Integrated Data Repository Toolkit: Werkzeuge zur Nachnutzung medizinischer Daten für die Forschung

Ganslandt T[1], Sax U[2], Löbe M[3], Drepper J[4], Bauer C[2], Baum B[2], Christoph J[5], Mate S[5], Quade M[2], Stäubert S[3], Prokosch HU[5]

[1] Medizinisches Zentrum für Information & Kommunikation, Uniklinik Erlangen
[2] Abteilung Medizinische Informatik, Universitätsmedizin Göttingen
[3] Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig
[4] TMF e.V., Berlin
[5] Lehrstuhl für Medizinische Informatik, Universität Erlangen-Nürnberg
thomas.ganslandt@uk-erlangen.de
ulrich.sax@med.uni-goettingen.de

Abstract: Medizinische Daten werden in Forschung und Routineversorgung zunehmend elektronisch erhoben und gespeichert. Hierdurch ergeben sich erhebliche Nachnutzungspotentiale z.B. im Sinne von Feasibility-Analysen oder einer Rekrutierungsunterstützung. Die notwendige Aufbereitung der Rohdaten und Bereitstellung in geeigneten Auswertungswerkzeugen stellen jedoch eine Hürde für die Realisierung dieser Potentiale dar. Im Rahmen des beschriebenen Projekts wurden zunächst Anforderungen und Anwendungsszenarien für die Nachnutzung medizinischer Daten erhoben. Anschließend wurde das Integrated Data Repository Toolkit auf Basis der Open Source-Komponenten i2b2 und Talend Open Studio konzipiert und implementiert. Das Toolkit unterstützt Forschungsverbünde bei der Inbetriebnahme der i2b2-Plattform sowie der Aufbereitung medizinischer Rohdaten durch einen flexibel anpassbaren ETL-Prozess (Extraction, Transformation & Loading).

1 Einleitung

Der zunehmende Einsatz von Electronic Data Capture-Systemen (EDC) in der medizinischen Forschung sowie der Ausbau klinischer Informationssysteme in der Routineversorgung führen zu einem stetig wachsenden Pool elektronisch verfügbarer medizinischer Daten. Obwohl die Erhebung der Informationen primär für einen bestimmten Zweck erfolgt (z.B. Wirksamkeitsnachweis eines Medikaments, Leistungsabrechnung), bieten die so entstandenen Datensammlungen erhebliche Potentiale für eine weitergehende Nachnutzung [PG09]. Aus den Routinedaten einer Klinik können beispielsweise Feasibility-Analysen zur Durchführbarkeit und Kollektivgröße einer zukünftigen Studie durchgeführt oder der Rekrutierungsprozess einer laufenden Studie unterstützt werden. Aus den in einer Forschungsdatenbank gesammelten Erhebungsbögen einer abgeschlossenen Studie können neue Hypothesen generiert oder Kollektive für Substudien ermittelt werden, die über das ursprüngliche

Studienziel hinausgehen. Die beschriebenen sowie weitere Anwendungsszenarien werden unter dem Begriff "Secondary Use" bzw. "Single-Source-Ansatz" zusammengefasst.

