Dokumente 5

Gesundheitswissenschaften

Nutzenpotenzial von Krankenversicherungsdaten am Beispiel der Oberösterreichischen Gebietskrankenkasse

Mag. (FH) Joachim Hagleitner

Herausgegeben von

Univ.- Prof. Dr.

Josef Weidenholzer,

Institut für Gesellschafts-

und Sozialpolitik,

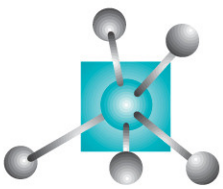
Johannes Kepler

Universität Linz in

Zusammenarbeit mit

der Oberösterreichischen

Gebietskrankenkasse.



Linz, 2005

Fachhochschul-Studiengänge
Betriebs- und Forschungseinrichtungen
der Wiener Wirtschaft GesmbH

Fachhochschul-Studiengang

Unternehmensführung

Titel der Diplomarbeit:

Nutzenpotenzial von Krankenversicherungsdaten am Beispiel der Oberösterreichischen Gebietskrankenkasse

Verfasst von: **Joachim Hagleitner**

Betreut von: **a.o. Univ. Prof. Dr. Franz Hörmann**

Ich versichere:

- dass ich die Diplomarbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfe bedient habe.
- dass ich dieses Diplomarbeitsthema bisher weder im In- noch im Ausland in irgendeiner Form als Prüfungsarbeit vorgelegt habe.

_____ _____

Datum Unterschrift

Inhalt

1 Ausgangssituation und Forschungsfrage

In den unterschiedlichsten Wirtschafts- und Wissenschaftsgebieten nimmt die Menge elektronisch verfügbarer Daten rapide zu. An den Scannerkassen des amerikanischen Unternehmens WalMart werden beispielsweise täglich 20 Millionen Transaktionen erfasst und in eine Datenbank eingespeist, die eine Größe von 24 Terrabyte erreicht. Die Satelliten des Earth Observing System der NASA senden stündlich 50 Gigabyte an Bildmaterial zur Erde.[1] Das Data Warehouse der oberösterreichischen Gebietskrankenkasse, in dem die Daten von rund einer Million Versicherten verwaltet werden, umfasst bereits 600 Gigabyte.[2] Die Aufzählung von Beispielen könnte fast beliebig fortgesetzt werden und verdeutlicht, dass derart große Datenmengen nach ausgefeilten, oft automatisierten Verarbeitungs- und Auswertungsmethoden verlangen, um entscheidungsrelevante und strategisch wichtige Informationen aus der unüberschaubaren Menge an Daten zu extrahieren.

Krankenversicherungen stehen heute vor der Herausforderung, die hohe Qualität der bestehenden medizinischen Angebote zu gewährleisten und durch neue Leistungen zu ergänzen und gleichzeitig die verfügbaren finanziellen Mittel effizient einzusetzen und das Wachstum der Ausgaben möglichst einzudämmen. Die vorliegende Diplomarbeit beschäftigt sich mit der Frage, welche Methoden und Techniken im Rahmen betrieblicher Informationssysteme von Krankenversicherungen genutzt werden können, um dieser Herausforderung zu begegnen.

Zur Beantwortung der Fragestellung werden betriebliche Informationssysteme und moderne Werkzeuge zur Auswertung von Daten dargestellt. Internationale Studien zeigen den Nutzen aber auch die Gefahren der unterschiedlichen Anwendungsgebiete im Krankenversicherungsbereich auf. Am Beispiel der Oberösterreichischen

[1] vgl. Runkler, T.A.: Information Mining: Methoden, Algorithmen und Anwendungen intelligenter Datenanalyse. Braunschweig/Wiesbaden: Vieweg&Sohn 2000, S.1

[2] vgl. Hofer, P.: Kosteneinsparung und Qualitätssicherung in der Behandlungsökonomie durch Folgekostenanalyse mit SAS bei der oberösterreichischen Gebietskrankenkasse. Tagungsbericht: ICV Forum Gesundheitswesen Österreich 2004, 24.09.2004, Wien 2004, o.S.

Gebietskrankenkasse werden die Entwicklung und der Aufbau eines Data Warehouse beleuchtet, das sowohl der Kontrolle der Kosten als auch der Verbesserung der Versorgungsqualität und Forschungszwecken dient.

2 Aufbau der Arbeit

Der Aufbau der Arbeit gliedert sich in drei Teile. Im ersten Teil werden methodische Ansätze und Werkzeuge zur Analyse von Daten im Rahmen betrieblicher Informationssysteme beschrieben. Die Anforderungen an ein gut funktionierendes Data Warehouse, das die Grundlage für gute Auswertungen darstellt, werden erläutert. Die vorliegende Arbeit konzentriert sich auf Anwendungen im Rahmen des Online Analytical Processing (OLAP), Knowledge Discovery in Databases und Data Mining, da diese im Bereich der Krankenversicherungen eine wichtige und zukunftsweisende Rolle spielen. Vor allem die Methoden des Data Mining und deren betriebliche Nutzung werden näher beschrieben.

Der zweite Teil der Arbeit widmet sich dem breiten Spektrum der Datennutzung durch Krankenversicherungen. Anhand von internationalen Studien werden die Beratung von bestimmten Patientengruppen, die Analyse bestimmter Krankheitsbilder, die Aufdeckung von Betrugsfällen und die Versorgungsforschung näher beleuchtet. Die Nutzung von Krankenversicherungsdaten birgt – vor allem im Bereich des Datenschutzes – Gefahren in sich, diese werden am Ende des zweiten Teils behandelt.

Der dritte und abschließende Teil der Arbeit widmet sich dem Fallbeispiel der Oberösterreichischen Gebietskrankenkasse. Zunächst werden der Aufbau des Data Warehouse der OÖGKK und die Implementierung einer Software zur Analyse der verrechneten Leistungen (kurz FOKO) beschrieben. Besonders wichtige Anwendungsgebiete sind die Analyse von Kosten und die Aufdeckung von Einsparungspotenzialen (v.a. im Bereich der Heilmittel), die Bereitstellung von Informationen für Ärzte und Patienten, die Entwicklung von Behandlungsleitlinien und Forschungsprojekte zu aktuellen Themen.

3 Betriebliche Informationssysteme

Unter Informationssystemen werden soziotechnische Systeme verstanden, deren Zweck die Schaffung einer Kongruenz von Informationsangebot und –nachfrage ist. Betriebliche Informationssysteme werden häufig mit Hilfe von Informationssystempyramiden (**Abbildung 1**) veranschaulicht, wobei die einzelnen Ebenen der Pyramide den von unten nach oben zunehmenden Grad der Informationsverdichtung darstellen. Die Spitze der Pyramide bildet die für die Unternehmensleitung notwendige Führungsinformation, während die Basis aus den Massendaten im Produktionsprozess entsteht. Die Doppelpfeile verdeutlichen sowohl die horizontale Integration von Systemen in der gleichen Ebene als auch vertikale Integration über- und untergeordneter Ebenen. Mit der Integration soll erreicht werden, dass alle Systeme ihre Informationen untereinander vollautomatisch austauschen. Im Idealfall kann dadurch ein Schnittstellenproblem zwischen den Systemen ausgeschlossen werden, da alle Systeme eine gemeinsame redundanzfreie Datenbasis besitzen.[1]

[1] vgl. Totok, A.: Modellierung von OLAP- und Data-Warehouse-Systemen. Wiesbaden: Gabler 2000, S. 37

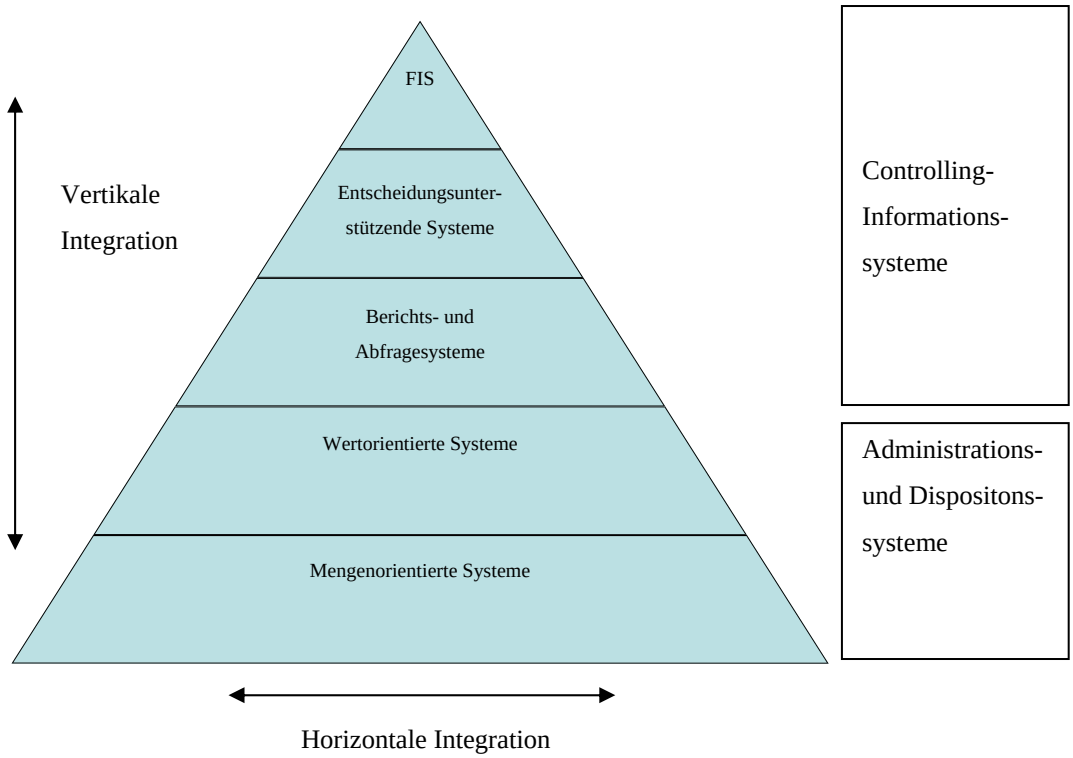


Abbildung 1: Integrierte Informationssysteme

(Quelle: Totok, 2000, S.38)

Die unteren beiden Ebenen der Pyramide werden unter dem Begriff *Administrations- und Dispositionssysteme* (Kapitel 3.1) zusammengefasst, *Controllinginformationssysteme* (Kapitel 3.3) sind im oberen Teil der Pyramide abgebildet.[1]

Die ersten Ansätze, mittels entscheidungsunterstützender Systeme Unternehmensdaten strategisch zu nutzen, um Wettbewerbsvorteile zu erzielen, gehen in die 1970er Jahre zurück. Information kam zusehends die Bedeutung eines Produktionsfaktors zu.[2]

[1] vgl. Totok, 2000, S.38

[2] vgl. Martin, W.: Data Warehouse, Data Mining und OLAP: Von der Datenquelle zum Informationsverbraucher. In: Martin, W. (Hrsg.): Data Warehousing: Data Mining – OLAP. Bonn: International Thompson1998, S.20

Voraussetzung für das Funktionieren eines Controllinginformationssystems ist die Einrichtung eines Data Warehouse, das die Nutzung unterschiedlicher Daten im Unternehmen überhaupt erst ermöglicht.[1] Das Data-Warehouse-Konzept wird in Kapitel 3.2 genauer ausgeführt.

