CLOUD-ARCHITEKTUR FÜR DIE DATENSCHUTZKONFORME SEKUNDÄRNUTZUNG STRUKTURIERTER UND FREITEXTLICHER DATEN

Griebel L[1], Leb I[1], Christoph J[1], Laufer J[2], Marquardt K[2], Prokosch HU[1], Toddenroth D[1], Sedlmayr M[1]

Kurzfassung

Bislang konnte eine Vielzahl medizinischer Daten nicht automatisiert für weitere Analysen verwendet werden, da die Informationen in Freitextform (z.B. in Arztbriefen oder OP-Berichten) vorlagen. Im Projekt „cloud4health" werden große medizinische Rohdatenbestände erschlossen, indem Data-Warehouse- und Data-Mining-Technologien mit Werkzeugen der Textanalyse kombiniert werden. Eine spezialisierte Cloud-Infrastruktur stellt dazu bedarfsgerecht und datenschutzkonform erforderliche Ressourcen wie Rechenleistung oder Speicherkapazitäten bereit.

Abstract

So far, medical data can often not be used for further analysis, since the information is in free-text form (e.g. in discharge letters or surgical reports). The project "cloud4health" will unlock large medical raw data files. The approach combines data warehousing and data mining technologies with tools of textual analysis. A specialized cloud infrastructure provides on-demand resources such as processing power and storage capacity in a data protection conformant environment.

Keywords – Cloud-Computing, Systemarchitektur, Sekundärnutzung, NLP, Datenschutz

1. Einleitung

Durch die wachsende Digitalisierung der Medizin fallen im klinischen Umfeld heutzutage enorme Datenmengen an. Diese Daten unterstützen den Behandlungsprozess oder den organisatorischen Rahmen im Behandlungsumfeld. Zunehmend werden Routinedaten aber auch außerhalb ihrer ursprünglichen Zweckbestimmung verwendet, beispielsweise zum Qualitätsmanagement und insbesondere in der klinischen Forschung [5,7,8].

Derartige Routinedaten können dabei in strukturierter als auch in unstrukturierter Form vorliegen. Vielfach sind für die Forschung interessante Angaben zu Symptomen oder Behandlungen nur als Freitext beispielsweise in Arztbriefen, OP-Berichten oder Pathologieberichten vorhanden. Natural Language Processing (NLP) kann solche Textinformationen semantisch erschließen. Die großen

[1] Lehrstuhl für Medizinische Informatik, FAU, Erlangen, Deutschland
[2] RHÖN-KLINIKUM AG, Bad Neustadt/Saale, Deutschland

Datenmengen, welche dabei analysiert werden müssen, erfordern große Computerressourcen [2], die aber nur zeitweise für eine konkrete Analyse notwendig sind.

Genau dafür bietet sich Cloud-Computing an [9]. Cloud-Computing ist ein vielgestaltiges Modell, in dem der Nutzer auf einen Pool konfigurierbarer, potentiell unendlicher Rechnerressourcen innerhalb der eigenen Organisation (private Cloud) und/oder außerhalb (public Cloud) zugreifen kann. Eine Cloud kann dabei sowohl virtuelle Hardware zur Verfügung stellen (IaaS: Infrastructure as a Service), aber auch vorinstallierte Dienste, beispielsweise Analysewerkzeuge für klinische Forscher (SaaS: Software as a Service) anbieten. Eine besondere Herausforderung für die Nutzung externer Clouds ist allerdings die besondere Schutzwürdigkeit medizinischer Daten [4].

Ein flexibler, cloudbasierter Lösungsansatz zur datenschutzkonformen Erschließung sowohl strukturierter als auch unstrukturierter Daten in medizinischen Freitexten wird seit November 2011 im Projekt cloud4health (c4h) realisiert. Die Partner Averbis GmbH, RHÖN-KLINIKUM AG, Fraunhofer SCAI, TMF e.V. und die Universität Erlangen-Nürnberg kombinieren Data-Warehouse- und Data-Mining-Technologien mit Werkzeugen der Textanalyse um große medizinische Rohdatenbestände zu erschließen und für Studien in einem zentralen Studienportal zur Verfügung zu stellen.

