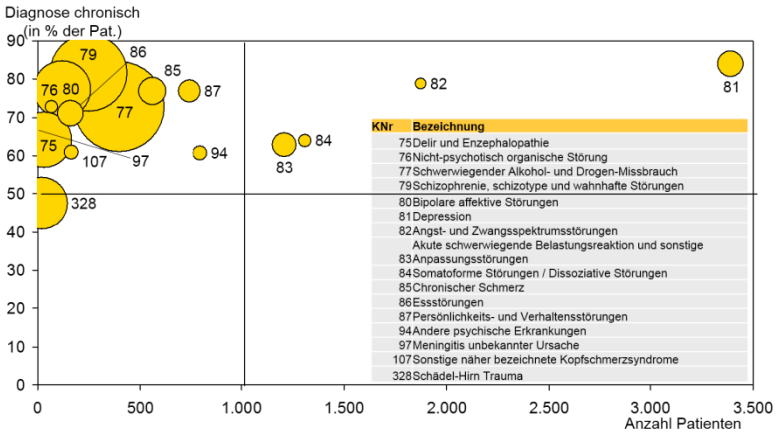
Bachelorprojekt für 2014/2015
Fachgebiet Informationssysteme
Prof. Dr. Felix Naumann

Big Data Analytics für Gesundheitsdaten

Die Nutzung statistischer Analysen von Daten im Gesundheitswesen hat in den letzten Jahren stark zugenommen. Die Erhebung und Speicherung der Daten erfolgt dabei aktuell noch stark fragmentiert – Ärzte, Krankenhäuser, Apotheken etc. betreiben eigene Systeme. Die Daten aus allen Teil-Systemen laufen aber bei den gesetzlichen Krankenkassen für die Abrechnung zusammen und stellen aktuell die einzige Möglichkeit zur breiten Analyse medizinischer Versorgung dar.

Die Analyse dieser „GKV-Daten“ kann dazu dienen, die medizinische Versorgung der Patienten in Deutschland zu verbessern, zum Beispiel durch

- Identifizierung von medizinischer Unter- oder Überversorgung, oder Schnittstellenproblemen (z.B. zwischen Hausarzt, Facharzt, und Krankenhaus).
- Untersuchung der Effektivität von Behandlungen und Therapien, zum Beispiel die Frage, ob Chirotherapie („Knacken und Drücken“) für die Behandlung von Rückenschmerz geeignet ist.
- Identifizierung von Patienten mit hohem Risiko, an bestimmten Krankheiten zu erkranken.
- Analyse von Krankheitsverläufen und –schweregraden



Projektbeschreibung

Elsevier hat zur statistischen Analyse der GKV-Daten den Elsevier Versorgungsmanagement Assistent (EVA) entwickelt. Die EVA-Software ermöglicht Krankenkassen, selbst die Konzeption, Evaluation und Optimierung der medizinischen Versorgung durchzuführen, ohne Kenntnis von Datenabfragesprachen (SQL) und ohne Kenntnis statistischer Methoden.

Für die nächste Generation der EVA – der Version 4.x – wird der Technology Stack komplett erneuert. Um die höchstmögliche Performance zu erzielen, sollen leistungsfähigste Open Source Systeme eingesetzt werden:

a. Das Clustersystem „High Performance Computing Cluster“ (HPCC)
b. Pentaho Data Integration
c. Mondrian OLAP Server
d. Saiku Analytics
e. Eine selbstentwickelte GUI auf Grundlage von Java und gwt

Das Projekt wird den Fokus auf zwei wesentliche Punkte legen:

1. **Integration und Visualisierung**: Die verwendeten Systeme sollen unter ein gemeinsames User Interface zusammengefügt werden. Das Interface soll so einfach gehalten sein, dass den Business Anwendern ein Data Mining ohne SQL- oder Statistik-Kenntnisse ermöglicht wird. Und es soll die State-of-the-Art graphischen Darstellungen integrieren, die Saiku Analytics mitbringt. Zusätzlich soll eine zentrale Ablage und Verwaltung von Datenabfragen und Nutzer-definierten Methoden über die Systeme hinweg konzipiert und entwickelt werden.

2. **Optimierung**: Ziel ist, auch für große Datenmengen (z.B. 7 Millionen Patienten mit 1,2 Milliarden Arztbesuchen in den letzten 7 Jahren) eine fast „real-time" Analyse zu ermöglichen. Um dies zu erreichen, müssen alle Teilsysteme optimiert werden:
   a. Interaktive Visualisierung mit Saiku Analytics
   b. ETL Funktionalität für EVA in Pentaho Data Integration
   c. Optimierter Zugriff auf das HPCC Clustersystem via JDBC
   d. (Automatische) Skalierung des HPCC Clustersystems
   e. Optimierung der Multidimensional Expressions (MDX) Anfragen in Mondrian OLAP

Die fachliche Domäne „Gesundheitswesen" und die sehr strukturiert vorliegenden Daten ermöglichen eine starke Optimierung. Das Projekt wird fachlich von Elsevier Health Analytics in Berlin unterstützt, und technisch von bakdata. Es werden ein vollständiger, großer Entwicklungsdatensatz und eine zentrale Cloud-Infrastruktur auf Amazon Webservices zur Verfügung gestellt.

Um eine gemeinsame und flexible Konkretisierung der zu erreichenden Ziele zu ermöglichen, ist ein agiles Vorgehen wünschenswert. Als Vorgehensmodell wird Scrum empfohlen und unterstützt.

Projektpartner

Elsevier Health Analytics ist führend auf dem Gebiet Predictive Analytics & Data Mining auf deutschen Gesundheitsdaten. Wir identifizieren prospektiv Hochrisikopatienten und unterstützen Ärzte bei der Umsetzung von evidenzbasierten Best Practices, Leitlinien und Versorgungsprogrammen. Unser Erfolg stammt aus der engen Zusammenarbeit von Analytikern, Medizinern, Statistikern, Informatikern und Gesundheitsökonomen.

Wir gehören zur **Reed Elsevier Gruppe** mit weltweit ca. 32.000 Mitarbeitern und einem Gesamtumsatz von rund 7,5 Milliarden Euro. Wir haben die HPCC Super-Computer Plattform entwickelt (jetzt open source). HPCC ermöglicht es, sehr große Datenmengen mit hoher Geschwindigkeit auf verteilten Systemen parallel zu verarbeiten. Seit 1992 verbinden wir Data Mining auf Routine-Gesundheitsdaten mit fortgeschrittener Analytik für US-amerikanische Krankenversicherungen und Leistungserbringer.

Das Projekt für 6-8 Studenten beginnt am 1. Oktober 2014 und wird durch Prof. Dr. Felix Naumann, Dr. Ralf Krestel und Toni Grütze betreut. Fragen können gerne an felix.naumann@hpi.uni-potsdam.de gerichtet werden.