

Die grafische Benutzeroberfläche PaDaWaN für das klinische Data Warehouse für Routinedaten an der Universitätsklinik Würzburg

Georg Fette^{1,2}, Georg Dietrich², Maximilian Ertl^{1,2}, Martin Toepfer², Mathias Kaspar¹, Stefan Störk¹, Frank Puppe²

¹ Comprehensive Heart Failure Center,
97078 Würzburg
{georg.fette}@uni-wuerzburg.de,
{ertl_m, stoerk_s, kaspar_m}@ukw.de

² Universität Würzburg, Lehrstuhl Informatik VI,
97074 Würzburg
{dietrich, toepfer, puppe}@uni-wuerzburg.de

1 Einleitung

Das klinische Data Warehouse dient dazu, die heterogen und nur teilweise strukturiert im klinischen Informationssystem der Universität Würzburg gespeicherten Daten aus der Diagnose und Therapie von Patienten für Information Retrieval und statistische Auswertungen verfügbar zu machen. Die Daten werden dafür in eine homogenen Datenstruktur transformiert und über die grafische Benutzeroberfläche PaDaWaN (Patienten Data Warehouse Navigator) den Benutzern am Klinikum zur Verfügung gestellt. Es gibt zwei Ebenen an Benutzerrechten: Eine niedrige Stufe erlaubt Häufigkeiten von Patientengruppenprofilen abzufragen. Eine höhere Berechtigungsstufe erlaubt die Suche nach individuellen, jedoch pseudonymisierten, Patientendaten. Die Zuordnung von Pseudonymen zu Patienten-IDs wird von einer dritten Stelle verwaltet, an die Anträge zur Depseudonymisierung gestellt werden können. Dadurch können mit dem Data Warehouse z.B. Studienpatienten rekrutiert werden.

Derzeit (31.3.2015) sind im Data Warehouse die Domänen Stammdaten, ICD10-kodierte Diagnosen, Laborwerte, Prozeduren, Arztbriefe, Echobefunde, Sonographiebefunde, Medikationen, Radiologiebefunde, Herzkatheterbefunde, Anamnesebefunde und Patientenbewegungen enthalten und können über den PaDaWaN abgefragt werden. Auf den textuellen Domänen wie z.B. aus Arztbriefen oder Echobefunde werden mithilfe von Informationsextraktionsmethoden [1, 2] zusätzliche strukturierte Daten gewonnen. Das Data Warehouse umfasst insgesamt einen Datenbestand von 280 Millionen Fakten, die aus 4.4 Millionen Fälle von 1 Millionen Patienten stammen.

2 Methoden

Klinischen Routinedaten werden aus den verschiedenen Quellsystemen durch automatische Services täglich exportiert, in eine SQL-Datenbank importiert und anschließend von einem Lucene-Index (lucene.apache.org) indiziert. Nutzer im Klinikum Würzburg können an diesen Index mit der grafischen Benutzeroberfläche PaDaWaN Suchanfragen stellen.

PaDaWaN besteht hauptsächlich aus drei grafischen Elementen: Ein Fenster stellt den hierarchischen Katalog dar, aus dem der Benutzer Attribute durch Browsen oder Suchen für eine gewünschte Suchanfrage auswählen kann. Aus dem Katalog werden Attribute in das Suchfenster gezogen und dort Werte und andere Konfigurationen der eigentlichen Suchanfrage angegeben. Im dritten Fenster wird das Ergebnis angezeigt. Konfigurationsmöglichkeiten im Suchfenster umfassen z.B. Einschränkungen auf Attributen (z.B. „<“, „>“ auf Zahlen- oder enthaltene Strings in Textattributen) oder boolesche Verknüpfungen von Attributen. Suchen können gespeichert und später wieder editiert werden oder auch als Makro in andere Suchen integriert werden. Die Art der Ergebnisse im Ergebnisfenster wird durch zwei auswählbare Suchmodi bestimmt:

Mit der „individuellen Patientensuche“ ist das Ergebnis eine Tabelle mit einer Zeile pro Fall bzw pro Patient und den Daten der angewählten Attribute in den Spalten. Mit der entsprechenden Genehmigung können die Patienten-IDs in diesen Daten depseudonymisiert werden, um z.B. die Patientenrekrutierung für Studien zu unterstützen.

Mit der „statistischen Suche“ ist das Ergebnis eine Tabelle die nur Patienten-Zahlen der gefundenen Mengen angibt. Attribute können als reiner Filter definiert werden oder Zahlen- oder Single-Choice-Attribute können so konfiguriert werden, dass ihre Ausprägungen (vordefinierte Intervallgrenzen, bzw. Ausprägungen von Single-Choice-Attributen) weitere Spalten und Zeilen der Ergebnistabelle ergeben. Zellen die unter einer k-Anonymitätsschwelle liegen werden leer gelassen.

3 Ergebnisse

Bei der Unterstützung von Studien mit dem Data Warehouse hat sich herausgestellt, dass für eine gegebene Suchanforderung die Auswahl der korrekten Attribute aus dem Katalog und deren geschickte Parametrisierung eine herausfordernde Aufgabe darstellt. Durch die schnellen Antwortzeiten des Lucene-Indexes dauert eine durchschnittlich komplexe Suche, die nur die Anzahl an Treffern zurückliefert, unter 100ms. Dadurch ist es möglich in einer explorativen Art und Weise die gewünschte Suchanfrage zusammenzustellen und somit schneller zur gewünschten Suche und ihrem Ergebnis zu kommen. Dies ist mit dem PaDaWaN besser möglich als mit ähnlichen Data Warehouse Tools wie z.B. I2B2 (www.i2b2.org), da dort die Suchen alleinig mit SQL Statements durchgeführt werden die bei vielen Data-Warehouse-typischen Anfragetypen durchschnittlich um ein bis zwei Größenordnungen langsamer sind als Anfragen über einen Lucene-Index [3]. Mit PaDaWaN kann, beginnend mit der statistischen Suche, ein schneller Überblick über gewünschte Patientenkollektive gewonnen werden. Mit der individuellen Patientensuche können die zurückgegebenen Ergebnisse zu Beginn der Sucherstellung auf kleine Stichprobenmengen eingeschränkt werden, so dass auch hier die Antwortzeiten sehr kurz bleiben. Lediglich komplexe Suchanfragen mit vielen Schlüsselwörtern und regulären Ausdrücken auf Textdokumenten können im Minutenbereich liegen.

4 Diskussion

Da die Funktionalität und das Design des PaDaWaNs bei der Unterstützung von Studien im Klinikum Würzburg stetig weiter entwickelt werden, hat sich die Oberfläche zu einem sehr nützlichen Werkzeug entwickelt, mit dem bereits viele Studien unterstützt wurden.

Diese Arbeit wurde unterstützt durch die Förderung des Bundesministeriums für Bildung und Forschung (BMBF01 EO1004)

5 Referenzen

[1] Toepfer M., Beck PD, Dietrich G., Ertl M., Fette G., Kluegl P., Störk S., Puppe F.: Extraktion kodierter Daten aus textuellen Befundberichten: Eine Fallstudie zu Echokardiographieberichten; GMDS14, 2014

[2] Fette G., Ertl M., Wörner A., Kluegl P., Störk S., Puppe F.: Information Extraction from Unstructured Electronic Health Records and Integration into a Data Warehouse;Lecture Notes in Informatics (LNI) 2012

[3] Dietrich G., Fette G., Puppe F.: A Comparison of Search Engine Technologies for a Clinical Data Warehouse; Proceedings LWA2014