Datenschutzkonforme Sekundärnutzung strukturierter und freitextlicher Daten mittels Cloud-Architektur

Data privacy compliant secondary use of structured and unstructered data via cloud-architecutre

Ines LEB [a,*], Lena GRIEBEL [a], Jan CHRISTOPH [a], Igor ENGEL [a], Julian LAUFER [b], Kurt MARQUARDT [b], Hans-Ulrich PROKOSCH [a], Dennis TODDENROTH [a], Martin SEDLMAYR [a]
[a] *Lehrstuhl für Medizinische Informatik, Friedrich-Alexander-Universität Erlangen-Nürnberg*
[b] *RHÖN-KLINIKUM AG, Bad Neustadt/Saale*

Zusammenfassung. In cloud4health werden medizinische Routinedaten, die wichtige Informationen über die Patientenversorgung enthalten, zusammengeführt und ausgewertet. Dafür realisiert cloud4health auf der Basis einer bedarfsgerechten Anonymisierung oder Pseudonymisierung einen flexiblen, cloudbasierten Lösungsansatz zur Erschließung sowohl strukturierter als auch freitextlicher Daten. Besonders herausfordernd ist hierbei, Data-Warehouse- und Data-Mining-Technologien mit Textanalyse-Werkzeugen datenschutzkonform in einer Cloud zu kombinieren, um umfangreiche medizinische Rohdatenbestände beispielweise zur Überprüfung der Behandlungsqualität und damit letztlich zur Verbesserung der Patientensicherheit in einem zentralen Studienportal zu integrieren.

Abstract. Within cloud4health clinical data which contain important information about patient care and treatment are consolidated and analysed. Based on an on-demand anonymization and pseudonymization cloud4health realizes a flexible, cloud-based solution to retrieve both structured and unstructured data. Particularly challenging is the data privacy compliant combination of data warehousing and data mining technologies with text analysis tools. With this combination comprehensive raw data is integrated in a central study portal which offers the possibility to check and improve the quality of care and eventual to increase patient safety.

Keywords. Cloud-Computing, Systemarchitektur, Sekundärnutzung, NLP, Datenschutz, Textmining

* Corresponding Author.

Einleitung

Im medizinischen Umfeld werden zunehmend Daten digital erfasst und gespeichert. Diese entstehen primär bei Aktivitäten rund um die Behandlung eines Patienten. So werden bei der Aufnahme des Patienten administrative Daten erfasst, die später zur Abrechnung dienen. Als Teil der Diagnostik werden Laborwerte oder Bilddaten erzeugt und ergänzen dann weitere Angaben zur Therapie, wie beispielsweise OP-Berichte. Solche Daten sollen neben der klinischen Routine auch vermehrt für die klinische Forschung genutzt werden [1, 2, 3].

Dabei liegen die für Forschungszwecke relevanten Daten oft in unstrukturierter Form wie in Arztbriefen vor. Natural Language Processing (NLP) ermöglicht die semantische Erschließung solcher Textinformationen. So kann hier z. B. die freitextlich gespeicherte Operationsmethode eines Patienten aus dem OP-Bericht erschlossen und für die Sekundärnutzung vorbereitet werden.

Um entsprechend leistungsfähige Computerressourcen zur Verarbeitung dieser u. U. umfangreichen Daten zur Verfügung zu stellen, bietet sich Cloud-Computing [4] an, welches die dynamische Nutzung großer Computerressourcen über mehrere Kliniken unabhängig von lokal vorhandener Hardware ermöglicht. Dabei muss allerdings stets der Datenschutz für die Sekundärnutzung der Daten gewährleistet sein [5].

Methoden

Mit Scrum [6] wurde eine agile Vorgehensmethode gewählt, um die vielfältigen Anforderungen der Technologien, des Datenschutzes und der medizinischen Fragestellungen iterativ zu erfassen und in Sprints konsekutiv umzusetzen. Dabei hat Scrum zum Ziel durch regelmäßiges Dokumentieren des Fortschritts, Überprüfen der Produktfunktionalitäten und Anpassung der Anforderungen je nach aktuellem Bedarf die Komplexität in einem Projekt zu reduzieren.