Um die gewünschte Nachnutzung zu ermöglichen, müssen die vorhandenen Daten jedoch zunächst in eine zweckgeeignete Form gebracht und in entsprechende Werkzeuge z.B. zur weitergehenden Auswertung importiert werden. Diese Verarbeitungsschritte werden unter dem Begriff Extraction, Transformation & Loading (ETL) zusammengefasst. Daten der klinischen Routine werden häufig mit Hilfe verteilter Informationssysteme erhoben, die sowohl aus unternehmensweit eingesetzten klinischen Arbeitsplatzsystemen als auch aus spezialisierten Abteilungssystemen bestehen können. Daten können zwar aus jedem einzelnen System exportiert werden, stehen dann häufig jedoch in unterschiedlichen Formaten und abweichender Codierung der erfassten Merkmale zur Verfügung. Informationen für klinische Studien werden zwar häufig mit einem übergreifenden EDC-System erfasst, können jedoch auch mit zusätzlichen Datenpools z.B. aus Laboranalysen oder daraus abgeleiteten Werten flankiert werden. Neben den Aspekten der Zusammenführung und semantischen Integration der heterogenen Quelldatensätze müssen auch die Anforderungen des Datenschutzes z.B. im Sinne der Anonymisierung oder Pseudonymisierung von Datensätzen beachtet werden. Zur Verwertung der aufbereiteten Daten sind wiederum Plattformen zur intuitiven Abfrage und Analyse erforderlich. Obwohl Werkzeuge für viele der beschriebenen Teilaspekte existieren, fehlt eine integrierte Lösung, die den Anwender über den gesamten Prozess der Datenaufbereitung unterstützt. Dieser Teilschritt ist daher häufig ressourcenaufwändig und stellt eine Hürde bei der Realisierung des Mehrwerts im Single-Source-Ansatz dar.

Die TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF)[1] unterstützt Forschungseinrichtungen und -verbünde durch die Erarbeitung von Konzepten und IT-Werkzeugen über den gesamten Lebenszyklus der medizinischen Forschung. Im Kontext dieses Artikels sind insbesondere Angebote zum Datenschutz (generische Datenschutzkonzepte [Re06][He10], PID-Generator [FP05], Pseudonymisierungsdienst [PS06]), das IT-Strategieprojekt sowie die Evaluation der KIS-basierten Rekrutierungsunterstützung erwähnenswert. Ziel des vorliegenden, von der TMF geförderten Projekts ist es, eine integrierte Plattform zur Datenaufbereitung für die Nachnutzung medizinischer Daten zu etablieren. In diesem Rahmen sollen zunächst prototypische Anwendungsfälle mit ihren Anforderungen erhoben werden. Im Anschluss soll eine flexible und erweiterungsfähige Plattform zur Datenaufbereitung zunächst konzipiert und dann implementiert werden.

2 Methoden

Auf einem vorbereitenden Workshop mit Vertretern interessierter Forschungsverbünde wurden am 29.10.2010 prototypische Anwendungsszenarien erhoben sowie relevante IT-Plattformen vorgestellt. Auf einem weiteren Workshop nach Projektbeginn am

[1] http://www.tmf-ev.de/ (abgerufen 13.05.2012)

27.02.2012 wurde den Verbünden das geplante Umsetzungskonzept präsentiert und auf dieser Basis die Anforderungserhebung aktualisiert. Anschließend wurde mit der Implementierung der Integrationsplattform auf Basis der ausgewählten Open Source-Komponenten begonnen.

3 Ergebnisse

Der angenommene Bedarf sowohl in Bezug auf die Nachnutzung medizinischer Daten an sich als auch in Bezug auf Werkzeuge zu deren Integration konnte im Rahmen der Workshops bestätigt werden. Als relevante Zielgruppen wurden Forschungsverbünde identifiziert, die im Rahmen ihrer Vorhaben Datenbestände aufgebaut haben bzw. momentan aufbauen, sowie Versorgungseinrichtungen, die mit Hilfe von klinischen Routinedaten Forschungsprojekte durchführen wollen. Mehrwerte können hierbei durch eine standardisierte und datenschutzkonforme Aufbereitung vorhandener Datenbestände für Forschungszwecke realisiert werden.

Die folgenden Zielgruppen und Anwendungsfälle wurden im Rahmen der Workshops diskutiert:

- Für multizentrische, dezentrale Forschungsnetze (z.B. Kompetenznetze in der Medizin oder Netzwerke seltener Erkrankungen): Integration verschiedener Studiendatenbanken; Metaanalysen; Langzeitverfügbarkeit nach Auslaufen der Förderung

- Für Klinische Studienzentren / Site Management Organisationen: Unterstützung der Patientenrekrutierung, Abschätzung von Feasibility-Anfragen aus der Pharmaforschung; Durchführung von Follow-Up-Studien

