Exposé zur Diplomarbeit

# Integration klinischer Befunddaten unter besonderer Berücksichtigung der Diagnose-Klassifikation

Johannes Starlinger

Januar 2010

Betreuer: Prof. Ulf Leser, Dr. Bernd Schmeck

## Einleitung

Während eines Krankenhausaufenthaltes werden oft verschiedene diagnostische Verfahren herangezogen, um zu ermitteln, an genau welcher Krankheit ein Patient leidet. Die Ergebnisse jeder durchgeführten Untersuchung werden dabei typischerweise zunächst separat erfasst und liegen anschließend dem Arzt zur summarischen Beurteilung vor – jeweils für den einzelnen Patienten. Zum Zweck der klinischen Forschung und zur Qualitätssicherung ist es notwendig, die Untersuchungsdaten vieler Patienten zu betrachten. So interessieren etwa die Korrelation von Verdachtsdiagnosen, die aufgrund von bestimmten Untersuchungen gestellt wurden, mit der endgültigen Diagnose oder auch ein mögliches Vorhandensein von Unterschieden in der Befundung ähnlicher Datenlagen.

Die Beantwortung dieser Fragestellungen wird dadurch behindert, dass im klinischen Alltag Untersuchungsergebnisse und andere Patientendaten oft in unterschiedlicher elektronischer Form festgehalten und abgelegt werden, etwa als Textverarbeitungsdokumente oder in Datenbanken. Als Konsequenz sind diese physisch getrennten und strukturell heterogenen Daten einer gemeinsamen Analyse zunächst entzogen. Selbst Zusammenhänge zwischen Merkmalen nur eines Krankheitsverlaufes, die in verschiedenen Untersuchungen erfasst wurden, lassen sich nicht oder nur mit erheblichem manuellen Aufwand erforschen. Im Sinne patientenübergreifender statistischer Erhebungen ist dies besonders ungünstig.

Für eine effiziente Verarbeitung der erhobenen Daten ist es somit wünschenswert, diese in eine gemeinsame Datenbasis zu überführen. Dieser Vorgang der

Integration von Informationen aus verschiedenen Quellen bringt eine Reihen von Schwierigkeiten mit sich [1]. So müssen die Daten zunächst aus den einzelnen Quellen und deren spezifischen Formaten eingelesen und in ein gemeinsames Format und Schema transformiert werden. Dieses gemeinsame Zielschema muss derart ausgelegt sein, dass es – unter Berücksichtigung der jeweiligen Anwendungsaufgabe – eine möglichst gute Repräsentation der Ursprungsdaten erlaubt. Des weiteren ist erforderlich, in Beziehung stehende Wertpaare aus verschiedenen Quellen zu identifizieren (Mapping), um auch eine semantische Integration der Daten zu erlauben. So können etwa mehrere Datensätze desselben Patienten nur durch Zuordnung der patientenbezogenen Daten miteinander in Verbindung gebracht werden. Eine weitere besondere Schwierigkeit bei der Integration von klinischen Daten ist oftmals die eindeutige Identifikation (Normalisierung) der (Verdachts-)Diagnosen, die von den befundenen Ärzten als freier Text formuliert sind, sofern nicht bereits manuell eine Kodierung vorgenommen wurde. Erst nach dieser Normalisierung ist das erstellen aussagekräftiger Statistiken über Diagnosen möglich.

## Zielsetzung

Das grundlegende Ziel der Arbeit ist das Ermöglichen von statistischen Untersuchungen auf Daten verschiedener Quellen an der Medizinischen Klinik mit Schwerpunkt Infektiologie und Pneumologie der Charité Berlin. Speziell die Beantwortung datenquellenübergreifender Fragestellungen, wie dem erwähnten Vergleich der in den BAL-Befunden gestellten Verdachtsdiagnosen mit den endgültigen Diagnosen der Arztbriefe, steht dabei im Mittelpunkt des Interesses. Dazu soll in dieser Diplomarbeit anhand des realen Beispielszenarios an der Charité ein System entwickelt werden, welches mehrere Quellen klinischer Daten in einer gemeinsamen Datenbank zusammenführt. Der Fokus liegt dabei auf dem genannten Problem der Normalisierung von Diagnosen. Dieses spezielle Problem ist nicht nur im Rahmen der Informationsintegration von Interesse, sondern auch für die Kodierung klinischer Diagnosen nach den Schemata der International Classification of Diseases (ICD10) [2] und der Diagnosis Related Groups (DRG) [3], die nicht zuletzt für die Abrechnung von Leistungen im medizinischen Bereich von Bedeutung sind, relevant. Diese Kodierung wird gegenwärtig manuell durchgeführt, was nicht nur einen erheblichen Kostenfaktor darstellt, sondern auch – in erster Linie aufgrund der Komplexität der genannten Klassifikationsschemata – eine nicht unerhebliche Rate an Fehlkodierungen mit sich bringt [4]. Es konnte gezeigt werden, dass automatische Systeme eine mindestens ebenso hohe Kodierungsgüte erzielen können wie menschliche Annotatoren [5]. Bei der Computational Medicine Challenge 2007 [6], mit der Aufgabenstellung der Diagnoseklassifikation von medizinischem Freitext wurden F-Scores von 89% erzielt.

Die bisher auf diesem Gebiet entwickelten Systeme verarbeiten zum allergrößten Teil englischsprachigen Text und beziehen sich auf den ICD9 Standard. Die im Rahmen dieser Arbeit zu untersuchende Normalisierung und Kodierung erfolgt auf deutschsprachigen Diagnosetexten nach ICD10.