3.1 Administrations- und Dispositionssysteme

Administrations- und Dispositionssysteme dienen der Massendatenverarbeitung der betrieblichen Produktionsfaktoren. Dementsprechend werden sie auch als operative Systeme bezeichnet, das Einsatzgebiet dieser Systeme wird häufig als *Enterprise Resource Planning* (ERP) bezeichnet.

Die Rationalisierung der Massendatenverarbeitung durch die Automatisierung von Routineaufgaben ist die Hauptfunktion der Administrationssysteme. Die Systeme werden vor allem auf mengenorientierte Prozesse (z.B. Beschaffung, Produktion) und wertorientierte Prozesse, (z.B. Finanz- oder Lagerbuchführung) angewendet. Dispositionssysteme werden auf unterer oder mittlerer Führungsebene eingesetzt und sollen die Lösung von gut strukturierten Problemen entweder vereinfachen oder überhaupt automatisieren. Ist die maschinelle Entscheidung besser als die menschliche, so entsteht ein Optimierungsnutzen. Ein Rationalisierungsnutzen ergibt sich, wenn die menschlichen Entscheidungsträger von Routineentscheidungen entlastet werden. Eingesetzt werden Dispositionssysteme beispielsweise im Bestellwesen zur Ermittlung optimaler Losgrößen.[2]

Operative Systeme greifen auf Produktionsdatenbanken zurück, die das laufende Geschäft unterstützen. In Produktionsdatenbanken werden Geschäftsvorgänge abgelegt, indem die einzelnen Transaktionen laufend, vollständig und redundanzarm fortgeschrieben werden. Ihrem Zweck entsprechend verarbeiten operative Systeme

[1] vgl. Lusti, M.: Data Warehousing und Data Mining: Eine Einführung in entscheidungsunterstützende Systeme. Berlin: Springer Verlag 1999, S.125

[2] vgl. Totok, 2000, S.38f.

auch eine sehr hohe Anzahl an Transaktionen sehr effizient.

Zur Entscheidungsunterstützung auf höherer Ebene sind sie aus mehreren Gründen ungeeignet:

- Die operativen Daten sind oft unübersichtlich, da eine Vielfalt von Details unterschiedlicher Anwendungen in ihnen enthalten ist.
- Die Produktionsdaten werden häufig überschrieben, für viele Auswertungen sind aber historische Daten erforderlich.
- Produktionsdatenbanken sind meist wenig benutzerfreundlich, da die Abfragewerkzeuge in erster Linie die Routineverarbeitung unterstützen und nicht auf eingehendere Analysen ausgerichtet sind.[1]
- Die Daten sind teilweise redundant und liegen in unterschiedlichen Formaten an verschiedenen Orten vor.[2]

3.2 Das Data Warehouse

3.2.1 Definitionen des Data Warehouse

Die eben beschriebenen Mängel der operativen Datenbanken und technologische Fortschritte führten Anfang der 1990er Jahre zum Konzept des Data Warehouse.[3] Das Aufkommen von Data Warehousing wird mit W. Inmon and E.F. Codd in Verbindung gebracht, die feststellten, dass operative Systeme und entscheidungsunterstützende Systeme nicht effizient in derselben Datenbankumgebung eingesetzt werden können, vor allem aufgrund ihrer unterschiedlichen Transaktionscharakteristiken.[4]

[1] vgl. Lusti, 1999, S.124f.

[2] vgl. Totok, 2000, S.41

[3] vgl. Lusti, 1999, S.125

[4] vgl. Jarke, M. et al.: Fundamentals of data warehouses. Berlin: Springer 2000, S.1

Die in der Literatur zu findenden Definitionen des Begriffs Data Warehouse sind nicht völlig identisch. Anahory und Murray bezeichnen Data Warehouse in einer einfachen Definition als *„eine Sammlung von Schlüsselinformationen, die verwendet werden, um eine Firma auf die profitabelste Art und Weise zu verwalten und zu führen“*.[1]
Ähnlich lautet auch eine Definition von Jarke et al.: *„A data warehouse is a collection of technologies aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. It is expected to have the right information in the right place at the right time with the right cost in order to support the right decision”*.[2]

Wörtlich wird Warehouse mit Lagerhaus oder Speicher übersetzt und nicht mit Warenhaus. Das Herzstück des Data-Warehouse-Konzepts bildet die zentrale Datenbasis, der Aufbau und die Gestaltung dieser zentralen Datenbasis entscheiden oft über Erfolg und Misserfolg eines Data Warehouse.[3] Die zentrale Datenbasis wird im Englischen manchmal auch als *primary* or *core* Data Warehouse bezeichnet.[4] Imnon gibt eine Anforderungsdefinition für die zentrale Datenbasis: *„A data warehouse is a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management’s decision”* .[5]

Die einzelnen in der Definition enthaltenen Anforderungen erklären die Eigenschaften der zentralen Datenbasis:

- *subject oriented*: Damit ist gemeint, dass die Datenbasis nach den relevanten Unternehmensbereichen geordnet ist und nicht nach funktionalen Gesichtspunkten wie in operativen Datenbanken.
- *integrated*: Die aus anderen Systemen übernommenen Datenbanken müssen im Data Warehouse integriert und vereinheitlicht werden (z.B. hinsichtlich

[1] Anahory, S., Murray, D.: Data Warehouse: Planung, Implementierung und Administration. Bonn: Addison-Wesley-Longman 1997, S.19

[2] Jarke et al. 2000, S.1

[3] vgl. Totok, 2000, S.43

[4] vgl. Jarke et al. 2000, S.4

[5] Inmon, 1996, S.33, zitiert nach Totok, 2000, S.43

Typ, Format, Bezeichnung,…)

- *nonvolatile*: Dieser Teil der Definition zielt auf die Unveränderbarkeit der Daten nach der Übertragung ins Data Warehouse ab, die Zugriffe erfolgen nur lesend. Einzige Ausnahme stellen die Bereiche Planung und Prognose dar, falls diese in die zentrale Datenbasis gespielt werden.
- *time variant:* Die Daten im Data Warehouse werden über längere Zeit gespeichert, um Veränderungen über die Zeit abbilden zu können. Um ein Überschreiben älterer Daten zu verhindern, werden diese jeweils mit einem Datum versehen.[1] In vielen Unternehmen hat sich die Praxis eingebürgert, die Daten zwei Jahre lang im Data Warehouse zu halten.[2]

Neben diesen Anforderungen sollte das Data Warehouse die Erweiterung um neue Datentypen oder neue Aggregationsstufen erlauben. Weiters sollte die Skalierbarkeit des Datenbanksystems für schnell wachsende Datenbestände berücksichtigt werden.[3] Dass sich viele Unternehmen von der Implementierung eines Data Warehouse große strategische Vorteile erhoffen, zeigen die getätigten Investitionen in diesem Bereich. Im Jahr 1998 erreichten die Umsätze in diesem Bereich bereits $ 14,6 Mrd.[4]

3.2.2 Komponenten und Architektur des Data Warehouse

In der Architektur eines Data Warehouse sind mehrere Datenebenen dargestellt, wobei jede Ebene von der jeweils darunter liegenden Ebene mit Daten gespeist wird. Die unterste Ebene bilden die operativen, meist unternehmensinternen Datenbanken. Häufig werden auch externe Daten eingespielt. Die wichtigste Ebene stellt die im vorigen Kapitel erläuterte zentrale Datenbasis – auch als *primary* oder *core data warehouse* bezeichnet – dar.[5]

---

[1] vgl. Totok, 2000, S.43

[2] vgl. Jarke et al. 2000, S.4

[3] vgl. Totok, 2000, S.43

[4] vgl. Totok, 2000, S.1

[5] vgl. Jarke et al. 2000, S.4

Eine wichtige Funktion üben die Metadatenbanksysteme aus, sie enthalten Informationen über die Objekte eines Data Warehouse, wie z.B. Erläuterungen zu den enthaltenen Daten, Datenstrukturen oder Benutzer. Sie ermöglichen bzw. erleichtern die Orientierung im Data Warehouse. Da die zentrale Datenbasis sehr umfangreich werden kann, werden häufig bereichsspezifische Datenbanken (*data marts*) für spezielle Auswertungen angelegt. Die Daten in den data marts stammen in der Regel aus der zentralen Datenbasis und werden redundant geführt. Vorteile der data marts sind geringere Zugriffszeiten und Entlastung der zentralen Datenbasis.

Als Zwischenebene (*middleware*) zwischen den operativen Datenbanken und der zentralen Datenbasis dient der *operational data store* (ODS). Durch Datenextraktion und verschiedene Transformationsprozesse werden die Daten entweder direkt oder über den operational data store in die zentrale Datenbasis übergeführt. Drei wesentliche Merkmale unterscheiden den ODS von der zentralen Datenbasis:

- Der ODS wird häufig verändert im Vergleich zur zentralen Datenbasis.
- Im ODS werden nur aktuelle Daten gehalten.
- Die Daten im ODS sind wenig aggregiert und von feiner Granularität.