- Für Register und Kohorten: Verknüpfung mit externen Partnern, die erweiterte, z.B. soziodemografische, lebensstil-assoziierte, genomische oder umweltbezogene Daten erfassen; Observationsstudien

- Für klinische Forschergruppen: Zusammenführung der verteilten lokalen Datenbestände aus der patientennahen Forschung bzw. der jeweiligen Routineversorgung an den Studienpatienten

- Für Integrierte Forschungs- und Behandlungszentren: Verwendung von Daten und Diensten der Patientenversorgung für die Forschung (Single Source); Patiententagging

Auf dieser Basis wurden Datenflussdiagramme für drei prototypische Anwendungsszenarien ausgearbeitet (Abbildung 1).

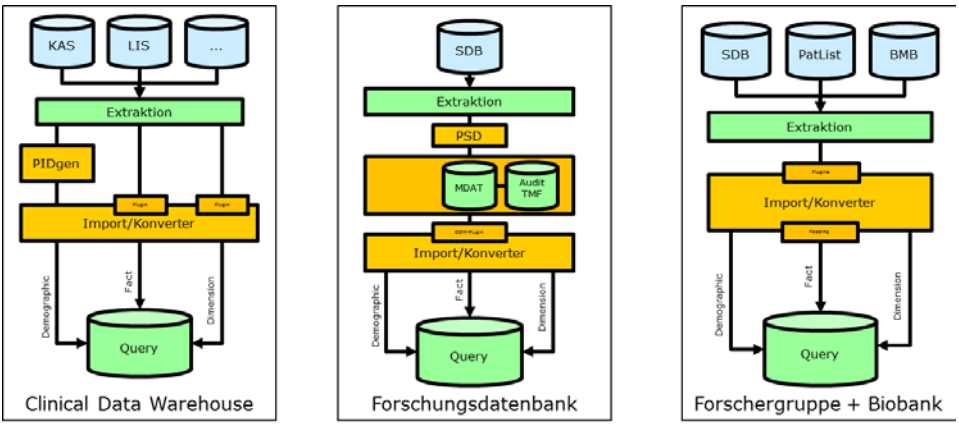

Abbildung 1: Datenflussdiagramme für die Anwendungsszenarien "Clinical Data Warehouse", "Forschungsdatenbank" und "Forschergruppe + Biobank" (KAS = Klinisches Arbeitsplatzsystem, LIS = Laborinformationssystem, PIDgen = PID-Generator, SDB = Studiendatenbank, PSD = Pseudonymisierungsdienst, MDAT = Medizinische Daten; TMF = Trial Master File, PatList = Patientenliste, BMB = Biomaterialbank)

Im Rahmen des ersten Workshops und im Vorfeld der Projektkonzeption wurden Evaluationsergebnisse des TMF-IT-Strategieprojekts und Anwendererfahrungen zu den internationalen Software- und Infrastrukturentwicklungen caBIG [EB07], OpEN.SC [Ha09] und i2b2 [Mu10][KCM12] präsentiert und diskutiert: caBIG (Cancer Bioinformatics Grid) verfolgte den umfassendsten Ansatz mit einer Vielzahl von Einzelkomponenten u.a. zur Verwaltung eines kontrollierten Vokabulars und von Bioproben sowie für High-Performance-Computing, setzte jedoch für den Betrieb die Etablierung einer umfangreichen Infrastruktur voraus, die für kleinere Forschungsverbünde schwer aufzubauen und zu betreiben wäre. OpEN.SC (Open European Nephrology Science Center) befand sich zum Zeitpunkt des Workshops noch in der Phase einer lokalen Entwicklung an der Charité. i2b2 (Informatics for Integrating Biology and the Bedside) stellte flexible und intuitiv nutzbare Funktionen zur Abfrage klinischer Datenbestände zur Verfügung, war über ein Plugin-Konzept erweiterbar und verfügt über eine große internationale Anwendercommunity, die in der i2b2 Academic User Group[2] organisiert ist. i2b2 wurde daraufhin im Rahmen eines Vorprojekts in Bezug auf die Verwendung in den identifizierten Anwendungsszenarien evaluiert [Ga10]. Die Anwendbarkeit für die Szenarien konnte hierbei bestätigt werden. Als Einschränkungen wurden jedoch der komplexe Installationsprozess sowie fehlende Werkzeuge für die Aufbereitung und den Import von Quelldaten identifiziert.

