Visualisierung und Analyse im Kontext onkologischer Daten

Daniel Nasseh[12], Diana Schweizer[2], Ulrich Mansmann[1]

(1) IBE, Ludwig-Maximilians-Universität München

(2) Comprehensive Cancer Center - LMU, Klinikum der Universität München

Das CCC (Comprehensive Cancer Center) Netzwerk besteht aktuell aus 13 onkologischen Spitzenzentren, deren Hauptaufgabe darin besteht, die einzelnen onkologischen Fachrichtungen institutionalisiert zusammenzuführen, sowie die gemeinsame medizinische Versorgung und Forschung zu verbessern (www.ccc-netzwerk.de).

Eines dieser Spitzenzentren ist das CCC-München (www.ccc-muenchen.de), organisatorisch untergliedert in das CCC-LMU (Ludwigs-Maximilians-Universität), mit Standort am Klinikum der Universität München bzw. das RHCCC (Roman Herzog CCC), mit Standort am Klinikum am Rechts der Isar. Beide CCCs lassen sich jeweils als ein Verbund verschiedener Organzentren der jeweiligen Standorte beschreiben.

Während der Patientenversorgung fallen eine Vielzahl patientenbezogener Daten an, die aus verschiedenen Quellen und Systemen stammen. Zu nennen wären beispielsweise Befunde aus Pathologie und Radiologie, Arztbriefe oder auch Laborberichte.

Über alle Zentren hinweg werden diese Daten jedoch standortspezifisch, zentral in einer gemeinschaftlichen Tumordokumentationssoftware (CREDOS) elektronisch und strukturiert gespeichert (Verweis zum Hersteller: www.uniklinik-ulm.de).

Zum Stand Anfang 2016 befinden sich in der Datenbank auf Seite des CCC-LMU aktuell Informationen zu c.a. 10.000 Patienten, mit einem Umfang von aktuell 65 Tabellen und mehreren hundert Datenfeldern. Jedes Datenfeld wiederrum kann teilweise hunderte von Ausprägungen besitzen (z.B. die Kodierung der Diagnosen oder Maßnahmen).

Hauptaufgabe der Tumordokumentation ist die alljährliche Zertifizierung nach Onkozert (www.onkozert.de), bei der es gilt, Kennzahlen als Beleg dafür zu generieren (z.B. Anzahl aller Primärfälle), dass das Zentrum die Auflagen eines Spitzenzentrums (nach den Kriterien der Deutschen Krebshilfe) erfüllt.

Die zentrale Erfassung der Daten ist aufwendig und ihre Zweitnutzung (Secondary Use) wäre, wie auch andere Projekte bereits demonstriert haben, erstrebenswert [1]. Da die Datenbank viele wichtige onkologische Parameter enthält würde es sich also anbieten sie zur Generierung und Analyse wissenschaftlicher Hypothesen, zur Analyse von Qualitätsindikatoren oder auch zur Prozessoptimierung zu verwenden (um nur ein paar Anwendungsfelder zu nennen).

Problematisch dabei ist die Komplexität der Datenbank. Klassische statistische Tools wie SPSS, SAS oder R können zwar auch eingesetzt werden um klar definierte Datenaggregationen, Visualisierungen und Übersichten aufzubereiten, das schnelle, oft iterative, durchforsten hochkomplexer Daten, ohne zunächst klarer Fragestellung, stellt sich jedoch eher als mühsam dar. Für den Umgang mit großen Daten wären also insbesondere flexible Systeme anzustreben, die einerseits klar definierte Fragestellungen beantworten können, andererseits und insbesondere jedoch auch erlauben in Echtzeit, reaktiv, ausgehend von der höchsten Komplexitätsstufe, beispielsweise über das wiederholte einschränken der Datenkohorte, in die Daten einzudringen.