Für die Realisierung von cloud4health wurden zunächst vier Anwendungsfälle spezifiziert, die als Basis für die Implementierung sowohl des Prototypen als auch der cloud4health-Infrastruktur dienen. Die Planung bzw. der Fortschritt der Implementierung wurde in 8-wöchentlichen Sprints dokumentiert und wöchentlich diskutiert.

Die cloud4health-Architektur ermöglicht, dass verschiedene Datenlieferanten ihre strukturierten und freitextlichen Daten jeweils lokal aufbereiten und anschließend die für den Anwendungsfall benötigten Daten strukturiert in eine zentrale Datenbank liefern. Dafür konnte auf bereits existierende Erfahrungen bezüglich Sekundärnutzung und Data-Warehousing zurückgegriffen werden (z. B. [1, 4, 7]). Um das System für die Datenlieferanten kostengünstig anzubieten, wurden für die Implementierung der Architektur frei verfügbare Werkzeuge und Standards verwendet.

Ergebnisse

1. Architektur

Der Architekturentwurf von cloud4health bildet ein strukturelles Rahmenwerk, um die einzelnen Komponenten mit ihren Verbindungen untereinander gesamtheitlich darzustellen. Entscheidend ist hierbei die Trennung der verschiedenen Elemente der Datenverarbeitung. So erfolgt der Extract-Transform-Load (ETL)-Prozess dezentral in der datenliefernden Klinik und die Cloud-Datenpools können bei dem verantwortlichen Betreiber betrieben werden.

Dafür wurde die Architektur für cloud4health in drei Bereiche untergliedert, die in Abbildung 1 detailliert gezeigt werden:

- Lokale cloud4health-Services: Erschließung und Deidentifizierung (Anonymisierung/Pseudonymisierung) der strukturierten und freitextlichen Rohdaten bei jedem Datenlieferanten vor Ort (ETL-Prozess).
- Textmining-Cloud: Annotation von Freitexten, geschützter Raum für Daten-lieferanten als Arbeitsumgebung für studienspezifische Instanziierung, Text-mining und Rückgabe strukturierter Ergebnisse an Lieferanten.
- Cloud4health-Studienportal: Zusammenführung der Daten mehrerer Lieferanten in einem Studienportal, das neben dem Zugriff auf die Daten auch Services zur Auswertung (z. B. Reporting, Datamining) zur Verfügung stellt.