2. Methoden

Für die Umsetzung des Projektes sollte eine Architektur gefunden werden, die auf existierenden Erfahrungen bezüglich Sekundärnutzung und Data-Warehousing aufbaut und flexibel bezüglich eingesetzter Technologien und der länderabhängigen Datenschutzbestimmungen ist.

Aufgrund der komplexen Rahmenbedingungen (z.B. Datenverfügbarkeit, Datenschutz, eingesetzte Technologien) wurde mit SCRUM eine agile Vorgehensmethode gewählt. In achtwöchigen Sprints und wöchentlichen Diskussionsrunden wurden zunächst vier Forschungsfragestellungen in Form von Anwendungsszenarien ausformuliert, die später als Prototypen auf Basis der cloud4health Infrastruktur implementiert werden. Für diese Szenarien wurden die Stakeholder identifiziert und deren Aufgaben organisatorischer, technischer, rechtlicher und medizinischer Art beschrieben.

Die Auswahl der Werkzeuge und Schnittstellen wurde unter Berücksichtigung freier Verfügbarkeit (z.B. Talend Open Studio, i2b2) und Standardisierung (HL7, ODM) getroffen.

3. Ergebnisse

Mit dem Architekturentwurf (*Abbildung 1*) wurde ein strukturelles Rahmenwerk für das System cloud4health erstellt. Dabei gibt es drei wesentliche Bereiche:
Lokale Services erschließen und deidentifizieren die strukturierten und freitextlichen Rohdaten bei jedem Datenlieferanten vor Ort (Extraction-Transfer-Loading-Prozess). Die Klinik verlassen damit nur Daten, die durch das Entfernen identifizierender Textstellen in den Freitexten bzw. das Anonymisieren der Metadaten keine Rückschlüsse auf den Patienten ermöglichen.

Um die Sekundärnutzung sowohl strukturierter als auch unstrukturierter Daten zu ermöglichen, werden letztere durch Annotationskomponenten in einer *Text-Mining-Cloud* aufbereitet. Hier werden Freitexte studienspezifisch und für jeden Datenlieferanten instanziiert und annotiert, bevor

die strukturierten Ergebnisse an den Datenlieferanten zurückgegeben werden. Da durch die vorherige Deidentifizierung die Schutzwürdigkeit der Daten berücksichtigt wird, ist es möglich, die Cloud sowohl privat als auch öffentlich zu betreiben. Eine spezielle Infrastruktur (Trusted Cloud) gewährleistet durch Maßnahmen, wie z.B. strengen Zugriffsbeschränkungen, die Wahrung der Sicherheit bei der Datenverarbeitung.

Die Bereitstellung der so aufbereiteten Daten zu weiterführenden Zwecken erfolgt im *zentralen cloud4health-Studienportal.* Hier befinden sich die von mehreren Lieferanten zusammengeführten anonymisierten Daten. Strukturierte Daten werden dabei direkt nach ihrer Deidentifizierung an das Studienportal übermittelt, während unstrukturierte Freitexte erst in der oben erklärten Textmining-Cloud aufbereitet werden müssen.

Um eine datenschutzkonforme Sekundärnutzung klinischer Routinedaten zu ermöglichen, werden die Deidentizierungsverfahren ausschließlich lokal nach folgendem Ablauf durchgeführt:

- Für eine klinische Studie notwendige Patienten werden in den teilnehmenden Kliniken ausgewählt (*Patient-Selector*) und lokal die dazu gehörigen Elemente und Dokumente aus den klinischen Systemen extrahiert (*Data-Collector*). Diese Daten werden das Krankenhaus nicht verlassen, bevor sie nicht um alle identifizierenden Merkmale bereinigt wurden.
- Dazu werden in einem ersten Schritt alle identifizierenden Textpassagen und Attribute (z.B. Namen, Datums- und Adressangaben =Personal Health Information/PHI) durch XML-Tags markiert (*PHI-Tagger*) und ersetzt (*Nonstructured PHI-Replacer*). Eine weitere Komponente (*IDAT-Translator*) liefert auf der Basis definierter Regeln einzusetzende Merkmalsausprägungen, die identifizierende Daten in den Dokumenten ersetzen.
- Um die Effizienz dieses Vorgangs zu erhöhen, werden die extrahierten Dokumente zuvor vom *Wrapper* dahingehend überprüft, ob sie durch vorhergehende Studien bereits anonymisiert vorliegen.
- Sofern die Daten strukturiert vorliegen, erhält der *Local Mapper* die Datensätze, um krankenhausspezifische Elemente auf standardisierte Terminologien (z.B. Laborwerte auf LOINC) automatisiert zu mappen.
- Im Anschluss daran wird der *Structured PHI-Replacer* aktiv, durch den die Datensätze und entsprechende Elemente, bei Bedarf nach Rückfrage an den *IDAT-Translator*, mit den einzusetzenden Merkmalsausprägungen ersetzt werden.
- Der *Metadata-Replacer* prüft anschließend die Headerfelder und filtert/löscht oder anonymisiert/pseudonymisiert personifizierte HL7-Felder. Sofern notwendig kommuniziert er mit dem *IDAT-Translator* um entsprechende Merkmalausprägungen auszutauschen.
- Dann werden die Daten in der *Transferdatenbank* gesichert und an den *Tempifier* zur Vorbereitung der Übertragung an die *Textmining-Cloud* weitergegeben.

Das Textmining zum Zwecke der Erschließung freitextlicher Dokumente erfolgt in dieser speziellen Textmining-Cloud. Ein *Tempifier* erzeugt dafür pro Dokument eine eindeutige temporäre Dokumenten-ID, die nur während der Bearbeitung in der Cloud gültig ist (doppelte Pseudonymisierung). Durch diese Sicherheitsmaßnahmen entsteht eine besondere Trusted Cloud-Struktur.

Wurden die Dokumente in der Textmining-Cloud erschlossen, gelangen sie als ODM-Dokument zurück an die *Transferdatenbank*. Das Ergebnis des studienspezifischen Mappings wird als

Ergebnis in der *Transferdatenbank* vorgehalten und kann nun, ggf. nach einer *K-Anonymisierung*, an das *c4h-Studienportal* übertragen werden.

In diesem Portal werden deidentifizierte Daten aus strukturieren und unstrukturierten klinischen Dokumenten verschiedener Kliniken für spezifizierte Nutzer für bestimmte Usecases zur Verfügung stehen. Ein *Study-Mapper* bereitet die Dokumente für die Übertragung vor. Angereichert kann die Bereitstellung von Daten außerdem um weitere Services (z.B. Reporting Engine), die in einer privaten Cloud bereitgestellt werden.

