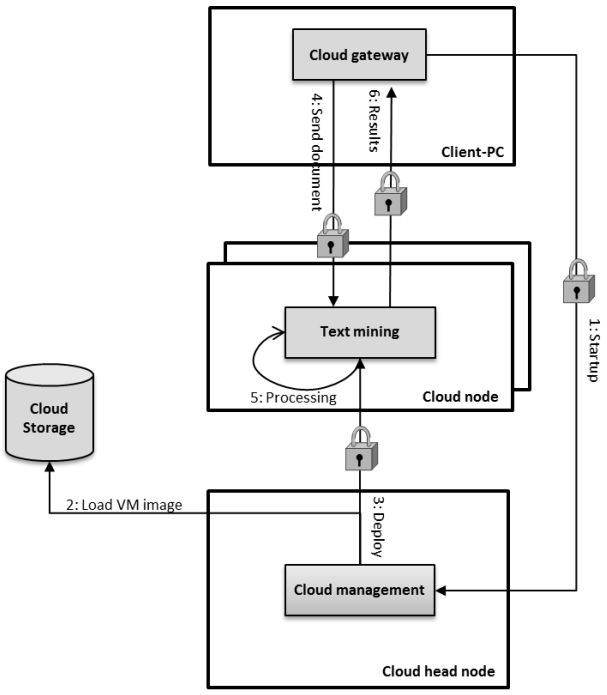


**Figure 2** Sequence of data processing in the Cloud

# 4 Conclusion and future work

The cloud4health project successfully implemented a prototype of the first solution architecture in March 2013. Based on this initial architecture, further use-cases will be realized. The initialization procedure is extended to support dynamic up- and down-scaling of virtual machines during runtime. Load tests with hundreds of thousands of patient documents will be used to benchmark this solution. With the integrated approach and the first prototypical architecture, cloud4health demonstrated the feasibility of processing sensitive patient data in a Cloud in a secure and trusted manner. Data protection and security concepts have been composed in close cooperation with clinical data protection officers and legal experts and thus serve as best practices for similar projects.

The project cloud4health is funded by the German Federal Ministry of Economics and Technology in the funding program "Trusted Cloud" (FKZ 01MD11009).

# 5 References

[1] Chard, K.; Russell, M.; Lussier, Y.; A; Mendonça, E. A; Silverstein, J. C.: A Cloud-based Approach to Medical NLP. AMIA Annual Symposium proceedings, 2011, S. 207–216.

[2] Carrell, D.: A Strategy for Deploying Secure Cloud-Based Natural Language Processing Systems for Applied Research Involving Clinical Text. Proceedings of the 44th Hawaii International Conference on System Sciences, IEEE Computer Society, 2011, S. 1-11.

[3] Chute, C.; Pathak, J.; Savova, G.: The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. AMIA Annual Symposium Proceedings, 2011, S. 248–256.

[4] Weber, G. M.; Murphy, S. N.; McMurry, A. J.; Macfadden, D.; Nigrin, D. J.; Churchill S.; Kohane, I. S.: The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. Journal of the American Medical Informatics Association 16(5), 2009, S. 624–630.

[5] EHR4CR Consortium: Electronic Health Records for Clinical Research, http://www.ehr4cr.eu, last visited 10.06.2013.

[6] Berliner Forschungsplattform Gesundheit, http://medinfo.charite.de/forschung/berliner_forschungsplattform_gesundheit/, last visited 21.06.2013.

[7] Ganslandt, T.; Mate, S.; Helbing, K.; Sax, U.; Prokosch, H. U.: Unlocking data for clinical research - The German i2b2 experience, Applied Clinical Informatics 1(4), 2011, S. 116-127.

[8] Elektronische Fallakte, http://www.fallakte.de/, last visited 21.06.2013.

[9] Hupperich, T.; Löhr, H.: Flexible patient-controlled security for electronic health records. International Health Informatics Symposium, 2012, S. 1–5.

[10] Deserno, T. M.; Deserno, V.; Lowitsch, V.; Franck, W.; Willems, J.; Löbner, H.: Aspekte des datenschutzgerechten Managements klinischer Forschungsdaten, GI-Jahrestagung, 2012, S. 1491-1505.

[11] Rahmouni, H. B.; Solomonides, T.; Mont, M. C.; Shiu, S.: Privacy compliance and enforcement on European healthgrids: an approach through ontology, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368(1926), 2010, S. 4057-4072.

[12] Cloudi/o – Sicheres Cloud-basiertes Datenmanagement im Umfeld der klinischen Forschung, http://www.cloudi-o.de/, last visited 21.06.2013.

[13] TRESOR – Trusted Ecosystem for Standardized and Open cloud-based Resources, http://www.cloud-tresor.com/about-tresor/, last visited 21.06.2013.

[14] Neuhaus, C., Wierschke, R., Löwis, M. von, Polze, A.: Secure Cloud-based Medical Data Exchange, GI-Jahrestagung, 2011

[15] Li, M.; Yu, S.; Zheng, Y.; Ren, K.; Lou, W.: Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption, IEEE Transactions on parallel and distributed systems, *24*(1), 2013, S. 131–143.

[16] Pommerening, K.; Helbing, K.; Ganslandt, T.; Drepper, J.: Identitätsmanagement für Patienten in medizinischen Forschungsverbünden, GI-Jahrestagung, 2012, S. 1520–1529.

[17] Dankar, F. K.; El Emam, K.; Neisa, A.; Roffey, T.: Estimating the re-identification risk of clinical data sets. BMC medical informatics and decision making, 2012, 12(1), 66.

[18] Arbeitskreis Technik und Medien: Orientierungshilfe – Cloud Computing, 2011.

[19] Arbeitskreis Technische und Organisatorische Datenschutzfragen: Technische und organisatorische Anforderungen an die Trennung von automatisierten Verfahren bei der Benutzung einer gemeinsamen Infrastruktur, 2012.

[20] Bundesamt für Sicherheit in der Informationstechnik: Eckpunktepapier - Sicherheitsempfehlungen für Cloud Computing Anbieter, Bonn, 2011.

[21] Bundesamt für Sicherheit in der Informationstechnik: Kryptographische Verfahren: Empfehlungen und Schlüssellängen, Bonn, 2013.
[22] Bundesamt für Sicherheit in der Informationstechnik: Kryptographische Verfahren: Empfehlungen und Schlüssellängen – Verwendung von TLS, Bonn, 2013.
[23] ENISA: Cloud Computing - Information Assurance Framework, 2009.
[24] Bundesamt für Sicherheit in der Informationstechnik: BSI-Standard 100-2, IT-Grundschutz-Vorgehensweise, Version 2.0, Bonn, 2008.
[25] ISO/IEC 27001:2005 - Information technology – Security techniques – Information security management systems requirements specification, 2005.
[26] ISO/IEC 27002:2005 - Information technology – Code of practice for information security management, 2005.
[27] OpenVPN, Version 2.3.1, 03/2013.
[28] Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein: EuroPriSe Criteria, Kiel, 2011.
[29] Griebel, L.; Leb, I.; Christoph, J.; Laufer, J.; Marquardt, K.; Prokosch, H.U.; Toddenroth, D.; Sedlmayr, M.: Cloud-Architektur für die datenschutzkonforme Sekundärnutzung strukturierter und freitextlicher Daten, Proceedings of the eHealth2013, 2013, S. 59-64.