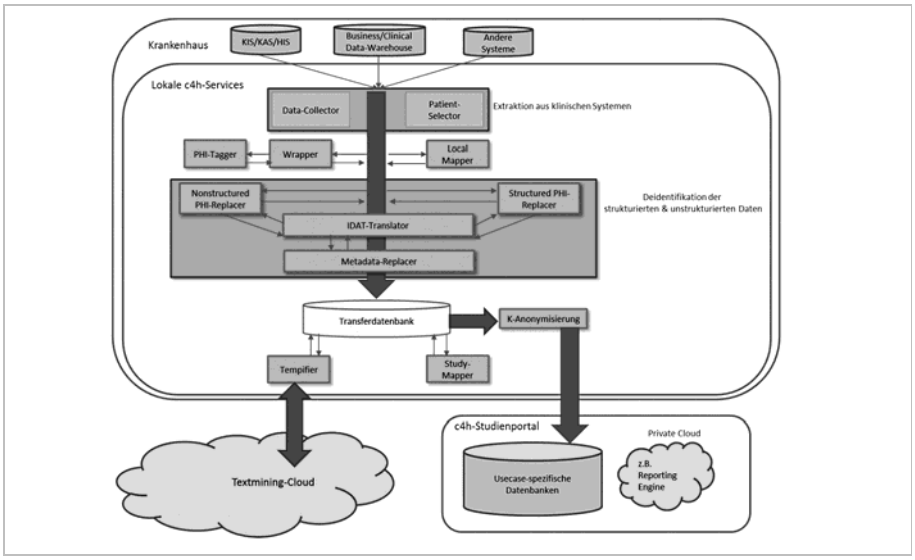

Abbildung 1. Architektur-Rahmenwerk von cloud4health

Die cloud4health-Architektur ermöglicht ein hohes Maß an Flexibilität mittels Skalierbarkeit. So können die lokalen Services als klassische Anwendungen oder auch als Appliances in einer privaten Cloud betrieben werden. Weiterhin kann die Textmining-Cloud als Community, öffentliche oder private Cloud realisiert werden. Dies ermöglicht beispielsweise großen Klinikkonzernen den Eigenbetrieb der Textmining-Cloud, aber auch kleinen Kliniken die Cloud als externe Dienstleistung zu nutzen.

Hohe Flexibilität bezüglich des Datenschutzes wird erreicht durch drei Ausbaustufen, sog. Architektur-Modelle, die unterschiedliche Schutzbedürfnisse implementieren [5]:

- Stufe 1: Anonymisierung: Die Daten aus den Quellsystemen werden lokal anonymisiert (Entfernen identifizierender Merkmale) und beim Export in das zentrale Studienportal zusätzlich k-anonymisiert.
- Stufe 2: Lokale Pseudonymisierung erlaubt einem Datenlieferanten den eige-nen Patienten zu reidentifizieren. Wenn ein Patient z. B. alle Kriterien für eine Studie erfüllt und in diese eingeschlossen werden soll, wird eine Reiden-tifizierung dieses Patienten notwendig. Vor dem Export in eine zentrale Stu-diendatenbank werden die Patienten jedoch wie in Stufe 1 k-anonymisiert.
- Stufe 3: Klinikübergreifende Pseudonymisierung erlaubt das Zusammenfüh-ren von Daten eines Patienten aus mehreren Kliniken (record linkage). Wenn ein Patient in mehreren Kliniken behandelt wird (z. B. Erstimplantation-Hüftprothese in Klinik 1, Hüftprothesenwechsel in Klinik 2), kann diese Zu-sammenführung von Daten notwendig sein. Um den entsprechend kritischen Datenschutz für diesen Fall zu gewährleisten, sollte beispielsweise das Einverständnis des Patienten dokumentiert werden.

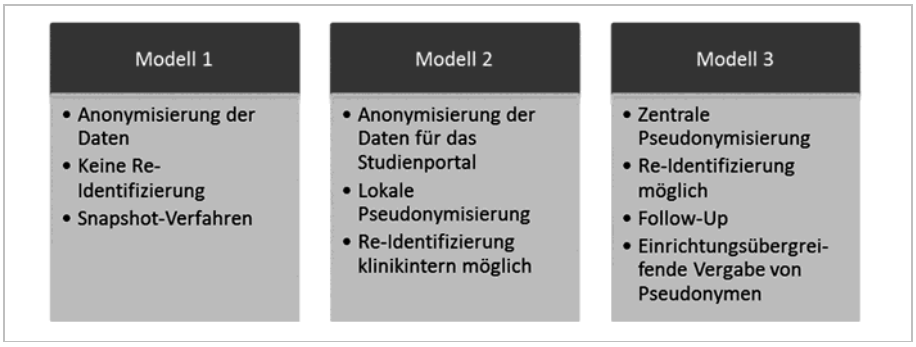

Abbildung 2. fasst die einzelnen Punkte der verschiedenen Modelle zusammen

2. Ablauf in den lokalen cloud4health-Services

Auf Grund rechtlicher Rahmenbedingungen z. B. BDSG §4, der eine Verarbeitung personenbezogener Daten ohne Einwilligung der betroffenen Person untersagt, sowie diverser Gesetze aus den Bundesländern (z. B. Landeskrankenhausgesetz, Landesdatenschutzgesetz), ist eine Anonymisierung oder Pseudonymisierung der personenbezogenen Daten für die wissenschaftliche Forschung unabdingbar. Der Bereich "Lokale c4h-Services" gewährleistet, dass alle strukturiert und unstrukturiert vorliegenden Daten nur ohne identifizierende Informationen das Klinikum verlassen und erfüllt somit diese rechtliche Anforderung.

