datenschutzkonforme Sekundärnutzung strukturierter und freitextlicher daten Mittels Cloud-Architektur

Ines Leb1, Lena Griebel1, Jan Christoph[[1]](#footnote-1), Julian Laufer[[2]](#footnote-2), Kurt Marquardt2, Hans-Ulrich Prokosch1, Dennis Toddenroth1, Martin Sedlmayr1

# Einleitung

Im medizinischen Umfeld werden zunehmend Daten digital erfasst und gespeichert. Diese sollen neben der klinischen Routine auch für die klinische Forschung genutzt werden [1, 2, 3].

Dabei liegen die für Forschungszwecke benötigten Daten, meist in unstrukturierter Form wie in Arztbriefen vor. Natural Language Processing (NLP) ermöglicht solche Textinformationen semantisch zu erschließen. Um die großen dafür notwendigen Computerressourcen zur Verfügung zu stellen, bietet sich Cloud-Computing [4] an, das dem Nutzer konfigurierbare und potentiell unendliche Ressourcen bereitstellt. Dabei muss stets der Datenschutz für die Sekundärnutzung gewährleistet sein [5].

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. Besonders herausfordernd ist hierbei, Data-Warehouse und Data-Mining-Technologien mit Textanalyse-Werkzeugen datenschutzkonform in der Cloud zu kombinieren, um große medizinische Rohdatenbestände zu erschließen und für Studien in einem zentralen Studienportal zur Verfügung zu stellen [6].

# Methoden

Mit SCRUM [7] 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.

Die cloud4health-Architektur ermöglicht, dass verschiedene Datenlieferanten ihre strukturierten und unstrukturierten Daten in eine zentrale Datenbank liefern. Dafür konnte auf bereits existierende Erfahrungen bezüglich Sekundärnutzung und Data-Warehousing zurückgegriffen werden (z.B. [1, 4, 8]). Um das System für die Datenlieferanten kostengünstig anzubieten, wurden für die Realisierung der Architektur frei verfügbare Werkzeuge und Standards verwendet.

# Ergebnisse

Die Architektur für cloud4health wurde in drei Bereiche untergliedert:

* Lokale Services: Erschließung und Deidentifizierung der strukturierten und freitextlichen Rohdaten bei jedem Datenlieferanten vor Ort (ETL-Prozess)
* Trusted Cloud: Annotation von Freitexten, geschützter Raum für Datenlieferanten als Arbeitsumgebung für studienspezifische Instanziierung, Textmining und Rückgabe strukturierter Ergebnisse an Lieferanten
* Cloud4health-Studienportal: Zusammenführung der Freitexte mehrerer Datenlieferanten in einer geschützten Public Cloud und intelligenter Raum für die Bereitstellung dieser Daten zur Auswertung als ganzheitliche Betrachtung

Die cloud4health-Architektur ermöglicht eine Flexibilität bezüglich der Skalierbarkeit und des Datenschutzes. So können die einzelnen Bereiche je nach Bedarf auch in einer privaten oder einen öffentlichen Cloud betrieben werden. Weiterhin können die Daten pseudonymisiert und/oder anonymisiert werden, um die jeweiligen Datenschutzbestimmungen zu erfüllen [5]. Durch diese Flexibilität sind auch Kooperationen mit anderen Systemen zur Sekundärnutzung möglich.

Bis März 2013 wurde ein erster vollständiger Prototyp entwickelt, der die Infrastruktur für die Realisierung des ersten Anwendungsfalls liefert. Dabei wird die Qualität von Hüftprothesenimplantationen (z.B. mit oder ohne Zement) 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.

# Diskussion

Die Sekundärnutzung klinischer Daten zu Forschungszwecken ist auch Ziel anderer Projekte. Dabei untersuchen diese die Identifikation von Patientenkollektiven auf Basis unstrukturierter Daten [9], Cloud-Technologien [4] und Zugang über webbasierte Frontends [10]. Keines der Projekte verbindet jedoch die Komponenten zu einer gemeinsamen Architektur wie cloud4health.

Durch den ersten Prototyp konnte demonstriert werden, dass das entwickelte System es ermöglicht strukturierte und unstrukturierte Daten aus unterschiedlichen Datenquellen gemäß den Datenschutzbestimmungen automatisch in ein zentrales Studienportal zu überführen. In den nächsten Schritten muss nun das System verfeinert und um weitere geplante Komponenten erweitert werden. Zur Evaluation des Systems sind weitere Anwendungsfälle vorgesehen.

# Danksagung

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

# 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, 0:1-5.

[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, 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.

[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] 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.

[6] cloud4health. Online verfügbar unter: <http://cloud4health.de/>, zuletzt geprüft am 05.03.2013.

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

[8] 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.

[9] 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.

[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 2012, 00:1-8.

1. Lehrstuhl für Medizinische Informatik, FAU, Erlangen [↑](#footnote-ref-1)
2. RHÖN-KLINIKUM AG, Bad Neustadt/Saale [↑](#footnote-ref-2)