Eine zweite Zwischenebene liegt zwischen der zentralen Datenbasis und der obersten Architekturebene, diese Zwischenebene stellt die Plattform für die Controllinginformationssysteme in der obersten Architekturebene dar.[1]
Eine idealtypische Architektur des Data Warehouse ist in **Abbildung 2** dargestellt.

[1] vgl. Totok, 2000, S.39ff.

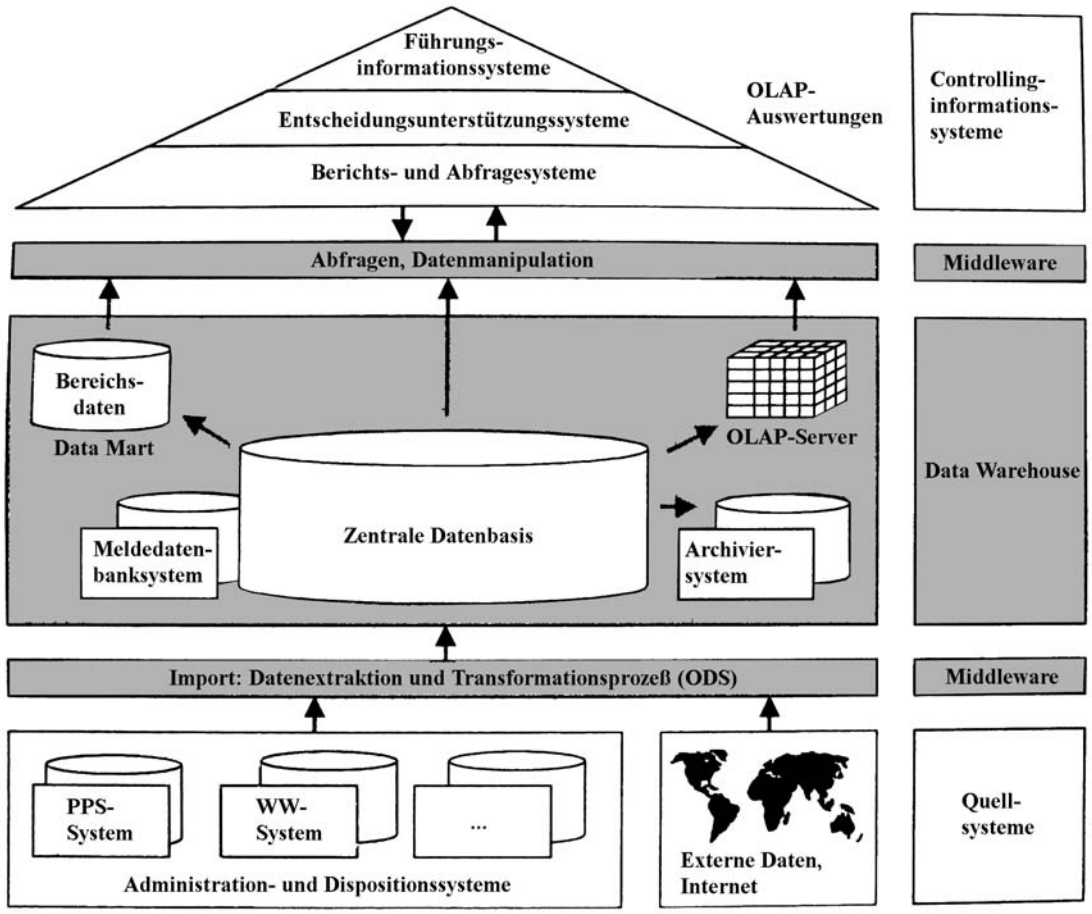


Abbildung 2: Schematische Darstellung eines Data Warehouse

(Quelle: Totok, 2000, S.40)

3.3 Controllinginformationssysteme

Controllinginformationssysteme werden häufig gleichgesetzt mit entscheidungsunterstützenden Systemen (EUS), tatsächlich unterscheiden sich oft eher die Nutzergruppen voneinander – z.B. Controlling oder Management – und nicht die Programmoberflächen.[1] Allgemein formuliert sind EUS computergestützte Informationssysteme, die Endbenutzern bei der Lösung komplexer Probleme helfen

[1] vgl. Totok, 2000, S.51

sollen. Häufig synonym verwendete Begriffe für EUS sind die englische Bezeichnung *Decision Support Systems* (DSS), *Management Informationsysteme* (MIS) und *Führungsinformationsysteme* (FIS) bzw. *Executive Information Systems* (EIS).[1] Anhand der Anwendungsgebiete können die Informationssysteme differenziert werden, wobei große Überschneidungen zwischen den Bereichen zu bedenken sind.

Management Reporting Systems (MRS) sind für die problemlose Erstellung von Berichten konzipiert. Sie ermöglichen die Ausgabe von Standardberichten, die in einem bestimmten Format und periodischen Abständen ausgegeben werden. Häufig erfolgt die automatische Ausgabe aufgrund der Über- oder Unterschreitung vorher festgelegter Schwellenwerte. Für den jeweiligen Anwender spezifische Bedarfsberichte und dialogisierte Abfrage- und Auskunftsmöglichkeiten fallen ebenfalls unter die MRS.[2]

Die Standardberichte markieren den Anfang der EUS und wurden unter dem Begriff Management Information Systems (MIS) bekannt, diese konnten sich damals aber nicht durchsetzen. Deswegen geriet der Begriff MIS auch etwas in Verruf. Dafür wird vor allem die durch die ersten MIS produzierte Informationsflut verantwortlich gemacht, die dem Management die Entscheidungen nicht erleichterte. Weiters fehlte neben den Zahlen aus den einzelnen Bereichen auch die Abstimmung zwischen den Bereichen und das Aufzeigen der wichtigsten Zusammenhänge, um eine zentrale Planung und Koordination zu ermöglichen.[3]

Die entscheidungsunterstützenden Systeme sollen vor allem bei der Lösung teilweise oder völlig unstrukturierter Probleme eine Hilfe sein. Im Gegensatz zu statischen Berichtssystemen kann interaktiv und iterativ an ein Problem herangegangen werden. Vor allem die Generierung von statistischen Kennzahlen und die Entwicklung von Szenarien spielen eine bedeutende Rolle bei den Decision Support Systems.[4]

---

[1] vgl. Lusti, 1999, S.3

[2] vgl. Küppers, B.: Data Mining in der Praxis: Ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld. Wien: Lang 1999, S.38

[3] vgl. Totok, 2000, S.48

[4] vgl. Totok, 2000, S.52

Führungsinformationssysteme (FIS) oder Executive Informations Systems (EIS) sind auf die Bedürfnisse der Führungsspitze eines Unternehmens zugeschnitten. Sie sollen einen schnellen und einfachen Zugriff auf zumeist hochverdichtete interne und externe Informationen ermöglichen. Großer Wert wird auf Benutzerfreundlichkeit und einfache Einarbeitung gelegt, damit die Führungskräfte schnell an die von ihnen gewünschte Information herankommen.[1]

[1] vgl. Küppers, 1999, S.38

4 Online Analytical Processing (OLAP)

Die Unternehmenswelt ist in vielen Bereichen durch hohe Dynamik gekennzeichnet, die Produktentwicklung muss rasch erfolgen, die Produktlebenszyklen sind oft kurz. Gesetzliche Anforderungen müssen schnell umgesetzt werden. OLAP soll das Management dabei unterstützen, Probleme in kurzer Zeit zu analysieren und adäquate Lösungen zu finden.

Das OLAP-Konzept geht auf E.F. Codd zurück, das er 1993 entwickelte, um Endbenutzern einen mehrdimensionalen und schnellen Zugriff auf Daten im Data Warehouse oder in abgegrenzten Data Marts zu ermöglichen und interaktive Analysen durchzuführen. OLAP erfordert die Haltung der Daten entlang von unternehmensrelevanten Dimensionen, die als Hyperwürfel (oder *Hypercube*) bezeichnet wird.[1] Die Bezeichnung Hyperwürfel rührt daher, dass ein Würfel streng genommen nur drei Dimensionen haben kann, in OLAP-Systemen wird die Anzahl von drei Dimensionen meist übertroffen.

Diese Art der entscheidungsorientierten Datenhaltung unterscheidet sich grundlegend von jener in transaktionsbasierten Datenbanken, **Tabelle 1** zeigt eine Gegenüberstellung wichtiger Eigenschaften beider Typen.

[1] Hönig, T.: Desktop OLAP in Theorie und Praxis. In: Martin, W. (Hrsg.): Data Warehousing: Data Mining - OLAP. Bonn: International Thompson 1998, S.171

Eigenschaft	Operativ	Entscheidungsorientiert
Anzahl paralleler Benutzer	Bis zu mehreren Tausend	Zweistelliger Bereich
Verarbeitung	Transaktionsbasiert	Analyseorientiert
Antwortzeiten	Millisekunden	Sekunden bis Minuten
Zugriffsfrequenz	Hoch	Mittel bis niedrig
Datenvolumen pro Zugriff	Niedrig	Hoch
Änderungen des Datenbestands	Häufig	Selten durch definierte Updates
Aktualität der Daten	Absolut	Tägliche, wöchentliche, oder monatliche Updates
Datenstrukturierung	Detailliert	Verdichtet
Datenbankgröße	10-100 Gigabyte	Bis zu Terrabytes

Tabelle 1: Anforderungen entscheidungsorientierter und operativer Datenhaltung

(Quelle: Totok, 2000, S.42)

4.1 Charakteristische Eigenschaften von OLAP

Die grundlegenden Anforderungen und Eigenschaften von OLAP-Systemen, die erstmals von Codd in Form von 12 Regeln formuliert wurden, geben einen Einblick in die Funktionsweise von OLAP. Ein OLAP-System kann selten alle Anforderungen optimal erfüllen, die Regeln stellen den Idealfall dar.[1]

Die folgende Beschreibung der 12 Regeln und Anforderungen von Codd zeigt die Eigenschaften und Vorteile eines funktionierenden OLAP-Ansatzes auf.