Nachfolgend werden die dafür notwendigen Schritte erläutert:

- Patienten, die in eine Studie eingeschlossen werden sollen, werden in dem Quellsystem entsprechend der Einschlusskriterien (z. B. ICD, Alter, Geschlecht) durch den *Patient-Selector* identifiziert. Anschließend werden alle benötigten Daten durch den *Data-Collector* extrahiert.
- Bei Freitexten werden alle identifizierenden Attribute (z. B. Name, Datum, Adresse = persönliche Gesundheitsdaten / PHI) im Text durch entsprechende XML-Tags *(PHI-Tagger)* markiert.
- Sofern hier strukturierte Daten vorliegen, erhält der *Local Mapper* die Datensätze, um krankenhausspezifische Elemente auf standardisierte Termi-nologien (z. B. Laborwerte auf LOINC) zu mappen.
- Identifizierende Daten können sowohl in strukturierten, unstrukturierten als auch in Metadaten der Kommunikationspakete vorkommen. Deswegen werden alle drei Arten von Daten betrachtet und durch dedizierte *Replacer*-Komponenten deidentifiziert (anonymisiert bzw. pseudonymisiert). Um einheitliche Ersetzungen sicherzustellen, werden Ersetzungsregeln eingesetzt, die der *IDAT-Translator* bereitstellt. Die *Replacer*-Komponenten kontak-tieren den *IDAT-Translator*, um die Ersetzungsregeln abzufragen.
- Nach der Extraktion und der Deidentifizierung aller Daten, werden diese in der *Transferdatenbank* gesichert und an den *Tempifier* zur Vorbereitung der Übertragung an die Textmining-Cloud weitergegeben.
- Um die Sicherheit zu steigern, generiert der *Tempifier* eine temporäre ID für jedes Dokument, die nur während der Bearbeitung in der Cloud gültig ist.
- Nach Abschluss der Bearbeitung der Dokumente in der Textmining-Cloud, übergibt der *Tempifier* alle Dokumente in strukturierter Form wieder an die *Transferdatenbank*.
- Vor dem finalen Export in das cloud4health-Studienportal kann ein zusätzliches Mapping auf studienspezifische Terminologien durchgeführt werden. Zusätzlich erfolgt eine k-Anonymisierung, um den Datenschutz der aggregierten Daten zu gewährleisten.

Somit verlassen die Klinik nur deidentifizierte Daten. Wenn diese Daten in unstrukturierter Freitextform vorliegen, wird deren weitere Verwendung durch Annotationskomponenten in einer Text-Mining-Cloud ermöglicht. Strukturierte Daten werden, nachdem sie durch den oben beschriebenen Ablauf anonymisiert wurden, direkt an das Studienportal übertragen. Ein Abzug aller anonymisierten bzw. pseudonymisierten Daten wird in das cloud4health-Studienportal kopiert.

3. Prototyp

Bis März 2013 wurde ein erster vollständiger Prototyp entwickelt, der die Infrastruktur für die Realisierung des ersten Anwendungsfalls liefert. Hierbei werden mögliche Determinanten der Behandlungsqualität von Hüftprothesenimplantationen (bspw. zementierte vs. nicht-zementierte Implantate) auf Basis der Auswertung von OP-Berichten und Arztbriefen untersucht. Bisher wurden 550 Arztbriefe und mehr als 580 OP-Berichte von ca. 250 Patienten annotiert. Weitere Anwendungsfälle beziehen sich auf die Erschließung von Pathologieberichten (500.000 Freitexte liegen vor), die Plausibilitätsprüfung von Abrechnungen bei Krankenkassen sowie die Pharmakovigilanz zur Identifizierung von unerwünschten Wirkungen bei Medikamenten.

Diskussion

Die Sekundärnutzung klinischer Daten zu Forschungszwecken ist auch das Thema anderer Forschungsprojekte. Dabei gibt es unterschiedliche Ansätze in der Architektur der einzelnen Systeme [3, 4, 7, 8, 9, 10].

