Fallstudie zur Validierung eines klinischen Data-Warehouse mit Hintergrundwissen

Georg Dietrich[1], Georg Fette[1,2], Maximilian Ertl[1,2], Martin Toepfer[1], Mathias Kaspar[2], Stefan Störk[2], Frank Puppe[1]

[1] Universität Würzburg, Institut für Informatik
97074 Würzburg
{georg.dietrich, martin.toepfer, frank.puppe}@uni-wuerzburg.de

[2] Universität Würzburg, Deutsches Zentrum für Herzinsuffizienz,
97078 Würzburg
{ertl_m, stoerk_s, kaspar_m}@ukw.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.
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 (Patienten Data Warehouse Navigator) [1] abgefragt werden. Auf den textuellen Domänen wie z.B. aus Arztbriefen oder Echobefunde werden mithilfe von Informationsextraktionsmethoden [2, 3] 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

Die Validierung der Datenqualität ist ein sehr wichtiger und zentraler Punkt bei der Erstellung eines Data-Warehouses. Ein Prozess für eine gute Datenqualität umfasst Schritte wie Datenbereinigung, Datenbearbeitung, Datenfilterung, und Qualitätstests. Qualitätstests können in zwei Kategorien eingeteilt werden: in Syntax- und Semantiktests. Während bei der Syntaxprüfung z. B. die Korrektheit der Datentypen, die Einhaltung von Zeichenmustern und die Verwendung unerlaubter Zeichen betrachtet werden, überprüfen Semantiktests zum einen die Integrität der Daten zu dem Model und zum anderen die inhaltlichen Beziehungen dieser Daten.

Die inhaltlichen Beziehungen der Daten kann als Hintergrundwissen in einem medizinischen wissensbasierten System hinterlegt werden [4]. Diese umfassen Beziehungen wie Symptom-Diagnose, Diagnose-Therapie und Diagnose-Untersuchung. Ein Beispiel sind Beziehungen zwischen Laborwerten und Diagnosen, da bei vielen Diagnosen bestimmte Laborwerte erhöht oder erniedrigt sind. Allerdings hängen die Grenzwerte häufig von Alter und Geschlecht der Patienten ab. Die zu erwartenden Laborwerte bei einer Diagnose werden auch durch den Schweregrad der Diagnose und durch die aktuelle Therapie beeinflusst. Daher verfolgen wir einen Ansatz, der Beziehungen differenziert modelliert. Dafür eignet sich eine regelbasierte Wissensbasis, deren Objekte auf die Objekte des Data-Warehouses abgebildet werden.

Als Fallstudie haben wir bekannte Beziehungen zwischen Laborwerten und Diagnosen gewählt, insbesondere die Beziehung zwischen einer chronischen Nierenerkrankung (ICD: N18, sowie deren Verfeinerungen) und einem erhöhten Kreatinin-Wert im Blut.

3 Ergebnisse

Von den knapp 900 000 Patientenfällen der Universitätsklinik, in denen der Creatinin-Wert bestimmt wurde, war er in ca. 24% der Fälle erhöht (bei einem Grenzwert von >= 1,1 mg/dl). Bei der Diagnose N18 mit ca. 80 000 Patientenfällen hatten ca. 77% einen erhöhten Creatinin-Wert. Für die Validierung ist die Aufschlüsselung in die Untergruppen von N18 und deren Anteil mit erhöhtem Creatinin-Wert interessant: N18.0 (terminale Niereninsuffizienz): 99%, N18.1 und N18.2: Chronische Nierenerkrankung Stadium 1 oder 2: 40%, N18.3 (Stadium 3): 89%, 18.4 und N18.5. (Stadium 4 oder 5): 99,6%, N18.81 und N18.82 (chronische Niereninsuffizienz, Stadium 1 oder 2): 69%, N18.83 (Stadium 3): 94%, N18.84 (Stadium 4): 98%. Insbesondere bei der chronischen Niereninsuffizienz in Stadium 3 oder 4 wäre ein Wert von 100% zu erwarten gewesen. Wir haben daher die zugehörigen Arztbriefe stichprobenartig überprüft: Es scheint tatsächlich in den meisten Fällen mit N18.84 oder N18.83 bei normalem Creatinin-Wert Fehler bei der Kodierung zu geben, da teilweise eine akute Niereninsuffizienz vorlag (ICD Code N17), aber im Arztbrief keine Hinweise auf Niereninsuffizienz enthalten waren.

4 Diskussion

Die Fallstudie zeigt, dass die Methode der Konsistenzüberprüfung von Diagnosen und Laborwerten das Potential hat, Inkonsistenzen bei der Dokumentation medizinischer Befunde zu finden. Um das systematisch durchführen zu können, müssen außer den ICD-Codes der Diagnosen auch die Extraktion von Diagnosen aus Arztbriefen hinzugezogen werden, da diese häufig zuverlässiger sind als die primär für die Abrechnung verwendeten ICD-Codes. Weiterhin sollten auch weitere Patientenparameter wie z.B. Alter, Geschlecht und Medikation zur Bewertung der Konsistenten genutzt werden, was über die Anbindung an eine medizinische Wissensbasis umgesetzt werden soll.

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

5 Referenzen

[1] Dietrich G, Fette G, Beck P, Ertl M, Toepfer M, Kluegl P et alAnfrage-spezifische Validierung in einem Data Warehouse für klinische Routinedaten an der Universitätsklinik Würzburg. [Internet] 2014 [cited 2015 March 25]; doi: 10.3205/14gmds07.
[2] Toepfer M, Fette G, Beck PD, Klügl P, Puppe F. Integrated Tools for Query-driven Development of Light-weight Ontologies and Information Extraction Components. In: Ide N, Grivolla J, editors. Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT; 2014 Aug 23; Dublin, Ireland. Association for Computational Linguistics and Dublin City University; 2014. p. 83-92.
[3] 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. In: Goltz U, Magnor M, Appelrath HJ, Matthies HK, Balke WT, Wolf L, editors. Lecture Notes in Informatics 208: Proceedings of Informatik 2012; 2012 Sep 16-21; Braunschweig, Germany; Bonn: Gesellschaft für Informatik; 2012. p. 1237-1251.
[4] Puppe, F. Medizinische Entscheidungsunterstützungssystem. Informatik Spektrum, 2014; 37/3: 246-249.