[1] vgl. Totok, 2000, S.57

Grundlegende Anforderungen

- *Multidimensionalität*: Die multidimensionale Sichtweise soll dem tatsächlichen Unternehmensumfeld möglichst nahe kommen und entscheidungsrelevante Dimensionen abbilden. Dies kommt der Realität näher als die Darstellung in flachen, zweidimensionalen Tabellen. Daten aus der so angelegten Datenbasis sollen von den Benutzern beliebig aggregiert bzw. verdichtet werden können. Die so entstehenden Verdichtungswege sollen aufgeteilt, zerlegt, gedreht oder rotiert werden können.
- *Intuitive Datenanalyse*: Die Benutzerschnittstelle des OLAP-Systems sollte so gestaltet sein, dass der Zugriff auf die Daten und die Nutzung der Funktionalitäten möglichst einfach und intuitiv erfolgen kann. Menüabfolgen zum Aufruf bestimmter Befehle sollten möglichst kurz gehalten werden.
- *Zugriffsmöglichkeiten*: Damit sich der Benutzer leicht im OLAP-System orientieren kann, müssen die Unternehmensdaten in einheitlicher und konsistenter Form dargestellt werden. Daher ist es oft erforderlich, unterschiedliche Formate aus den operativen Systemen automatisch zu konvertieren, bevor der Benutzer darauf zugreift. Die Einbeziehung unnötiger Daten sollte vom System möglichst vermieden werden.[1]
- *Transparenz*: Mit Transparenz ist gemeint, dass sich durch die Implementierung der OLAP-Anwendungen für den Endbenutzer möglichst nichts ändern sollte, der Übergang von der Einzelplatz-Lösung zur Client-Server-Architektur sollte sich auf das Verhalten der Arbeitsplatzrechner nicht auswirken. Die Benutzer sollen nicht bemerken, dass die Daten aus unterschiedlichen Quellen stammen bzw. sollen mit technischen Details verschont werden.[2]

[1] vgl. Totok, 2000, S.57f.

[2] vgl. Kirchner, J.: Online Analytical Processing. In: Martin, W. (Hrsg.): Data Warehousing: Data Mining – OLAP. Bonn: International Thompson 1998, S.154

Anforderungen bezüglich Berichtsgenerierung

- *Konsistentes Antwortzeitverhalten*: Die Leistungsfähigkeit der Hard- und Software muss so konzipiert sein, dass auch bei einer mehrdimensionalen Berichtsgenerierung die Antwortzeiten gering bleiben. Vor allem gegenüber der ein- oder zweidimensionalen Betrachtungsweise sollte keine Verzögerung entstehen, da sich dies unter anderem negativ auf die Akzeptanz der Anwender auswirken könnte.
- *Flexible Berichtsgenerierung*: Die Mehrdimensionalität der Auswertungen soll sich auch in den Berichten zeigen. Die Gruppierung und der Vergleich unterschiedlicher Teile des Modells müssen möglich sein.[1] Die Möglichkeit zum Abruf vordefinierter Standardberichte sollte gegeben sein.[2]

Anforderungen an die Dimensionsverwaltung

- *Einheitliche Struktur und Funktionalität der Datendimensionen*: Die verschiedenen Dimensionen des Datenmodells sollten in einheitlicher Form in Bezug auf Struktur und Funktionalität aufgebaut sein. Der Einheitlichkeit sollte gegenüber der Erweiterung der Funktionalität nur bestimmter Dimensionen der Vorzug gegeben werden.
- *Unbeschränkte Durchführung dimensionsübergreifender Operationen*: Eine weitere Forderung ist die nach der Möglichkeit, Daten von unterschiedlichen Dimensionsebenen miteinander zu verknüpfen (z.B. Quartalsdaten mit monatlichen Auswertungen). Für diesen Zweck sollten für den Benutzer transparente Ableitungsregeln hinterlegt sein.
- *Unbegrenzte Anzahl von Verdichtungsebenen und Dimensionen*: Die Forderung nach einer unbegrenzten Anzahl von Dimensionen in den Datenmodellen scheitert meist an Hard- und Softwaregegebenheiten. Die Anzahl der Verdichtungsebenen ist hingegen tatsächlich meist unbegrenzt.[3]

[1] vgl. Totok, 2000, S.59

[2] vgl. Kirchner, 1998, S.154

[3] vgl. Totok, 2000, S.57ff.

Anforderungen bezüglich physikalischer Aspekte

- *Client-Server-Architektur:* Die Client-Server-Architektur stellt die beste Methode zur Realisierung von OLAP dar, denn die Anforderungen an die Leistungsfähigkeit der Systeme sind sehr hoch.
- *Mehrbenutzerunterstützung:* Mehrere Benutzer sollen – lesend wie schreibend – Zugriff auf den OLAP-Server haben, dies erfordert geeignete Sicherheitskonzepte und die Synchronisation paralleler Transaktionen.
- *Dynamische Verwaltung unvollständig besetzter Matrizen:* Nicht jede Kombination unterschiedlicher Dimensionen enthält tatsächlich Werte. Bei der physikalischen Speicherung von Daten entsteht dadurch ungenutzter Speicherplatz an solchen Stellen. In diesem Fall wird von dünn besiedelten Matrizen gesprochen. Vertreibt beispielsweise ein Unternehmen nicht in allen Regionen dieselben Produkte, entstehen Lücken bei den Umsatzkennzahlen. Das OLAP-System soll dem Rechnung tragen und für eine effiziente und rationale Speicherung sorgen, selbst wenn sehr viele Zellen unbesetzt sind.[1]

Typische betriebliche Anwendungen mit rechnerischen und grafischen Analysen im Rahmen von OLAP zeigt die folgende Auflistung von Lusti:[2]

Finanz- und Rechnungswesen

- Kurzfristige Erfolgsrechnung
- Jahresabschlussanalyse
- Cash Flow-Analyse
- Kennziffernanalyse

Absatz

- Soll/Ist-Vergleiche
- Produkt- und Kundenvergleiche
- Qualität des Kundendienstes

[1] vgl. Kirchner, 1998, S.154f.

[2] Lusti, 1999, S.148

Beschaffung

- Bestandsanalysen
- Lieferfristenüberwachung

Produktion

- Kapazitätsanalysen
- Qualitätskontrolle

Personalwesen

- Personalverwaltung
- Mitarbeiterqualifikation

4.2 OLAP - Operationen

In OLAP-Systemen steht meist eine Vielzahl an Werkzeugen zur Analyse von Daten oder auch zur Simulation zur Verfügung. Für die meisten Benutzer sind die folgenden Funktionen besonders wichtig:

- *Drill Down*: Mit dieser Funktion können Informationen aus immer kleineren Detailebenen analysiert werden. Beispielsweise können die Verkaufszahlen eines bestimmten Produktes zunächst in einem bestimmten Land, mittels Drill Down können dann einzelne Bundesländer, Regionen und schließlich einzelne Filialen betrachtet werden.[1]

- *Roll Up:* Mittels Roll Up erreicht man das Gegenteil eines Drill Down. Informationen werden auf diese Weise verdichtet, von untergeordneten Ebenen wird zur nächst höheren Betrachtungsebene gewechselt, z.B. von der Betrachtung der Monate auf Quartale.

[1] vgl. Hönig, 1998, S.172f.

- *Drill Across*: Durch einen Drill Across können die Dimensionen auf der x-Achse und der y-Achse miteinander vertauscht werden. Dadurch erhält man beispielsweise aus einer Auswertung der Artikelumsätze nach Monaten eine Monatsauswertung nach Artikeln.[1]

- *Slice and Dice*: Durch diese Funktionen können bestimmte Ausschnitte aus dem Datenmodell „geschnitten“ werden. In jedem funktionalen Bereich des Unternehmens kann so die relevante Informationsebene betrachtet werden.[2]

Abbildung 3 zeigt die wichtigsten OLAP-Operationen im Überblick anhand der Auswertung von Verkaufszahlen.

[1] vgl. Totok, 2000, S.62

[2] vgl. Hönig, 1998, S.171

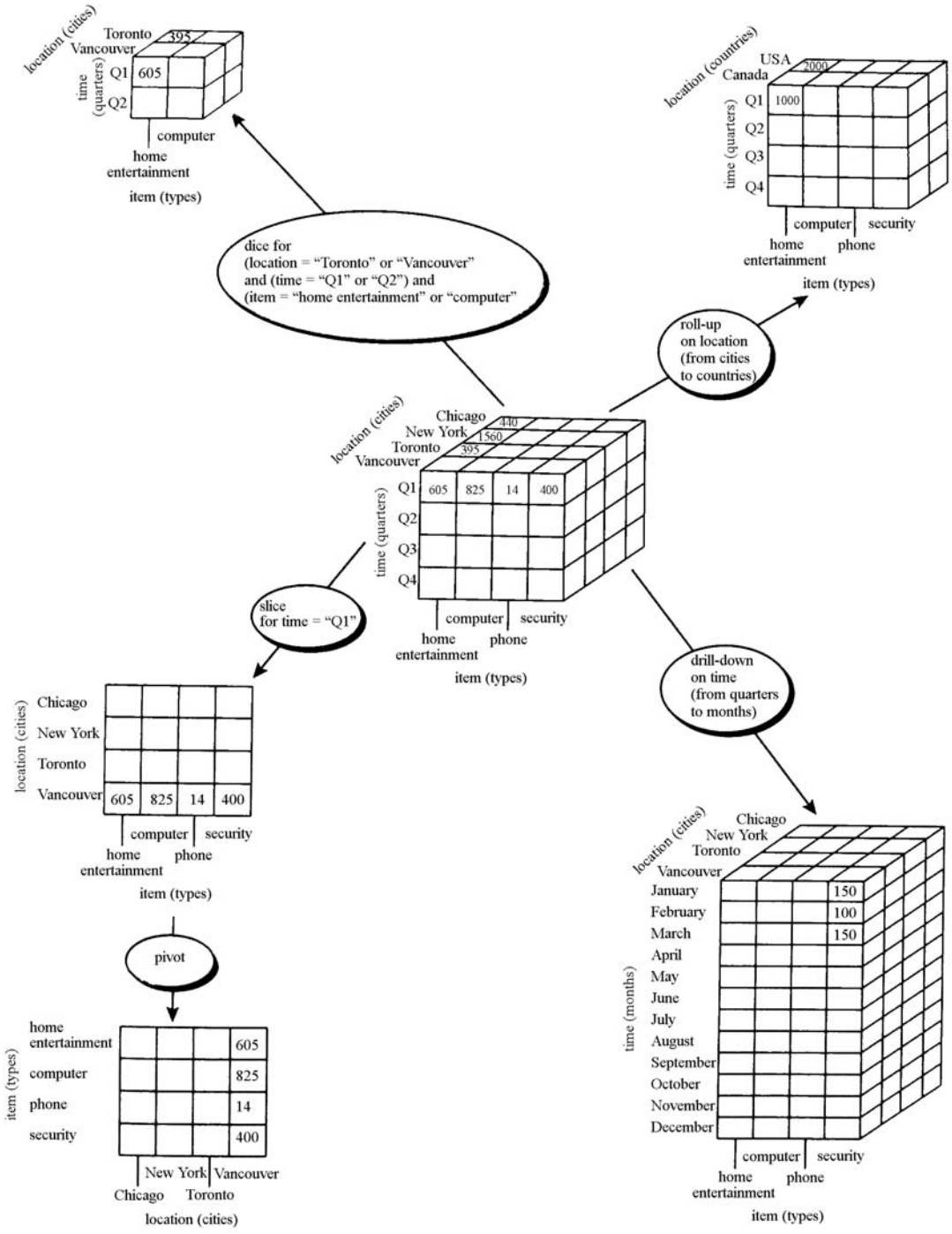


Abbildung 3: Überblick über die OLAP-Operationen

(Quelle: Han und Kamber, 2001, S.59)