Bisherige Forschungsprojekte zur Sekundärnutzung von Routinedaten haben es zum Ziel, Abfragen auf der Basis definierter Merkmale (z. B. demographische Merkmale, Diagnostik, Laborwerte) durchzuführen, um passende aggregierte Patientenkollektive zu identifizieren (z. B. *SHRINE – Shared Health Research Information Network* [7]) oder Daten aus elektronischen Krankenakten für die Forschung zu nutzen (*EHR4CR – Electronic Health Records for Clinical Research* [8]). Als Projektbasis dienen dabei, im Gegensatz zu cloud4health, jeweils nur strukturiert vorliegende Routinedaten. Im Rahmen von *SHRINE* können auf Grund der fehlenden Deidentifizierungsvorgänge einzelner personenbezogener Patientendaten diese nicht herausgegeben werden, so dass nur aggregierte Anzahlen extern verfügbar sind. Cloud4health mit seinen flexiblen und sicheren Deidentifizierungs- und Textmining-Komponenten hebt sich dadurch von diesen beiden Projekten ab, indem einerseits unstrukturierte Daten verfügbar gemacht werden sowie Zugriff auf die (anonymisierten) Daten selbst gewährt wird.

Auch Scott et al. [9] befassen sich mit der Sekundärnutzung klinischer Routinedaten, indem sie eine relationale Datenbank entwickelt haben, die Patientendaten

einer Intensivstation für Studien zur Verfügung stellt. Die sog. *Multiparameter Intelligence Monitoring in Intensive Care II (MIMIC-II)* Datenbank liefert hierfür ein webbasiertes Tool und eine virtuelle Maschine, um die Suchanfragen zu optimieren. Allerdings müssen die Suchanfragen durch SQL-Abfragen durchgeführt werden. Die Anwender der *MIMIC-II* müssen sich daher zuerst über ein mitgeliefertes Handbuch in SQL einarbeiten.

Chard et al. [4] bedienen sich ebenfalls der Vorteile einer Cloud, um die Abfragen für z. B. eine wissenschaftliche Studie zu optimieren. Die in Chicago entwickelte Architektur, die sog. *Smntx* ermöglicht eine hohe Skalierbarkeit, da die virtuellen Ressourcen erst bedarfsgerecht angelegt werden. Zusätzlich bietet Smntx drei verschiedene Anwenderinteraktionen. Je nach Bedarf kann Smntx über eine Weboberfläche, eine Programmiersprache oder Workflows gesteuert werden. Jedoch werden die Daten zurzeit noch nicht anonymisiert, so dass sie nur dem jeweiligen Datenlieferanten zur Verfügung gestellt werden dürfen.

Eine bereits in der Praxis erprobte Datenbank beschreiben Hurdle et al. [10]. Mit der *Utah Population Database Limited (UPDB-L)* stellen sie eine relationale Datenbank mit einem webbasierten Abfragetool zur Verfügung. Mittels der durch die *UPDB-L* ermittelten Kohorten wurden bereits Studien zu Brustkrebs und Spondylarthritis erfolgreich durchgeführt. Allerdings verwendet diese Architektur keine Cloud und kann somit auch nicht die entsprechenden Vorteile z. B. bei der Skalierbarkeit nutzen [10, 11].

Darüber hinaus analysierten Hruby et al. [1] den Nutzen für Systeme zur Sekundärnutzung von klinischen Rohdaten z. B. für retrospektive Studien. Dafür wurden Anzahl und Qualität (bestimmt durch den Impact-Faktor) der Veröffentlichungen der Columbia Urologie zwischen Januar 2005 bis Dezember 2011 betrachtet. In diesem Zeitraum stiegen durch die Nutzung des *Centralized Research Data Repository (CRDR)* sowohl die Anzahl als auch die Qualität der Veröffentlichungen.