Ausgehend von den Ergebnissen der Workshops wurde die IDRT-Plattform (Integrated Data Repository Toolkit) konzipiert, um Open Source Werkzeuge zur Unterstützung der Nachnutzung medizinischer Daten auf Basis von i2b2 zu entwickeln und bereitzustellen:

[2] https://www.i2b2.org/work/aug.html (abgerufen 13.05.2012)

- Setup- und Konfigurationswizard zur Vereinfachung von Installation und Betrieb der i2b2-Plattform

- Entwicklung einer auf i2b2 zugeschnittenen ETL-Plattform auf Basis des Open Source Produkts Talend Open Studio[3]

- Bereitstellung von Import-Schnittstellen für etablierte Datenformate aus der medizinischen Forschung und Routineversorgung bzw. generische Datenformate: CDISC ODM[4], Paragraph 21-Benchmarking-Datensatz[5], tabellarische Datenquellen (CSV, SQL)

- Aufbereitung häufig verwendeter Standardterminologien für die Nutzung in i2b2: z.B. ICD-Diagnosekatalog, Operationen- und Prozedurenschlüssel (OPS). Laboruntersuchungen (LOINC) u.a.

- Einbindung des PID-Generators der TMF zur einstufigen Pseudonymisierung sowie Unterstützung von Record-Linkage-Anforderungen

Der Setup- und Konfigurationswizard wurde als bash[6]-Shellscript implementiert und ist unter Ubuntu-Linux[7] einsatzfähig. Er unterstützt den Anwender über den gesamten Installationsprozess vom Download der nötigen Pakete über den Einspielprozess bis zur Konfiguration und zum Start einer lauffähigen i2b2-Instanz. Der Wizard stellt darüber hinaus Funktionen zur Verwaltung von i2b2-Projekten und Nutzern sowie zum Einspielen von Demodaten bereit. Der Wizard kann über die TMF-Homepage[8] als ausführbares Script sowie in einer vorbereiteten virtuellen Maschine heruntergeladen werden.

Der ETL-Prozess unter Talend Open Studio wurde als mehrstufige Pipeline konzipiert, in der Rohdaten zunächst über datentypspezifische Plugins in einen Staging-Bereich importiert werden (Abbildung 2). Für den Staging-Bereich wird hierbei ein normales i2b2-Datenbankschema verwendet, so dass die weiteren Verarbeitungsschritte auch auf Quelldaten angewendet werden können, die auf anderem Weg importiert wurden. Die ETL-Plattform unterstützt zurzeit die Importformate ODM 1.3 und Paragraph 21 sowie generische Datenquellen über CSV- und SQL-Quellen. Sie wird als Talend Open Studio-Projekt auf der TMF-Homepage[8] zum Download bereitgestellt.

Standardterminologien können über entsprechend angepasste Import-Plugins für die ETL-Plattform aufbereitet werden und unterstützen das Rohdatenformat der jeweiligen offiziellen Quelle. Die Terminologie-Plugins werden als Bestandteil der ETL-Plattform bereitgestellt. Die Terminologie-Rohdaten sind nicht Bestandteil der Distribution und müssen von den jeweiligen offiziellen Quellen heruntergeladen sowie ggf. individuell lizensiert werden.

[3] http://www.talend.com/products/open-studio-di.php (abgerufen 13.05.2012)
[4] http://www.cdisc.org/odm (abgerufen 13.05.2012)
[5] http://www.g-drg.de/cms/Datenlieferung_gem._21_KHEntgG (abgerufen 13.05.2012)
[6] http://tiswww.case.edu/php/chet/bash/bashtop.html (abgerufen 13.05.2012)
[7] http://www.ubuntu.com/ (abgerufen 13.05.2012)
[8] http://www.tmf-ev.de/idrt (abgerufen 13.05.2012)

Die Integration des TMF PID-Generators befindet sich z.Zt. noch in der Umsetzung.