5 Knowledge Discovery in Databases

5.1 Grundlagen und Begriffe

In der Literatur finden sich sehr unterschiedliche Definitionen von Knowledge Discovery in Databases (KDD). Vielfach werden Begriffe wie *Data Mining, Knowledge Extraction, Information Discovery, Knowledge Discovery* oder *Pattern Processing* synonym verwendet, was eine Definition zusätzlich erschwert. Die prominentesten Begriffe in der Literatur sind Data Mining und KDD. Die begrifflichen Unterschiede könnten darin begründet sein, dass KDD und Data Mining in mehreren Wissenschaftsdisziplinen ihre Wurzeln haben. Einige Definitionszugänge sind methodenorientiert, andere wiederum orientieren sich an den Problemen und Fragestellungen, die gelöst werden können.[1]

Fayyad et al. definieren KDD wie folgt: *„Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."* [2]

Die einzelnen Begriffe werden von Fayyad et al.[3] näher präzisiert:

- *Entdeckung von Mustern*: Daten sind eine Menge von Fakten F (z.B. Geschäftsfälle in einer Datenbank). Ein Muster ist ein Ausdruck E in einer Sprache L, der Fakten in einer Teilmenge F_e von F beschreibt. Bei E handelt es sich dann um ein Muster, wenn es in gewisser Weise einfacher ist als die Aufzählung aller Fakten in F. Diese bewusst sehr allgemeine Definition eines

[1] vgl. Knobloch, B.: Der Data-Mining-Ansatz zur Analyse betriebswirtschaftlicher Daten. Bamberg: Otto-Friedrich-Universität 2000, S.1f.

[2] vgl. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U.M. et al. (Hrsg.). Advances in knowledge discovery and data mining. AAAI Press/MIT Press 1996, S.6

[3] vgl. Fayyad et al. 1996, S.6f.

Musters schließt jede Beziehung zwischen Datensätzen, Datenfeldern, Daten innerhalb eines Satzes oder bestimmte Regelmäßigkeiten ein.

- *Nicht-Trivialität*: Dieser Aspekt bezieht sich auf die Datengetriebenheit der Analyse. Der Entdeckungsprozess ist nicht-trivial und erfordert daher ein gewisses Maß an Suchautonomie.
- *Gültigkeit*: Die aufgedeckten Muster sollen den Inhalt der Datenbasis beschreiben. Die Kenntnis der Gültigkeit ist wichtig zur Bestimmung des Vertrauens, das in die Analyseergebnisse gesetzt werden kann.
- *Neuartigkeit*: die Muster waren bislang unbekannt. Der Grad der Neuheit kann durch Vergleiche mit historischen oder prognostizierten Daten bestimmt werden.
- *Potenzielle Nützlichkeit*: Die Muster sollen zur Erreichung von Zielen der Anwender beitragen bzw. in nutzbringende Maßnahmen umgesetzt werden können.
- *Verständlichkeit*: Die gefundenen Muster sollen so aufbereitet werden, dass sie für den Menschen verständlich sind.

Die Ansicht, dass KDD einen mehrstufigen Prozess zur Entdeckung von neuem Wissen darstellt, hat sich weitgehend durchgesetzt. Data Mining ist demnach der Auswertungsschritt im übergeordneten KDD-Prozess. Die von verschiedenen Autoren vorgeschlagenen Prozessmodelle weisen meist eine ähnliche Struktur auf, Unterschiede gibt es vor allem hinsichtlich der Anzahl der Teilschritte.[1]

[1] vgl. Säuberlich, F.: KDD und Data Mining als Hilfsmittel zur Entscheidungsunterstützung. Wien: Lang 2000, S.9

5.2 Die einzelnen Schritte im KDD-Prozess

Runkler[1] unterteilt den Prozess in vier Stufen und listet einige Beispiele für Tätigkeiten auf, die in der jeweiligen Stufe erfolgen, was in **Abbildung 4** dargestellt wird.

Vorbereitung		**Vorverarbeitung**		**Mustererkennung**		**Nachbereitung**
Planung		Normalisieren		Korrelation		Interpretation
Daten-	➡	Säubern	➡	Regression	➡	Dokumentation
sammlung		Filtern		Modellierung		Auswertung
Merkmals-		Ergänzen		Klassifikation		
generierung		Korrigieren		Entscheidungsbäume		
Datenauswahl		Transformieren		Clusteranalyse		

Abbildung 4: Ablauf des Auswertungsprozesses
(Quelle: Runkler, 2000, S.2)

Fayyad et al.[2] - ähnlich wie Knobloch[3] - schlagen eine Untergliederung des KDD-Prozesses in fünf Teilschritte vor, die in der **Tabelle 2** dargestellt sind.

	Fayyad et al. (1996)	**Knobloch (2000)**
1.	Selection	Selektion
2.	Preprocessing	Exploration
3.	Transformation	Manipulation
4.	Data Mining	Analyse
5.	Interpretation/Evaluation	Interpretation

Tabelle 2: Darstellung der KDD-Prozesschritte

[1] vgl. Runkler, 2000, S.2

[2] vgl. Fayyad et al. 1996, S.10

[3] vgl. Knobloch, 2000, S.27

In Abhängigkeit vom Wissensgebiet, in dem KDD angewendet wird und in Abhängigkeit von der Komplexität der Datenbestände entscheidet sich, welche Schritte im KDD-Prozess mehr Zeit und Aufwand beanspruchen und welche weniger. Der gesamte Prozess kann mehrfach durchlaufen werden, bis ein befriedigendes Resultat erreicht wird, was **Abbildung 5** veranschaulicht.

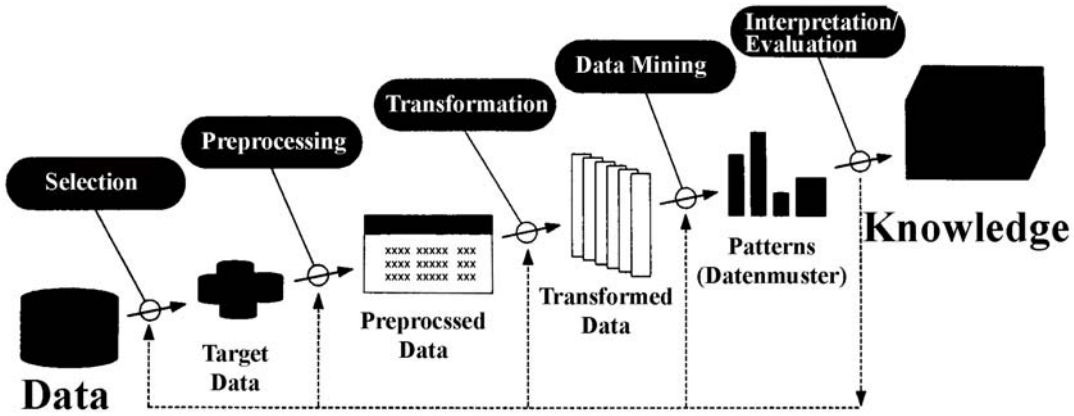


Abbildung 5: Die einzelnen Schritte des KDD-Prozesses

(Quelle: Fayyad et al. 1996, S.10)

5.2.1 Selektion der Daten

Den ersten Schritt im KDD-Prozess stellt die Auswahl der für die Beantwortung der Fragestellung relevanten Daten dar. Der Erfolg jeder Datenanalyse ist eng verknüpft mit der Qualität des verwendeten Datenmaterials. Im günstigsten Fall stammen die Daten aus einem Data Warehouse, dessen Bestände in der Regel bereits eine hohe Qualität aufweisen. Vielfach ist der Rückgriff auf externe Datenbestände oder operative Datenbanken erforderlich.

Werden längere Untersuchungszeiträume beobachtet, ist darauf zu achten, dass die Datenbestände in dieser Zeit keinen Veränderungen unterworfen wurden, die zu Ergebnisverfälschungen führen könnten.

Die Auswahl der Datensätze erfolgt in Hinblick auf die Fragestellung, schwieriger gestaltet sich Selektion relevanter Attribute. Vor der Analyse ist nicht bekannt, welche Attribute sich tatsächlich auf das Ergebnis auswirken. Die Einbeziehung möglichst aller vorhandenen Attribute hat den Vorteil, dass durch Data Mining der Einfluss bislang vernachlässigter Variablen deutlich wird. Allerdings können auch Sättigungseffekte auftreten, da sich die Genauigkeit durch Aufnahme immer neuer Attribute nicht beliebig steigern lässt.

Vielfach werden vor der eigentlichen Analyse Stichproben gezogen, um mit den Daten vertraut zu werden, Fehler aufzudecken und Verfahren zu testen. Wichtig dabei ist, dass die Stichprobe den gesamten Datenbestand möglichst gut zu repräsentieren vermag.[1]

5.2.2 Exploration der Daten

Mit Hilfe der Exploration der Daten vor der eigentlichen Datenanalyse soll ein Verständnis für die Daten entwickelt werden. So können Fehler und Mängel in den Daten frühzeitig erkannt werden und spätere Fehlinterpretationen vermieden werden. Häufig werden statistische Kennzahlen, tabellarische Auflistungen oder Verteilungsinformationen ermittelt, um die Struktur des Datenbestandes besser sichtbar zu machen.

