INTEGRATION BIOMEDIZINISCHER DATEN IN EINER KLINISCHEN BIOINFORMATIK PLATTFORM

Pfeifer B[1], Baumgartner C[1], Plant C[1], Modre R[2], Schreier G[2], Tilg B[1]

Kurzfassung

In diesem Beitrag wird die klinische Bioinformatik Plattform zur Speicherung und Integration von biomedizinischen Datenbeständen dargestellt, um systembiologische Ansätze vorantreiben, Therapien verbessern und Medikamentenentwicklung unterstützen zu können. Speziell bei interdisziplinären Aufgabestellungen ist eine Plattform, die diverse Institutionen miteinander einheitlich verbindet, unerlässlich, um bestmögliche Daten-qualität, sowie Sicherheit gewährleisten zu können.

1. Einleitung

Speziell in interdisziplinären Forschungsfeldern, wie der biomedizinischen Forschung, wird eine Plattform benötigt, mit der systembiologische Ansätze untersucht werden können. Bei einer systembiologischen Untersuchung wird der Organismus eines Individuums als ein integriertes und interagierendes Netzwerk von Genen, Proteinen und Reaktionen angesehen. Um Krankheiten besser verstehen zu können, um neue biologische Marker finden, um biostatistische Analysen durchführen zu können, wird eine klinische Bioinformatik Plattform (CBP) zur strukturierten und qualitätsgesicherten Speicherung von biomedizinischen Daten benötigt. Mittels einer solchen Plattform werden sowohl klinische, Hochdurchsatzdaten (-omics) sowie Literaturdaten und externe Biodatenbanken qualitätsgesichert in einem Repository gehalten.

Die schnell wachsende Zahl an internen und externen biomedizinischen Datenquellen stellt dabei eine große Herausforderung bei der Integration der Daten in ein klinisches Data Warehouse dar. Besonders die Verknüpfung von klinischen, experimentellen Daten (-omics) und Literaturdaten ist eine Notwendigkeit, um Data Mining und Wissensmanagement betreiben zu können. Hinzu kommt, dass biomedizinisches Wissen aufgrund mangelnder Strukturierung in Texten natürlichsprachlich kodiert ist. Auch die geringe Vernetzung von Patientendaten, Literaturdaten und Hochdurchsatzdaten verlangen nach einer stärkeren Verknüpfung, um neues Wissen aus dem analytischen System extrahieren zu können. Eine weitere Herausforderung stellt die Multilingualität der Wissensdomänen und der Textkollektionen dar.

Das CBP Projekt zielt auf die Integration klinischer sowie experimenteller Daten bei klinischen Studien und Forschungsprojekten ab, um eine analytische Sichtweise auf die verknüpften und integ-

1 Private Universität für Gesundheitswissenschaften, Medizinische Informatik und Technik, Institut für Biomedizinische Technik, UMIT, Hall in Tirol

2 Biomedical Engineering / eHealth systems, Austrian Research Centers GmbH - ARC, Graz

rierten Datenbestände bekommen zu können, und um mittels Biostatistik, Data Mining Verfahren und systembiologischen Ansätzen neue Biomarker identifizieren und Behandlungsmethoden und Medikamentenentwicklung verbessern zu können.

2. Methoden & Ergebnisse

Das Gesamtsystem besteht aus einem GCP konformen Electronic Data Capture (EDC) System [1]. Projektpartner können mittels diesem System alle für die Studie relevanten Daten und Informationen integrieren, d.h es können Patienten im System registriert und Hochdurchsatzdaten mit den Patientenproben verlinkt werden. Weiters regelt dieses System das Benutzer / Rechte Management und stellt eine Kommunikationsschnittstelle bereit.

Als IBM im Jahre 1980 das Konzept des Data Warehouses [2] geprägt hat, wurde dieses speziell in der Betriebsinformatik zur Analyse von Wirtschaftssystemen eingesetzt. Ein Data Warehouse und dessen integriertes Schema kann als mehr oder weniger stabiles System zur Durchführung von Business Intelligence (BI) angesehen werden. Üblicherweise wird ein Data Warehouse bei so genannten Tagesabschlüssen mit neuen Daten bestückt, in vergangene Daten wird nicht mehr eingegriffen. Die Anforderungen an ein biomedizinisches Data Warehouse sind jedoch anders, da die Integration der diversen heterogenen Datenquellen wesentlich komplexer ist, und es sich um hoch dynamische Datenrepräsentationen handelt. Durch das, wie oben beschriebene, rasante Wachstum an Daten (Publikationen, Biodatenbanken, ...) benötigt ein biomedizinisches Data Warehouse eine andere Konzeption, wie ein traditionelles Warehouse System.

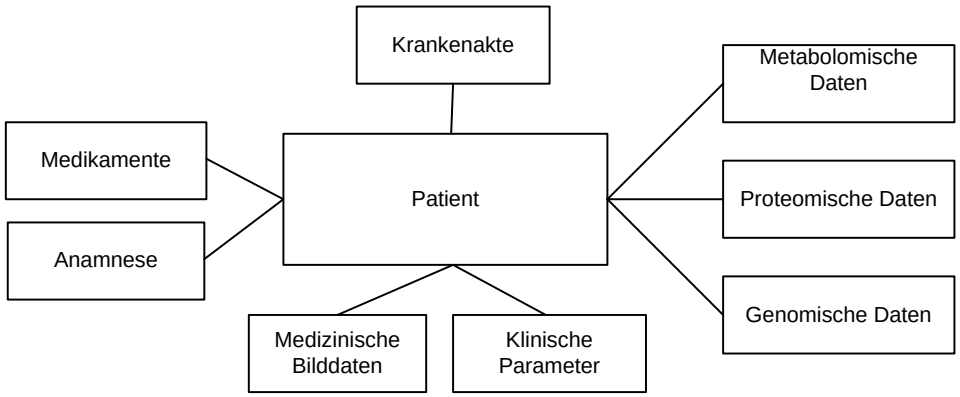

Abbildung 1: Mittels Star-Schema modelliertes Datenbankmodell. Aufgrund der einfachen Strukturierung ist dieses leicht verständlich. Eine Erweiterung des Schemas ist mit relativ wenig Aufwand möglich.