Zudem wäre es wichtig Werkzeuge zur Verfügung zu stellen, die auch ohne informatische, oder statistische Fachkenntnisse nutzbar und verständlich sind. Insbesondere wäre anzustreben, dass sich der Facharzt bzw. Onkologe (der das notwendige Fachwissen besitzt um beispielsweise Auffälligkeiten zu erkennen) selber und direkt mit solch einer Lösung auseinander setzen kann, denn die Kommunikation und Vermittlung von Inhalten zwischen IT / Statistik und Arzt kann zu Missverständnissen und somit Blockaden führen.

Am CCCM wurde deswegen beschlossen die Auswertung über ein fortgeschrittenes Dataanalysis Tool der Firma Qlik in Form einer umfangreichen Dataanalysis Plattform umzusetzen (www.qlik.com).

Bei der Software die zum Aufbau verwendet wird handelt es sich um Qlik-View, sehr gut geeignet zur Echtzeitanalyse und Datenvisualisierung und im renommierten Gartner Quadrant, der sich mit der Bewertung von Softwarelösungen beschäftigt hoch eingruppiert (www.gartner.com).

Das wissenschaftliche Feld, das sich mit der angestrebten Art der Aufarbeitung von großen Datenmengen beschäftigt und in deren Domäne die Software einzuordnen wäre trägt den Namen Business Intelligence.

Der Einsatz von BI Lösungen in der Medizin ist zwar nicht neu, so werden Daten an Krankenhäusern über ETL (Extract-Load-Transfer) Prozesse oftmals in Datawarehouse-Lösungen zusammengeführt [3,4], der Grad an Aufbereitung ist jedoch oft rudimentär, so dass eine Verringerung der Komplexität kaum erreicht werden kann. Die Präsentationsebene solcher DWH beschränkt sich hier oft auf einfache Suchmasken und einige wenige Filter [5,6].

Bei der Architektur der Datawarehouse-Lösungen findet man in vielen Projekten zwei Ebenen. Eine Integrationsebene und eine Analyseebene. Die Integrationsebene, oft realisiert über ein OLTP System, dient als Datenlager und führt die Daten aus verschiedenen Quellen zusammen [7]. In gegebenem Projekt könnte man CREDOS als Umsetzung solch einer Integrationsebene/Persistenzebene interpretieren welches als verknüpftes, relationales Datenmodell in QV überführt wird.

Die Analyseebene wird klassischer Weise über ein sogenanntes OLAP System umgesetzt, welches Daten gezielt aufbereitet und auch in Echtzeit nutzbar macht [8]. Informatisch verbirgt sich hinter einem OLAP System ein sogenannter OLAP Würfel, der insbesondere der schnellen Datenauswertung dient, dessen Erstellung jedoch aufwendig und komplex sein kann. Das Datenbankmodel ist hierbei multidimensional.

Qlik-View hingegen überführt die komplette (im Regelfall) relationale Datenbank in den Arbeitsspeicher (In-Memory Datenbank) und ermöglicht hierdurch Echtzeitanalysen ohne vorhergehende Datenaufbereitung, wie sie z.B. für die Erstellung eines OLAP-Würfels notwendig wäre. Auch andere fortgeschrittene Systeme wie SAP-HANA setzen deswegen nun vermehrt auf In-Memory Datenbanken anstelle von OLAP Systemen. [9]

Unabhängig von den benutzten Systemen müssen die Daten natürlich informatisch aufbereitet werden und Module müssen erstellt werden, mit denen man in der Lage ist beispielsweise Datenkohorten schnell einzuschränken.

Reine Tabellenansichten und Suchmasken sind dabei problematisch, da die Komplexität der Datenbank groß ist und Tabellenansichten kaum zur Verringerung der Komplexität beitragen.

Der Versuch des CCCM ist es deswegen innovative und einfach benutzbare Module zu kreieren, die insbesondere auf visuelle, fast schon spielerische Art und Weise versuchen dem Nutzer einen schnellen Einstieg in die Benutzung zu erlauben. Das Schlagwort "Usability" hat hierbei also größte

Priorität. Inspiration hierbei stammt auch aus modernen Smart-Phone Anwendungen die oftmals durch einfache und intuitive Benutzung überzeugen sowie erfolgreichen Pilotanwendungen am eigenen Standort. [10,11]

Beispielhaft soll eines dieser Module vorgestellt werden. Bei gegebenem Modul handelt es sich um eine dynamische Erkrankungskarte (siehe Abbildung 1).