Die Exploration hat die Aufgabe, Syntax und Semantik der Daten auf Korrektheit zu testen, Plausibilitätsprüfungen sollten durchgeführt werden. Eine gute Kenntnis des Datenmaterials ist auch bei der Auswahl von Zielvariablen im Data Mining von Vorteil. Die Gefahr, Zielgrößen durch Variablen erklären zu lassen, von denen die Zielvariable kausal abhängig ist, kann so ausgeschaltet werden. Die Exploration für sich kann schon interessante Ergebnisse liefern, beispielsweise können Mängel in den die Daten generierenden Systemen entdeckt und Verbesserungen der betrieblichen Systeme angeregt werden.

[1] vgl. Knobloch, 2000, S.29f.

Eine gründliche Exploration hat den Vorteil, dass dadurch die Interpretation der Data Mining Verfahren erleichtert wird, vor allem was die Bewertung der Relevanz, Gültigkeit und Plausibilität der Ergebnisse betrifft.

Die Explorationsphase steht in engem Zusammenhang mit der Manipulationsphase, die sorgfältige Exploration der Daten kann den Aufwand und die Anzahl notwendiger Manipulationen deutlich verringern. Aufgrund der vielfältigen Aufgaben sollte für die Exploration der Daten ein angemessenes Zeitbudget reserviert werden.[1]

5.2.3 Manipulation der Daten

Die Manipulation bzw. Datenvorverarbeitung ist erforderlich, da die verfügbaren Datenbestände in ihrer Ursprungsform häufig nicht unmittelbar für Data Mining-Methoden geeignet sind oder zu fehlerhaften Ergebnissen führen können.

Die möglichen Probleme, die es zu bereinigen gilt, beziehen sich auf Verfügbarkeit, Inhalt und Qualität sowie Repräsentation des Datenmaterials. Unter dem Aspekt der Datenverfügbarkeit sind vor allem fehlende Datensätze und -felder zu behandeln, diesem Problem kann durch Anreicherung begegnet werden. Eine unzureichende Qualität der Inhalte der Daten kann zu invaliden Mustern beim Data Mining führen, insbesondere unsichere, ungenaue, fehlende oder fehlerhafte Werte, Redundanzen sowie semantisch inkonsistente Daten. Durch Bereinigungsprozeduren und Konsolidierung kann die notwendige Qualität erreicht werden. Probleme der Datenrepräsentation ergeben sich aus unterschiedlichen Formaten und Darstellungsformen, vor allem wenn die Daten aus mehreren Quellen extrahiert werden. Durch verschiedene Transformationsprozesse können die Daten vereinheitlicht werden. Falls ein längerer Untersuchungszeitraum beobachtet wird, muss gewährleistet sein, dass keine gravierenden Änderungen in der Datenerfassung erfolgten.

[1] vgl. Knobloch, 2000, S.30f.

Fehlen notwendige Informationen, einzelne Attribute oder ganze Datensätze, ist eine Anreicherung des Datenbestandes erforderlich. Dies kann durch die Integration anderer im Unternehmen vorhandener Informationen, die Aufnahme externer Daten oder Berechnung aus den bestehenden Daten geschehen.

Durch Bereinigung sollen Verzerrungen des Analyseergebnisses durch Datenfehler weitgehend ausgeschlossen werden. Das sind beispielsweise aufgrund nicht verfügbarer Werte nicht gefüllte Felder (Null-Werte), beliebige Werte, die aufgrund Unkenntnis der tatsächlichen Werte als Platzhalter eingefügt werden (z.B. 01.01.1900 als Geburtsdatum) und „verrauschte“ Werte, die unsicher oder ungenau sind. Das Fehlen von Information kann auch wichtige Informationen enthalten, ebenso wie Ausreißer Hinweise auf bestimmte Sachverhalte liefern können. Daher sollte vor der Eliminierung bzw. Bereinigung stets geprüft werden, wie die Fehler und Abweichungen entstanden sein könnten. Sind die Datenfehler einmal identifiziert, können sie meist durch automatisierte Bereinigungsprozeduren behoben werden.

Als Konsolidierung wird die Beseitigung von Inkonsistenzen und redundanten Informationen verstanden. Beispielsweise tritt häufig der Fall ein, dass ein Kunde durch mehrere Datensätze mit verschiedenen Primärschlüsseln in den Stammdatenbeständen vertreten ist. Dies ist auch sehr häufig der Fall, wenn Daten aus mehreren Quellen zusammengespielt werden. Eine weitere Aufgabe der Konsolidierung ist die Angleichung unterschiedlicher Datenformate bei der Nutzung mehrerer Datenquellen, z.B. werden im einen Fall für das Geschlecht die Werte {0,1} und im anderen die Kategorien {m,w} verwendet.

Durch Transformation werden die Daten in die für die jeweiligen Verfahren des Data Mining erforderlichen Darstellungsformen und Formate übergeführt. Das kann beispielsweise die Übersetzung von Textinformation in eindeutige Schlüssel oder Codes oder die Einschränkung von Wertebereichen zur Verringerung der Anzahl zu betrachtender Ausprägungen bedeuten. Numerische Werte müssen häufig standardisiert oder skaliert werden.

Einige typische Transformationsprozesse sind:

- Verallgemeinerung von Adressinformationen auf Regionen oder Bezirke
- Darstellung des Geburtsdatums durch Berechnung des Alters
- Skalierung von Einkommenswerten durch Division mit dem Faktor 1000
- Transformation binärer kategorialer Merkmalswerte in die {1,0}-Form
- Fortlaufende Nummerierung von Zeitintervallen

Es ist davon auszugehen, dass die Auswahl, Exploration und Vorverarbeitung bis zu 80% der gesamten Projektressourcen beanspruchen, da die Qualität der gesamten Auswertung maßgeblich von diesen Schritten beeinflusst wird.[1]

5.2.4 Analyse der Daten

Nach den umfassenden Vorbereitungsarbeiten kann die eigentliche Analyse, die Auswahl und Durchführung der geeigneten Data Mining Methode, erfolgen. Selten kann auf Anhieb eine zufriedenstellende Verfahrenskonfiguration gefunden werden. Häufig werden mehrere Verfahren erprobt. Bei überwachten Methoden besteht die Gefahr, dass zufällige Phänomene, die in der Datenbasis enthalten sind, verallgemeinert werden. Diese Überanpassung des Modells wird als *Overfitting* bezeichnet. Daher werden bei überwachten Verfahren ein Trainingsdatensatz und ein Testdatensatz gebildet, um die optimale Trainingsintensität zu ermitteln.[2]

5.2.5 Interpretation der Ergebnisse

Der abschließende Schritt im KDD-Prozess besteht in der Interpretation der ermittelten Muster und Beziehungen. Zunächst muss geprüft werden, ob die gefundenen Muster den Anforderungen der Gültigkeit, Neuartigkeit, Nützlichkeit und Verständlichkeit genügen. Die Muster, die insignifikant, trivial, bereits bekannt

[1] vgl. Knobloch, 2000, S.31ff.

[2] vgl. Knobloch, 2000, S.37f.

sind, von denen kein Nutzen zu erwarten ist oder unverständlich und nicht nachvollziehbar sind, müssen gefiltert werden.

Die verbleibenden Ergebnisse sollen interpretiert und in konkrete Maßnahmen umgesetzt werden. Die Bildung eines Teams von Experten kann am ehesten gewährleisten, dass die Bewertung korrekt erfolgt und die gewonnen Erkenntnisse der optimalen Nutzung zugeführt werden. Häufig ergeben sich aus den Ergebnissen neue Fragestellungen und der gesamte KDD-Prozess beginnt von vorne oder es erfolgt zumindest ein Rücksprung auf eine frühere Stufe, was das iterative Vorgehen der Analyse nochmals verdeutlicht.[1]

[1] vgl. Knobloch, 2000, S.39ff.

6 Data Mining

6.1 Grundlagen und Begriffe

Der englische Begriff Data Mining wird mit *Datenmustererkennung* übersetzt und stellt den Auswertungsschritt im übergeordneten KDD-Prozess dar. Es gibt sehr unterschiedliche Sichtweisen, welche Methoden unter diesem Begriff subsumiert werden, da Ansätze aus der klassischen Statistik, dem maschinellen Lernen und der künstlichen Intelligenz zur Anwendung kommen.[1]

6.2 Charakteristische Eigenschaften des Data Mining

Es stellt sich die Frage, welche Eigenschaften Data Mining von anderen Ansätzen abhebt. Im Gegensatz zur klassischen Statistik erfolgt die Analyse oft datengetrieben und hypothesenfrei. Die statistischen Parameter in der klassischen Statistik werden aus dem ganzen Datensatz errechnet und die Signifikanzniveaus werden sofort ausgegeben. Der datengetriebene Zugang im Data Mining erfordert eine andere Vorgehensweise. Der Datensatz wird unterteilt, d.h. mit unterschiedlichen Methoden werden Regeln oder Muster in einem Testdatensatz ermittelt, die Gültigkeit wird in einem abgetrennten Datensatz überprüft.

Im Data Mining werden häufig – im Gegensatz zu anderen Verfahren – die Datenbestände automatisch mit verschiedenen Suchalgorithmen durchforstet, d.h. es wird keine bestimmte Datenbankabfrage durchgeführt, sondern die Verfahren laufen ständig im Hintergrund. Dadurch können Auffälligkeiten entdeckt werden. Diese Techniken werden beispielsweise in Industrieprozessen eingesetzt oder zur Überwachung von Kreditkartentransaktionen.

---

[1] vgl. Fayyad et al. 1996, S.4

Data Mining wird meist auf sehr große Datenbestände angewendet, die ansonsten schwer zu handhaben sind, da einfache Abfragen aufgrund der Größe der Datensätze oft Stunden an Rechnerleistung beanspruchen.

Die beschriebenen Eigenschaften sind weder völlig trennscharf noch erschöpfend, besonders hervorzuheben ist das datengetriebene, hypothesenfreie Vorgehen bei vielen Verfahren des Data Mining.