Die Verwendung des (Bio)Star Schemas [2-5] erlaubt das Erfassen von semantischen Informationen und das einfache Erweitern des Schemas, um weitere Datenquellen abbilden zu können. Daher wurde ein Schema mittels Entity-Relationship Modellierung erstellt [6], welches die Integration verschiedenster Datenquellen erlaubt. Eine schematische Darstellung ist in Abbildung 1 zu finden.

Der Extraktion-Transformation-Lade (ETL) Prozess [3] verarbeitet die im transaktionsorientierten EDC-System vorhandenen Daten und integriert diese in das analytische System. Dazu werden die Rohdaten transformiert (Schemaadaption und interne Datenrepräsentation), bereinigt (entfernen von fehlerhaften Daten, ...) und mit zusätzlichen Informationen und Annotationen versehen. Mittels Standard Abfragetools kann dann das Data Warehouse abgefragt und die Ergebnismenge der Analysepipeline übergeben werden. Abbildung 2 zeigt das System in einer schematischen Ansicht.
Alle in Forschungsprojekten und Studien anfallenden Daten werden im EDC-System erfasst, mittels ETL Prozessen bereinigt und in das Data Warehouse integriert. Das Data Warehouse dient als Datenpool für Analysen und systembiologische Modellbildung.

3. Diskussion & Schlussfolgerung

Ein einheitliches, integriertes System erlaubt den bei einer klinischen Studie teilnehmenden Forschungspartnern einzelne Schritte nach einem bestimmten Workflow abarbeiten zu können. Speziell ist zu erwähnen, dass biomedizinische Daten anders zu behandeln sind als Daten aus Wirtschaftssystemen. Aus dieser Tatsache motiviert, wurde ein leicht auf biomedizinische Projekte adaptierbares und erweiterbares Schema entwickelt. Die qualitätsgesicherten und integrierten Daten können mittels „database query processing“ abgefragt, extrahiert und systembiologischen Ansätzen als Eingabe dienen.

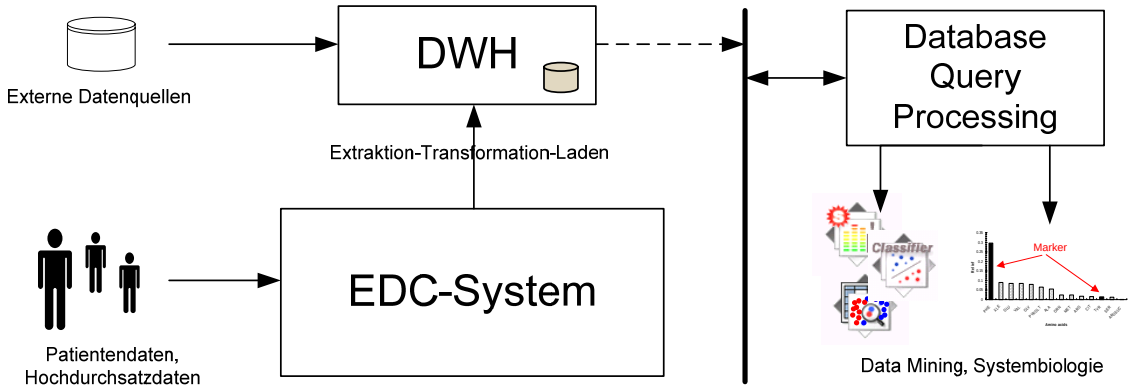

Abbildung 2: Schematische Darstellung der Plattform. Mittels EDC System Schema – Architektur

Mit dem vorgestellten Konzept wird eine Brücke gebaut zwischen der klinischen Domäne einerseits und der biomolekularen Domäne andererseits, zwei Welten, zwischen denen nach wie vor eine große Lücke klafft, die es für eine integrierte Betrachtung von Geno- und Phänotyp zu überwinden gilt.

4. Danksagung

Diese Arbeit wurde gefördert durch das Bundesministerium für Wirtschaft und Arbeit, durch die Tiroler Zukunftsstiftung und durch das Kompetenzzentrum HITT-health information technologies tyrol.

5. Referenzen

[1] Messmer, J., Ambros, P., Beiske, K., Hormann, M., Lewington, V., Helfre, S., Potschger, U., Burchil, I., Ladenstein, R., Schreier, G.: Eine web-basierte Plattform zur kollaborativen klinischen Forschung im Bereich der Onkologie. Tagungsband der 51. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds) (September 10.-14, Leipzig. 2006)

[2] Bauer, A., Gunzel, H.: Data Warehouse Systeme. 2. dpunkt.verlag (2004)

[3] Kimball, R., Caserta, J.: The Data Warehouse ETL Toolkit. Wiley Publishing Inc. (2004)

[4] Bloor: Data Warehousing Tools and Solutions. IT-Verlag (1997)

[5] Wang, L., Zhang, A.: Biostar models of clinical and genomic data for biomedical data warehouse design. Int. J. Bioinformatics Research and Applications 1(1) (2005) 63–80

[6] Thalheim, B.: Entity-Relationship Modeling - Foundadtions of Database Technology. Springer (2000)