## Vorgehen

Ziel der Integration ist die automatisierte Zusammenführung der Daten in einer gemeinsamen Datenbank. Um die leichte Erweiterbarkeit der Menge der Datenquellen zu gewährleisten, soll hierzu eine modulare Integrationsplatform entwickelt werden, in die neue Quellmodule einfach eingebunden werden können. Hierbei wird ein einfacher Ansatz verfolgt: Jede Quelle (bzw jedes Quellmodul) wird unter Definition eines eigenen Teilschemas direkt in die Datenbank abgebildet. Dies ist nicht zuletzt deshalb wünschenswert, weil einzelne Datenquellen in der Zieldatenbank weiterhin identifizierbar bleiben sollen.

Als primäre Eingangsdaten dienen zunächst zwei Quellen: Befunde der Bronchio-alveolären Lavage (BAL) und die korrespondierenden Arztbriefe. Beide liegen in Form von Word-Dokumenten vor. Erstere enthalten ein Formular, in dem die Untersuchungsergebnisse erfasst sind und zudem die Befunddiagnose in Form eines Freitextes festgehalten ist. Die Arztbriefe sind als semi-strukturierter Freitext verfasst. Weitere Datenquellen sind vorgesehen - so die Einbeziehung der SAP-Datenbank der Charité mit weiteren diagnostischen Ergebnissen.

Um die Datensätze aus den einzelnen Quellen zusammenzuführen, muss zunächst eine Identifizierung der Datensätze über die personenbezogen Patientendaten (Name, Geburtsdatum, Identifikationsnummer) erfolgen. Vor Eintrag in die Datenbank sollen diese Daten aus Gründen des Datenschutzes allerdings entfernt werden, so dass aus der endgültigen Datenbank keine personenbezogen Daten gewonnen werden können. In welchem Umfang diese (Pseudo-)Anonymisierung stattzufinden hat, bleibt zu klären.

Wie bereits erwähnt, ist die entscheidende Herausforderung die Extraktion von Diagnosen aus den Freitexten der Eingangsdokumente (insbesondere der BAL Befunde) und deren Normalisierung auf die Kodierungsschemata der International Classification of Diseases [2] und der Diagnosis Related Groups [3]. Support Vector Machines und Bayes Klassifikatoren haben sich als Verfahren des maschinellen Lernens bei vergleichbaren Aufgaben als zuverlässige Wahl erwiesen [7]. Gleichzeitig zeigen die Resultate von [6], dass Systeme mit regelbasierten Komponenten [8, 9] bei der automatischen Diagnosekodierung besonders gut abschneiden. Die genannten Verfahren sollen, einzeln und in Kombination, auf die vorliegenden deutschspachigen Texte (die Befundungs- und Beurteilungsfelder der BAL-Befunde und die Diagnosetexte der Arztbriefe) angewendet und verglichen werden. Je nach Ergebnis erfolgen anschließend weitere Optimierungen. Die Eingangsdaten werden in Trainings- und Testsätze geteilt. Die Annotation der Trainingdatensätze muss manuell erfolgen, wozu gegebenenfalls ein Hilfswerkzeug zu entwickeln ist.

Das Hinzufügen neuer Datensätze soll regelmäßig und automatisch erfolgen. Je nach Grad der vorgenommenen (Pseudo-)Anonymisierung kann dies durch Überwachung der Quellen in Echtzeit zum Zeitpunkt der Verfügbarkeit neuer Daten erfolgen, oder intervallbasiert. Ersteres ist möglich, wenn eine Identifikation der in der Datenbank vorliegenden Datensätze anhand der personenbezogen

Merkmale neuer Datensätze möglich bleibt (ohne jedoch eine rückgerichtete Identifikation zu erlauben). Ist dies nicht der Fall, müssen die Quellen als ganzes erfasst werden, um die vollständige Zuordnung neuer Datensätze zu bereits vorhandenen zu gewährleisten. Ob dies in Echtzeit sinnvoll ist, bleibt zu prüfen.

## Referenzen


[1] P. Ziegler, K. R. Dittrich,
*Three Decades of Data Integration — all Problems Solved?,*
IFIP Congress Topical Sessions, 2004, Springer

[2] International Classification of Diseases
http://www.who.int/classifications/icd/en/

[3] Diagnosis Related Groups
http://www.g-drg.de

[4] C. Benesch, D. W. Jr, A. Wilder, P. Duncan, G. Samsa, D. Matchar,
*Inaccuracy of the international classification of diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease,*
Neurology, pages 660–664, 1997

[5] A.F. Sonel, C.B. Good, H. Rao, A. Macioce, L.J. Wall, R.S., Niculescu, S. Sandilya, P. Giang, S. Krishnan, P. Aloni, R.B. Rao
*Use of remind artificial intelligence software for rapid assessment of adherence to disease specific management guidelines in acute coronary syndromes,*
AHRQ, 2006

[6] Computational Medicine Center, Medical NLP Challenge, 2007
http://www.computationalmedicine.org/challenge/index.php

[7] L. V. Lita, S. Yu, S. Niculescu, J. Bi,
*Large Scale Code Classification for Medical Patient Records*,
Proceedings of the Third International Joint Conference on Natural Language Processing, 2008

[8] R. Farkas, G. Szarvas
*Automatic construction of rule-based ICD-9-CM coding systems.*
BMC Bioinformatics, 9S3, S10. 2008

[9] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar,
*Automatic code assignment to medical text.*
Proceedings of ACL’07 BioNLP workshop, 2007