Nach Krahl et al. wird die vermehrte Nutzung von Data Mining durch folgende Entwicklungen unterstützt:[1]

- Die Menge an elektronisch gespeicherter Information nimmt exponentiell in fast allen Wirtschafts- und Verwaltungsbereichen zu.
- Die Etablierung von Data Warehouses (Konsolidierung von Datenbeständen, Aufbau von Historien) erleichtert komplexe Analysevorgänge.
- Die Leistungssteigerung der Rechnersysteme wird den hohen Anforderungen an rechnerische Analyseprozesse gerecht, gleichzeitig sind die Hardwarepreise (Prozessor- und Speicherpreise) kontinuierlich gesunken.
- Data Mining-Software ist in der Lage, sehr große Datenbestände zu verarbeiten.
- Immer mehr Unternehmen können Data Mining als umfassende Dienstleistung anbieten.
- Nach Ausschöpfung der sonstigen Möglichkeiten bietet Data Mining noch Möglichkeiten zur Kostenreduzierung oder Ergebnisverbesserung.

[1] vgl. Krahl, D. et al.: Data Mining: Einsatz in der Praxis. Bonn: Addison-Wesley-Longman 1998, S.25

6.3 Zielstellungen des Data Mining

Mit Data Mining können sehr unterschiedliche Fragestellungen in vielen verschiedenen Disziplinen bearbeitet werden. In diesem Kapitel werden die wichtigsten Zielstellungen erläutert und durch mögliche – überwiegend betriebliche Anwendungsbeispiele – ergänzt. Einen Überblick über die Zielstellungen und Methodenbeispiele zeigt **Tabelle 3**.

Ziel	Aufgabe	Methodenbeispiele
Klassifikation	Individuen/Variablen bekannten Klassen zuordnen	Entscheidungsbäume, Neuronale Netze
Segmentierung	Identifikation von Gruppen aufgrund von Ähnlichkeiten zwischen Individuen/Variablen	Neuronale Netze, Clusteranalyse
Prognose	Berechnung zukünftiger Werte aus unabhängigen Variablen	Neuronale Netze, Regression
Assoziation	Abhängigkeiten aufdecken und quantifizieren	Statistische Zusammen-hangsanalyse
Abweichungsanalyse	Identifikation von auffälligen Werten/Ausreißern	Boxplots, Verteilungs-analysen
Text Mining	Textmuster suchen	Suchalgorithmen
Visualisierung	Visuelle Darstellung von Mustern/Zusammenhängen	Dreidimensionale Streudiagramme

Tabelle 3: Aufgaben und Zielstellungen des Data Mining

(Quelle: Eigene Darstellung in Anlehnung an Lusti, 1999, S.252)

6.3.1 Klassifikation

Die Aufgabe von Klassifikationsverfahren ist die Zuordnung von Objekten zu vorgegebenen bzw. bekannten Klassen. Die Verfahren in diesem Bereich stammen zum überwiegenden Teil aus der Statistik und dem maschinellen Lernen. Aus der Statistik stammen z.B. die Diskriminanzanalyse und das K-Nächste-Nachbarn-Verfahren. Methoden aus dem maschinellen Lernen sind Neuronale Netze, Entscheidungsbaumverfahren und regelbasierte Verfahren.[1]

Aus historischen Daten, in denen die Klassenzugehörigkeit der Objekte bekannt ist, ermitteln die Verfahren Regeln und Eigenschaften, die die Klassenzugehörigkeit erklären können. Da die Klassenzugehörigkeit in den historischen Daten bekannt ist, handelt es sich um überwachte Verfahren.

Klassifikationsaufgaben sind sehr häufig in der Unternehmenswelt. Da die Gewinnung von Neukunden sehr teuer ist, sind die Unternehmen bestrebt, möglichst viele Stammkunden zu halten. Telekommunikationsunternehmen möchten beispielsweise verstehen, welche Kunden am ehesten geneigt sind, den Anbieter zu wechseln und welche Kunden dem Unternehmen am ehesten die Treue halten werden.[2] Im Bankensektor ist ein zentrales Thema die Beurteilung der Kreditwürdigkeit potenzieller Kunden. Durch Klassifikation werden Scoring-Modelle entwickelt, die zum Rating der Kunden eingesetzt werden. Ziel ist die Minimierung der Forderungsausfälle.[3] Zum Teil werden Klassifikationsregeln auch zur Prognose von Aktienkursen eingesetzt.[4]

---

[1] vgl. Säuberlich, 2000, S.44

[2] vgl. Groth, R.: Data Mining: A Hands-On Approach for Business Professionals. New Jersey: Prentice Hall 1998, S.4f.

[3] vgl. Küppers, 1999, S.142f.

[4] vgl. Knobloch, 2000, S.17

6.3.2 Segmentierung

Ziel der Segmentierung ist die Unterteilung einer großen Datenmenge in kleinere, homogene und zweckmäßige Teilmengen. Dieser Vorgang wird auch als *Clustering* bezeichnet. Die Objekte innerhalb einer resultierenden Teilmenge sollen einander möglichst ähnlich (Anforderung der Homogenität), Objekte aus unterschiedlichen Klassen hingegen einander möglichst unähnlich (Anforderung der Heterogenität) sein. Anwendungen von Clusteranalysen finden sich in den unterschiedlichsten Bereichen wie z.B. Biologie, Chemie, Geologie, Soziologie, Psychologie oder Marketing. Die meisten Clusteranalysen gehen in der Regel von Datenmatrizen aus. Die Zeilen enthalten die einzelnen Fälle bzw. Beobachtungen und die Spalten die Variablen (Merkmale, Eigenschaften). Grundsätzlich ist zwischen überwachten und unüberwachten Verfahren zu unterscheiden. Bei überwachten Verfahren ist die Klassenzugehörigkeit der Objekte bereits bekannt, dies kann sowohl für die Generierung als auch die Evaluierung von Klassifikatoren verwendet werden.

Häufiger sind unüberwachte Verfahren, d.h. weder eventuell vorhandene Cluster noch deren Anzahl sind bekannt. Nach der Durchführung eines unüberwachten Verfahrens ist zu prüfen, welche der gefundenen Klassen als gesichert angesehen werden können und wie die Güte der Ergebnisse bewertet werden kann. Das Ergebnis unüberwachter Verfahren kann in weiterer Folge mittels überwachter Verfahren geprüft werden.[1]

In Unternehmen spielt die Segmentierung vor allem im Marketing und in der Produktentwicklung eine große Rolle. Im Marketing können Data Minig Verfahren herkömmliche Methoden der Marktforschung unterstützen, um Kunden mit ähnlichen Bedürfnissen zusammenzufassen. Vor allem wenn es um die Identifizierung sehr kleiner Teilsegmente geht, bietet Data Mining Vorteile.
Die Bearbeitung bestimmter Kundengruppen wird dadurch verbessert, aber auch die

[1] vgl. Grimmer, U., Mucha, H.-J.: Datensegmentierung mittels Clusteranalyse. In: Nakhaeizadeh, G. (Hrsg.): Data Mining: Theoretische Aspekte und Anwendungen. Heidelberg: Physica-Verlag 1998, S.109ff.

Produktentwicklung kann davon profitieren. Weitere Ziele sind beispielsweise der Ausbau bestehender Kundenbeziehungen oder die Verhinderung von Abwanderungen von Kunden.[1]

6.3.3 Prognose

Die möglichst genaue Abschätzung von Entwicklungen und Trends ist der Zweck der Prognoseverfahren. Im Gegensatz zur Klassifikation sind die Zielvariablen nicht kategorisch sondern numerisch skaliert. Die Prognosen werden meist aus historischen Daten abgeleitet. Einige Modelle besitzen die Fähigkeit, Eingabevariablen zu variieren bzw. konstant zu halten und so verschiedene Szenarien zu simulieren.[2]

Typische Methoden zur Prognose aus der Statistik sind die einfache und die multiple Regressionsanalyse und das Box-Jenkins-Verfahren. Es werden aber auch neuronale Netze eingesetzt. Werden die stetigen Zielvariablen in Intervalle zusammengefasst und als Klassenwerte verwendet, können auch typische Klassifikationsverfahren wie Entscheidungsbäume eingesetzt werden.[3]

Die unternehmerischen Anwendungen von Prognoseverfahren sind mannigfaltig, z.B. zur Abschätzung zukünftiger Absatz- und Umsatzzahlen oder Kursentwicklungen.

---

[1] vgl. Küppers, 1999, S.126ff.

[2] vgl. Groth, 1998, S.21f.

[3] vgl. Nakhaeizadeh et al.: Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick. In: Nakhaeizadeh, G. (Hrsg.): Data Mining: Theoretische Aspekte und Anwendungen. Heidelberg: Physica Verlag 1997, S.18

6.3.4 Assoziation und Verknüpfung

Vielfach ist es interessant, die Beziehungen zwischen Objekten näher zu analysieren. Sind bestimmte Beziehungen zwischen einer Menge von Objekten häufiger zu beobachten, entstehen Verknüpfungsmuster. Betriebswirtschaftliche Anwendungen sind vor allem im Bereich Marketing und Vertrieb anzutreffen. Auf Basis des Kaufverhaltens der Kunden in der Vergangenheit können maßgeschneiderte Produkt- und Leistungsangebote erstellt werden. Ein typisches Anwendungsgebiet abseits der Wirtschaft ist die Verknüpfungsanalyse in der Kriminalistik, die der Aufklärung von Verbrechen dient. Die Analyse von Assoziationen oder Abhängigkeiten geht über die Betrachtung der Verknüpfungsstruktur hinaus. Es wird untersucht, welche Richtung die Abhängigkeiten aufweisen. Damit kann aufgezeigt werden, welche Größen voneinander abhängen und die Stärke der Abhängigkeit wird quantifiziert.[1]

Gefundene Beziehungen in den Daten werden oft in Form von „Wenn-dann-Regeln“ ausgedrückt. Ein prominentes Beispiel für die Aufdeckung von Abhängigkeiten ist die Warenkorbanalyse. Eine typisches Ergebnis einer Warenkorbanalyse könnte lauten: „Wenn Kunden einen Hammer kaufen, kaufen sie mit der Wahrscheinlichkeit von X auch Nägel.“ Da es gerade im Handel zu einer außerordentlich hohen Anzahl von Assoziationen kommt, ist die Einführung von Einschränkungen oder Aggregationen notwendig. Beispielsweise könnten alle Arten von Hämmern zur Kategorie „Hammer“ zusammengefasst werden oder nur Assoziationen ermittelt werden, die eine bestimmte Stärke aufweisen.[2]

Über die Assoziationsanalysen hinausgehende Möglichkeiten ergeben sich, wenn die Ereignisse im Zeitablauf bestimmten Zeitpunkten zugeordnet werden. Dadurch ist die Analyse der daraus resultierenden Sequenzen möglich. So wird beispielsweise das Kaufverhalten von Versandhandelskunden im Zeitablauf untersucht oder die

---

[1] vgl. Knobloch, 2000, S.18f.