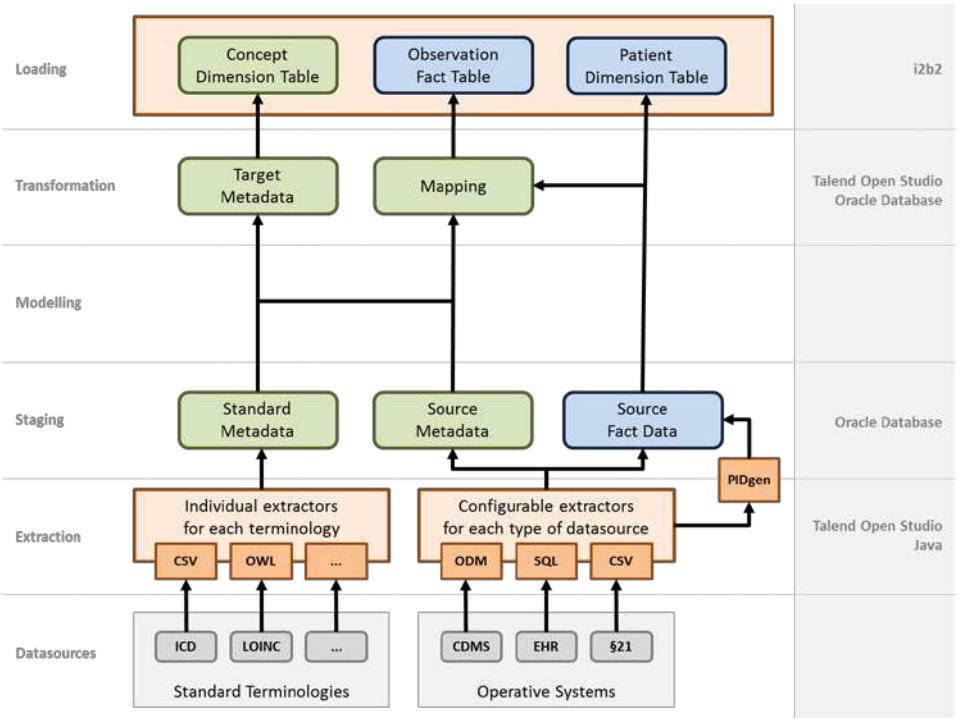

Abbildung 2: ETL-Prozess im beschriebenen Ansatz: die verschiedenen Datenquellen (unten) werden über Talend Open Studio erschlossen und in einen Staging-Bereich übertragen. Vor dort aus erfolgt nach Transformation und ggf. Mapping von Ontologie- bzw. Nutzdaten das Laden in das Abfrage- und Filterwerkzeug i2b2.

4 Diskussion

Das gewählte Plattformkonzept stellt durch Auswahl von Open Source-Komponenten (i2b2, Talend Open Studio) sicher, dass die Ergebnisse des Projekts ohne Lizensierungskosten genutzt werden können. Beide Plattformen verfügen über eine große, international aktive Anwendercommunity, so dass ihr langfristiger Fortbestand gesichert erscheint. Der Setup- und Konfigurationswizard vereinfacht und verkürzt die Installation erheblich und trägt damit zur Akzeptanzverbesserung und weiteren Verbreitung der i2b2-Plattform bei. Die Implementierung einer generischen ETL-Struktur mit datentypspezifische Plugins für den Import erlaubt die flexible Erweiterung der Plattform um zusätzliche Datenformate sowie die einfache Anpassung an zukünftige Änderungen bei den bereits unterstützten Formaten. Die Bereitstellung von Import-Plugins für verbreitete Standardterminologien reduziert wiederum den Aufwand für die produktive Nutzung von i2b2 in den verschiedenen Forschungsverbünden.