Abbildung 1: *Darstellung der relativen Häufigkeit des tumorbedingten Organbefalles auf Basis der CREDOS Daten. Es handelt sich hierbei um eine noch in Überarbeitung befindliche Version für Demonstrationszwecke.*

Die Farbsättigung der Organe entspricht dabei der relativen Häufigkeit des Organbefalles. So wird schnell ersichtlich, dass beispielsweise die Prostata (mit kräftigem rot hinterlegt) am Standort sehr häufig behandelt wird, wohingegen sich der Ösophagus / Speiseröhre (mit einem hellen rot hinterlegt) eher als ein weniger behandeltes Organ darstellt.

Bei der gegebenen Abbildung handelt es sich nun nicht um eine statische Grafik sondern um ein dynamisches Modul. Einzelne Organe lassen sich per Mausklick und, dank In Memory Technologie, in Echtzeit selektieren. Die Auswahl resultiert in einer differenzierteren Ansicht des Organes, wie nachfolgend gezeigt am Beispiel des Magens (siehe Abbildung 2).

Abbildung 2: *Darstellung der relativen Häufigkeit von Tumorerkrankungen am Magen auf Basis der CREDOS Daten. Die zugrunde liegende Datenkohorte wurde hierbei durch vernetzte Tabellen jeweils auf Patienten mit männlichem, bzw. weiblichem Geschlecht eingeschränkt. Die Kardia des Magens, deren Befall in den beiden unterschiedlichen Datenkohorten einen stark unterschiedlichen relativen Befall aufweist wurde mit einem orangen Kreis markiert*.

Eine Besonderheit von Qlik-View dabei ist, dass alle Tabellen und Objekte bzw. Grafiken miteinander verknüpft sind. Das heißt, über einfache Mauseingaben, in einer der integrierten Tabellen, lassen sich Filter setzen, wie beispielsweise die Einschränkung der Datenkohorte auf das Patientengeschlecht. In Echtzeit färbt sich das Organ nun entsprechend ein.

Wie man anhand vorhergehender Abbildung erkennt, ist die relative Erkrankungsrate der Kardia (Magenöffnung) bei Männern, zumindest auf Basis des Datenbestandes, höher als bei den in der Datenkohorte eingeschlossenen Frauen. Auch ohne Vorwissen können so top-down mit den Daten in kürzester Zeit wissenschaftliche Indizien aufgedeckt werden.

Jedoch dürfen die Ergebnisse die hier beispielhaft vermittelt wurden nicht als wissenschaftlich fundierte Tatsachen fehlinterpretiert werden. Zum einen spielt die Datengrundlage eine große Rolle (z.B. Sample Size, Beschaffenheit der Datenkohorte) zum anderen handelt es sich um Daten aus der Versorgung bei denen eine perfekte und abgesicherte Datenqualität nicht garantiert werden kann (z.B. Genauigkeit der Kodierung der Diagnose). Grundsätzlich muss man sich auch im Klaren darüber sein, dass Störfaktoren berücksichtigt werden müssen (Confounding) [12]. Bzgl. vorhergehendem Beispiel ist es zwar bekannt, dass es geschlechterspezifische Unterschiede bei der Häufigkeit von Erkrankungen am Magen gibt (Männer erkranken öfter als Frauen) [13], zum differenzierten Vergleich fällt es allerdings schwerer Literatur zu identifizieren. Möchte man nun also beginnen nach

Ursachen zu forschen wäre es also beginnend angebracht den Fund auch beispielsweise erst auf Basis anderer umfangreicherer Daten (z.B. denen aus epidemiologischen Krebsregistern) zu validieren.

Festhalten kann man jedoch, dass das Tool, so wie es im Moment konzipiert wurde sicherlich zur Hypothesengenerierung bzw. zur Aufdeckung interessanten Wissens (Data Discovery) [10] genutzt werden kann. Jedoch sollten interessanten Auffälligkeiten mit einer tiefgreifenden statistischen Analyse überprüft werden.