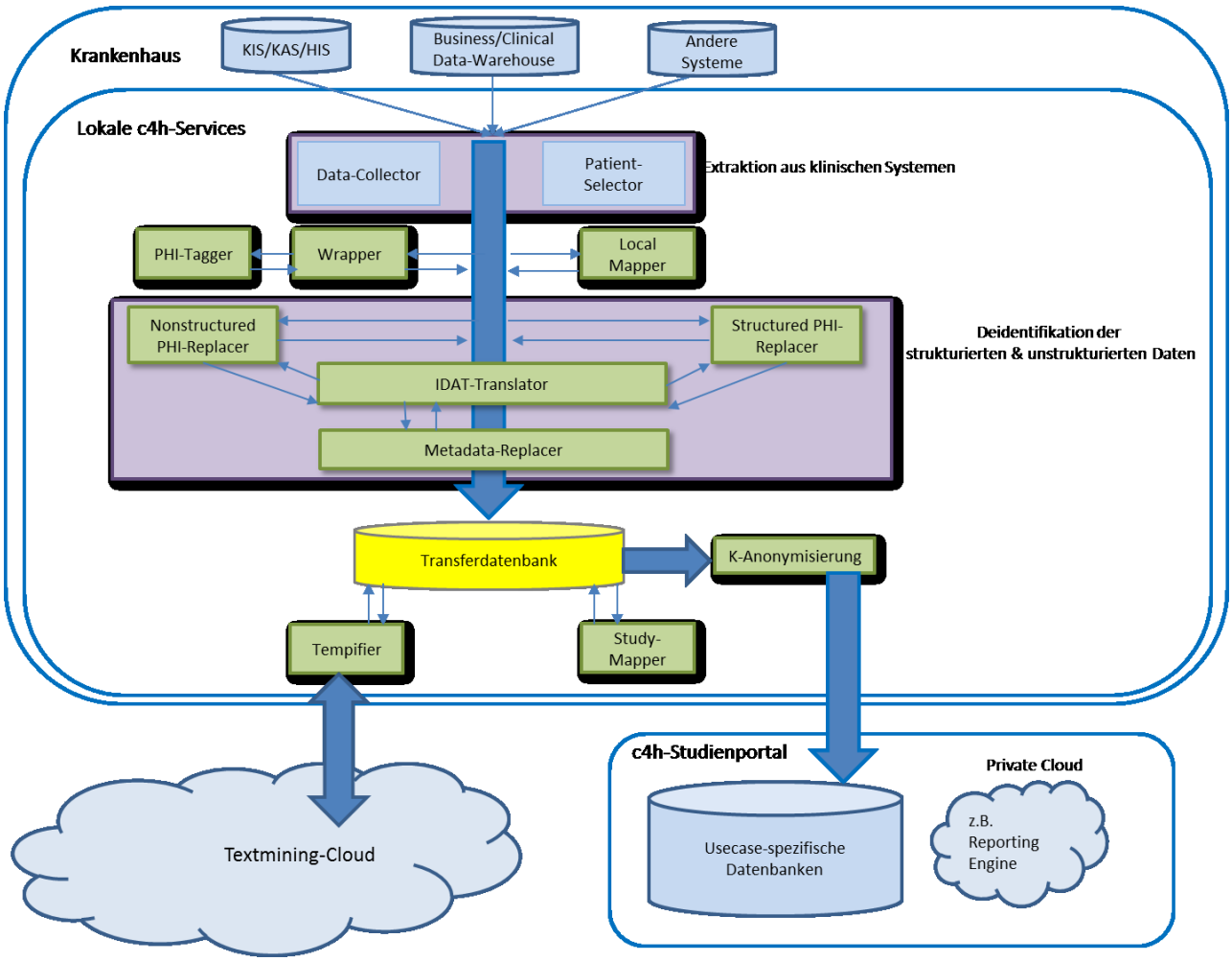

Abbildung 1: Architektur-Rahmenwerk

Die *Flexibilität bezüglich der Skalierung* äußert sich in der Möglichkeit, die drei Bereiche lokale Services, Textminingservices der Trusted Cloud und des Studienportals sowohl in privaten als auch öffentlichen Clouds zu betreiben. So können beispielsweise Klinikkonzerne eigene Textminingclouds betreiben während kleinere Kliniken externe Dienstleister nutzen können. Neben der Verlagerung eines ganzen Bereichs ist auch die Nutzung von Teildienste möglich, wie z. B. ein Terminologiemanagement oder Datenaufbereitung im Studienportal (Reporting als SaaS).

Die *Flexibilität bezüglich des Datenschutzes* zeigt sich vor allem durch drei Ausbaustufen, die unterschiedliche Schutzbedürfnisse implementieren [4]:

- 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 eigenen Patienten zu reidentifizieren. Vor dem Export in eine zentrale Studiendatenbank werden die Patienten jedoch wie in Stufe 1 k-anonymisiert.
- Stufe 3: Klinikübergreifende Pseudonymisierung erlaubt das Zusammenführen von Daten eines Patienten aus mehreren Kliniken (record linkage), ist datenschutzrechlich jedoch besonders kritisch.

Durch diese Flexibilität und die Modularisierung der Komponenten ermöglicht cloud4health auch die Kooperation mit anderen Systemen zur Sekundärnutzung und die Anpassung an andere Datenschutzbestimmungen beispielsweise in anderen Ländern.