Auch wenn mit diesen Forschungsprojekten bereits eine Basis für Systeme zur Sekundärnutzung von klinischen Rohdaten geschaffen wurde, so beschäftigen sich diese nur mit den Anforderungen des amerikanischen Gesundheitssystems und der Erfüllung der Richtlinien des *Health Insurance Portability and Accountability Act (HIPAA)* [12]. Cloud4health spezialisiert sich nun auf das deutsche Gesundheitssystem, um dessen Gesetze und Richtlinien zu erfüllen. Dabei sind Vorgaben von verschiedenen Bundesbehörden und Ländern sowie viele unterschiedliche Gesetze zu berücksichtigen. Da sich diese von den amerikanischen Richtlinien und Gesetzen unterscheiden, wäre eine Übertragbarkeit nur begrenzt möglich.

Zusätzlich gelingt es cloud4health, alle Komponenten für eine effiziente Sekundärnutzung von Routinedaten wie Cloud-Technologien, Textmining-Anwendungen und Datenschutz-Anforderungen in seiner Architektur in einen einzigen zentralen Prozessablauf zu kombinieren.

Danksagung

Das cloud4health-Konsortium besteht aus der Averbis GmbH (Konsortialführer), der RHÖN-KLINIKUM AG, der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V., dem Fraunhofer-Institut SCAI und der Friedrich-Alexander-Universität Erlangen-Nürnberg. Das Projekt wird im Zeitraum 12 / 2011 – 11 / 2014 vom Bundesministerium für Wirtschaft und Technologie (BMWi) im Förderprogramm Trusted Cloud gefördert.

Referenzen

[1] Hruby GW, McKiernan J, Bakken S, Weng C. A centralized research data repository enhances retrospective outcomes research capacity: a case report. Journal of American Medical Informatics Association 2012, 20:1-5.

[2] Klein A, Ganslandt T, Brinkmann L, Spitzer M, Ueckert F, Prokosch HU. Experiences with an interoperable data acquisition platform for multi-centric research networks based on HL7 CDA. AMIA Annu Symp Proc, 2006: 986.

[3] Li, Z., et al., ClinData Express – A Metadata Driven Clinical Research Data Management System for Secondary Use of Clinical Data. AMIA Annu Symp Proc, 2012. 2012: p. 552-7.

[4] Chard KM, Russell M, Lussier YA, Mendonca EA, Silverstein JC. A cloud-based approach to medical NLP. AMIA Annu Symp Proc 2011: 207-216.

[5] Pommering K., Drepper J., Ganslandt T., Helbing K., Müller T., Sax U., Semler S., Speer R. Das TMF-Datenschutzkonzept für medizinische Datensammlungen und Biobanken. INFORMATIK 2009 – Im Fokus das Leben. Lecture Notes in Informatics 154 (2009), 197; 1744-57.

[6] Scrum.org. Improving the Profession of Software Development (2013). Online verfügbar unter: http://www.scrum.org/, zuletzt geprüft am 06.03.2013.

[7] Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, Kohane IS. 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): 624-630.

[8] EHR4CR-Konsortium. Electronic Health Records for Clinical Research. Online verfügbar unter http://www.ehr4cr.eu, zuletzt geprüft am 11.06.2013.

[9] Scott DJ, Lee J, Silva I, Park S, Moody GB, Celi LA, Mark RG. et al. Accessing the public MIMIC-II intensive care relational database for clinical research. BMC Medical Informatics and Decision Making 2013: 13:9.

[10] Hurdle JF, Haroldsen SC, Hammer A, Spigle C, Fraser AM, Courdy SJ. Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database. Journal of American Medical Informatics Association 2013; 20(1):164-171.

[11] Mell, P. and T. Grance, The NIST definition of cloud computing. Recommendations of the National Institute of Standards and Technology., N.I.o.S.a.T.U.S.D.o. Commerce., Editor 2011, National Institute of Standards and Technology. Gaithersburg.

[12] U.S. Department of Health & Human Services (2012). Health Information Privacy. Online verfügbar unter http://www.hhs.gov/ocr/privacy/, zuletzt akutalisiert am 14.03.2012, zuletzt geprüft am 06.05.2013.