Aktuell befindet sich die Auswertungsplattform noch in Entwicklung, soll aber zum Ende des Jahres 2016 für den internen Gebrauch freigegeben werden. Wichtige Fragen die noch Klärungsbedarf benötigen sind insbesondere der Datenschutz und das Berechtigungskonzept.

Voraussichtlich wird ein zweckorientiertes Berechtigungssystem verwendet werden. Dies bedeutet, dass nur dann Daten freigegeben werden dürfen, wenn die daran gekoppelte Datenfreigabe klar einen Zweck erfüllen. Über die Freigabe wird ein gemeinschaftliches Gremium entscheiden. Um die Sache etwas zu erleichtern, würden alle Ärzte initial vollen Zugriff auf Daten bekommen, die Ihnen auf Grund ihres Behandlungsauftrages auch bereits in den Klinikumsinformationssystemen freigegeben wurden.

Bezüglich des Datenschutzes würden Ärzte zu jenen Patienten auch Klartextangaben zu identifizierenden Attributen lesen dürfen. Zweckbedingte Freigabe weiterer Patienten/Daten würde eine Pseudonymisierung erster Stufe voraussetzen.

Die exakte Umsetzung des Berechtigungssystemes ist jedoch noch in Planung und muss vom ansässigen Datenschützer geprüft werden.

Literatur:

[1]: Mosa AS, Yoo I, Apathy NC, Ko KJ, Parker JC. Secondary Use of Clinical Data to Enable Data-Driven Translational Science with Trustworthy Access Management. Mo Med. 2015 Nov-Dec;112(6):443-8.

[2]: Ross M. The Bottom-Up Misnomer - DecisionWorks Consulting. DecisionWorks Consulting. Online unter: http://completedwh.blogspot.de/2012/12/top-down-vs-bottom-up-in-data.html [Stand: 02.05.2016]

[3]: Majeed RW, Röhrig R. Automated realtime data import for the i2b2 clinical data warehouse: introducing the HL7 ETL cell. Stud Health Technol Inform. 2012;180:270-4.

[4]: Pecoraro F, Luzi D, Ricci FL. Designing ETL Tools to Feed a Data Warehouse Based on Electronic Healthcare Record Infrastructure. Stud Health Technol Inform. 2015;210:929-33.

[5]: Meineke FA, Stäubert S, Löbe M, Winter A. A Comprehensive Clinical Research Database based on CDISC ODM and i2b2. Stud Health Technol Inform. 2014;205:1115-9.

[6]: Rubin DL, Desser TS. A Data Warehouse for Integrating Radiologic and Pathologic Data

[7]: Gabriel R, Pastwa A, Gluchowski P.: Data Warehouse & Data Mining.

[8]: Gluchowski P, Chamoni P. Entwicklungslinien und Architekturkonzepte des On-Line Analytical Processing. In: Analytische Informationssysteme: Business Intelligence-Technologien und -Anwendungen. 4., vollständig überarbeitete Auflage, 2010, S. 200-202

[9]: Sevilla M. OLAP databases are being killed by In-Memory solutions. Online unter: https://www.capgemini.com/blog/capping-it-off/2011/09/olap-databases-are-being-killed-by-in-memory-solutions [Stand: 29.04.2016]

[10] Martin W. Data Discovery – BI im Zeitalter von Apps und Social Media. Online unter: http://drmartin.dbxl.de/Ausgabe-72 [Stand: 03.05.2016]

[11]: Nasseh D, Müller M, Ahlborn B, Kortüm K, Kampik A, Mansmann U, Kreutzer T. SMEYEDAT (SMART-EYE-DATA): Zusammenführung und Nutzbarmachung ophtalmologischer Daten. GMDS Jahrestagung 2015, Sep 2015, Krefeld.

[12]: Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. International journal of epidemiology, 15(3), 413-419.

[13]: Tiing LA, Kwong MF. Clinical epidemiology of gastric cancer. Singapore Med J. 2014 Dec; 55(12): 621–628.