Bis November 2012 erfolgte die Entwicklung der einzelnen Komponenten, die bis März 2013 zu einem vollständigen Prototyp kombiniert werden. Der erste Anwendungsfall realisiert eine Fragestellung des deutschen Endoprothesenregisters: Auf Basis der Auswertung von OP-Berichten und Arztbriefen soll die Frage nach besten OP-Techniken für den Hüftprotheseneinsatz (z.B. mit oder ohne Zement) beantwortet werden. Bisher wurden 550 Arztbriefe und mehr als 580 OP-Berichte von ca. 250 Patienten annotiert und durch oben beschriebenen Ablauf um identifizierende Elemente bereinigt, womit die Funktionsweise der datenschutzkonformen Aufbereitung unstrukturierter klinischer Daten verifiziert wurde. Weitere Anwendungsfälle beziehen sich auf die Erschließung freitextlicher Pathologieberichten (580.000 Freitexte liegen vor), eine Plausibilitätsprüfung von Abrechnungen bei Krankenkassen sowie auf Fragestellungen der Pharmakovigilanz.

4. Diskussion

Die Sekundärnutzung klinischer Daten zu Forschungszwecken ist auch das Ziel anderer Projekte, die jeweils unterschiedliche Ziele adressieren. Beispielsweise sind dies die Identifikation von Patientenkollektiven auf Basis unstrukturierter Daten [3, 11], Cloud-Technologien [2] und Zugang über webbasierte Frontends [6]. Keines der Projekte verbindet jedoch die Komponenten zu einer gemeinsamen Architektur wie cloud4health.

Zudem fokussiert cloud4health auf das deutsche Gesundheitssystem und dessen spezifische Gesetze und Richtlinien [1], denn länderspezifische Anforderungen wie bspw. der Health Insurance Portability and Accountability Act (HIPAA) [10] schränken die Übertragbarkeit ausländischer Lösungen stark ein.

Cloud4health verbindet die automatisierte Nutzung von strukturierten und unstrukturierten Daten und Cloud Computing zu einer flexiblen und datenschutzkonforme Infrastruktur zur institutionsübergreifenden Sekundärnutzung medizinischer Daten.

5. Danksagung

Das Projekt cloud4health wird vom Bundesministerium für Wirtschaft und Technologie im Rahmen des Schwerpunkts Trusted Cloud gefördert (FKZ 01MD11009).

6. Referenzen

[1] Bundesministerium für Gesundheit. Gesundheitssystem. Online verfügbar unter http://www.bmg.bund.de/gesundheitssystem.html, zuletzt geprüft am 17.01.2013.

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

[3] EHR4CR-Konsortium. Electronic Health Records for Clinical Research. Online verfügbar unter http://www.ehr4cr.eu, zuletzt geprüft am 16.01.2013.

[4] Glock J., Herold R., Pommerening K. Personen-Identifikatoren in medizinischen Forschungsnetzen: Evaluation des Personen-Identifikator-Generators im Kompetenznetz Pädiatrische Onkologie und Hämatologie. GMS Med Inform Biom Epidemiol 2006;2(2):Doc06.

[5] 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.

[6] Hurdle JF, Haroldsen SC, Hammer A, Spigle C, Fraser AM, Mineau GP, et al. Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database. Journal of the American Medical Informatics Association : JAMIA. 2013;20(1):164-71. Epub 2012/10/13.

[7] 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.

[8] Li Z, Wen J, Zhang X, Wu C, Li C, Li Z, Liu L. ClinData Express - A Metadata Driven Clinical Research Data Management System for Secondary Use of Clinical Data. AMIA Annu Symp Proc, 2012: 552-7.

[9] Mell P, Grance, T. 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.

[10] 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 17.01.2013.

[11] Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. Journal of the American Medical Informatics Association. 2009;16(5):624-30.

Corresponding Author
Martin Sedlmayr
Lehrstuhl für Medizinische Informatik
Friedrich-Alexander Universität Erlangen-Nürnberg
Krankenhausstrasse 12
D-91054 Erlangen
Email: martin.sedlmayr@imi.med.uni-erlangen.de