Die Nutzung der Plattform setzt jedoch weiterhin eine intensive Beschäftigung mit den zu integrierenden medizinischen Rohdaten voraus. Trotz der Bereitstellung von Import-Plugins ist weiterhin IT-Fachwissen sowie die Einarbeitung in die Talend Open Studio-Plattform erforderlich, um die nötigen Parametrierungsschritte umsetzen zu können. Durch die begleitende Dokumentation sowie die Bereitstellung einer integrierten Umgebung wird die Hürde für den produktiven Einsatz jedoch erheblich herabgesetzt.

5 Schlussfolgerungen und Ausblick

Im Rahmen des Projekts wurde eine integrierte Plattform zur Unterstützung Aufbereitung medizinischer Rohdaten für ihre Nachnutzung entwickelt. Das IDR-Toolkit nutzt etablierte Open Source-Komponenten und ergänzt sie durch gezielte Hilfsmittel über den gesamten Prozess von der Installation bis zur Datenaufbereitung. Die Einstiegshürde zu ihrer produktiven Nutzung und damit der Realisierung von Nachnutzungspotentialen wird dadurch abgesenkt.

Der flexible Ansatz erlaubt zukünftige Erweiterungen des IDR-Toolkits. Neben der Erschließung weiterer Quelldatentypen durch geeignete Plugins sind die Integration des TMF-Pseudonymisierungsdienstes sowie eine Anbindung des TMF Metadatenrepositories (MDR[9]) relevante Erweiterungsmöglichkeiten. Der Ausbau von i2b2 zur SHRINE-Architektur [We09] für Abfragen über verteilte Instanzen erweitert die Einsatzmöglichkeiten der Plattform erheblich. Das IDR-Toolkit könnte hierbei zukünftig Funktionen zur semantischen Integration der verteilten i2b2-Instanzen beitragen.

Danksagung

Das vorliegende Projekt wurde von der TMF unter dem Projektkennzeichen V091-MI aus Mitteln des BMBF-Projekts MethInfraNet (Förderkennzeichen 01GI1003) gefördert.

Literaturverzeichnis

[EB07] Eschenbach, A.C.; Buetow, K.: Cancer informatics vision: caBIG. Cancer Inform. 2007 ;2:22-4

[FP05] Faldum, A.; Pommerening, K.: An optimal code for patient identifiers. Comput Methods Programs Biomed. 2005;79(1):81-8

[Ga10] Ganslandt, T. et al.: Unlocking Data for Clinical Research – The German i2b2 Experience. Appl Clin Inform. 2010;2(1);116-127

[KCM12] Kohane, I.S.; Churchill, S.E.; Murphy, S.N.: A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc. 2012 Mar-Apr;19(2):181-5

[9] http://www.tmf-ev.de/mdr (abgerufen 13.05.2012)

[Ha09] Hanss, S. et al.: Integration of decentralized clinical data in a data warehouse: a service-oriented design and realization. Methods Inf Med. 2009;48(5):414-8

[He10] Helbing, K. et al.: A data protection scheme for medical research networks. Review after five years of operation. Methods Inf Med. 2010;49(6):601-7

[Mu10] Murphy, S.N. et al.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-30

[PG09] Prokosch, H.U.; Ganslandt, T.: Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48(1):38-44

[PS06] Pommerening, K., M. Schröder, et al.: Pseudonymization Service and Data Custodians in Medical Research Networks and Biobanks. Informatik 2006 - Informatik für den Menschen. Beiträge der 36. Jahrestagung der Gesellschaft für Informatik e.V. Bonn, Gesellschaft für Informatik. 2006; P-93: 715-721.

[Re06] Reng, C.M. et al.: Generische Lösungen zum Datenschutz für die Forschungsnetze in der Medizin. Mwv Medizinisch Wissenschaftliche Verlagsgesellschaft, Berlin, 1. Auflage 2006, ISBN 978-3939069041

[We09] Weber, G.M. et al.: The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624-30