[2] vgl. Two Crows, Introduction to Data Mining and Knowledge Discovery. Dritte Ausgabe. Potomac: Two Crows Corporation 1999, S.7f.

Erkennung von Betrugsdelikten bei der Kreditkartennutzung erleichtert.[1]

Assoziationen werden häufig visualisiert, wobei dickere Linien starke Verbindungen aufzeigen und dünnere Linien auf schwache Assoziationen hindeuten (**Abbildung 6**).

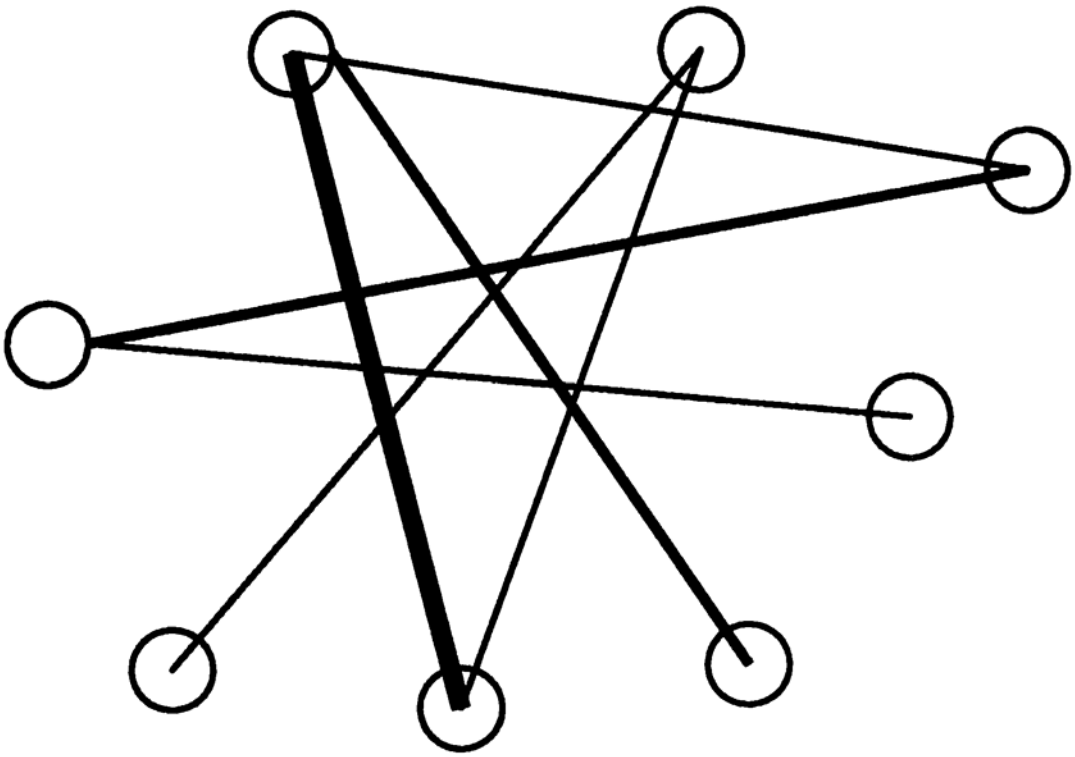


Abbildung 6: Visualisierung von Assoziationen

(Quelle: Two Crows, 1999, S.7)

[1] vgl. Knobloch, 2000, S.19

6.3.5 Abweichungsanalyse

Ziel von Abweichungsanalysen ist die Entdeckung von signifikanten Abweichungen in Datensätzen. Die Abweichungen können sich auf Normwerte oder früher gemessene Werte beziehen.[1] Die Analyse von Abweichungen kann Aufschluss über Probleme mit den Daten geben. Dies hat oft eine Bedeutung in der Vorbereitung weiterer Analysen, da die Unstimmigkeiten in den Daten so rechtzeitig bereinigt werden können. Weiters können Abweichungen auf notwendige Änderungen in einem bestehenden Modell hinweisen.[2]

Abweichungsanalysen (auch deviation detection genannt) dienen aber nicht nur der Aufdeckung von Fehlern. Abweichende Objekte – oft auch als Ausreißer bezeichnet – können fehlerfreie, interessante Eigenschaften aufweisen. Die Analyse und Interpretation der Ausreißer fördert oft neue und interessante Phänomene zu Tage, die ansonsten unentdeckt blieben.[3]

Das Ausmaß der Abweichung kann numerisch mit statistischen Methoden ermittelt werden, z.B. können alle Werte angezeigt werden, die außerhalb eines bestimmten Konfidenzintervalls liegen. Sehr hilfreich sind in diesem Zusammenhang Methoden der Visualisierung, die Auffälligkeiten oft rascher sichtbar machen.[4]

Abweichungsanalysen spielen eine wichtige Rolle in der Qualitätskontrolle, Fehler können rasch identifiziert und behoben werden. Im Absatzbereich können starke Umsatzschwankungen entdeckt und analysiert werden. Auch in der Aufdeckung und Bekämpfung von Betrugsfällen haben Abweichungsanalysen einen hohen Stellenwert. Vor allem Versicherungsunternehmen und Kreditkartenfirmen können davon profitieren.

---

[1] vgl. Fayyad et al. 1996, S.16

[2] vgl. Nakhaeizadeh et al. 1997, S.10

[3] vgl. Schommer, C.: Anwendung von Data Mining. Aachen: Shaker 2003, S.25

[4] vgl. Nakhaeizadeh et al. 1997, S.19

6.3.6 Text Mining

Zwischen Data Mining und Text Mining bestehen einige Analogien: Ziel ist die Analyse großer Datenmengen, die eingesetzten Verfahren stammen aus unterschiedlichen Disziplinen und es kommt zur Anwendung von überwachten und unüberwachten Verfahren.

Große Datenmengen sind in Form von Text gespeichert und elektronisch verfügbar, z.B. im Internet oder auch in Intranets vieler Unternehmen. Das Ziel des Text Mining besteht darin, diese Informationsquellen einer maschinellen und inhaltlichen Analyse zugänglich zu machen.[1]

Grundsätzlich wird zwischen dokumentbasierten Techniken und inhaltsbasierten Techniken unterschieden. Dokumentbasierte Techniken arbeiten mit einer ganzen Dokumentenkollektion, die Details eines einzelnen Dokuments spielen dabei eine untergeordnete Rolle. Mit Hilfe dieser Techniken können Texte kategorisiert (Automated Text Categorization) oder Dokumente gleichen Inhalts gebündelt werden (Clustering). Weitere Aufgabenstellungen sind das Suchen von Dokumenten (Document Retrieval) oder die Visualisierung von Zusammenhängen zwischen Dokumenten einer Kollektion. Inhaltsbasierte Techniken dienen der Analyse eines spezifischen Dokuments und nicht einer ganzen Kollektion. Diese Verfahren dienen beispielsweise der automatischen Zusammenfassung von Texten (Document Summarization) oder der automatischen Erkennung der Sprache, in der ein Dokument geschrieben wurde.[2]

[1] vgl. http://www.kuenstliche-intelligenz.de/Thema/Text-Mining.htm, 08.03.2005

[2] vgl. http://www.ie.iwi.unibe.ch/forschung/km/textmining.php, 08.03.2005

6.3.7 Visualisierung

Visualisierungstechniken sind nicht als eigener Ansatz des Data Mining zu verstehen, sie sollen vielmehr die Darstellung und Interpretation von Ergebnissen unterstützen bzw. erleichtern. Zusätzlich besteht die Möglichkeit für den Anwender, Auffälligkeiten durch Visualisierung zu entdecken, die aus den Zahlen nicht unmittelbar „ins Auge stechen“.[1]

Hinsichtlich der Qualität werden drei Anforderungen an die Visualisierung gestellt:[2]

- Effektivität: Für die Visualisierung von Daten/Informationen soll die beste Methode gewählt werden.
- Expressivität: Die Daten sollen durch die Visualisierung nicht verfälscht werden.
- Angemessenheit: Der Nutzen für die Visualisierung soll die Kosten dafür rechtfertigen.

Bei den neueren Visualisierungsmethoden wird versucht, Zusammenhänge in drei Dimensionen darzustellen z.B. mit Hilfe unterschiedlicher Farben und Formen. Bei einer statischen Darstellung bereitet es allerdings oft Schwierigkeiten, die räumliche Lage einzelner Punkte richtig zu identifizieren. Abhilfe bieten Visualisierungen in Form von dreidimensionalen Landschaften, durch die sich der Anwender navigieren und unterschiedliche Visualisierungsarten wählen kann.[3]

[1] vgl. Küppers, 1999, S.70

[2] vgl. Schommer, 2003, S.27

[3] vgl. Schommer, 2003, S.28

6.4 Methoden des Data Mining

Für die in Kapitel 6.3 beschriebenen Zielstellungen kommen sehr viele Methoden in Frage, einige methodische Ansätze können für mehrere Aufgabenstellungen eingesetzt werden. Beispielsweise dienen neuronale Netze sowohl der Segmentierung als auch der Klassifizierung. In diesem Kapitel sollen einige der in der Literatur am häufigsten beschriebenen Ansätze näher erläutert werden.

6.4.1 Neuronale Netze

Biologische neuronale Netze bestehen aus Milliarden von Neuronen. Jedes Neuron erhält Informationen von anderen Neuronen oder der Umwelt, verarbeitet diese und leitet Informationen an andere Neuronen weiter. Die Grundidee der neuronalen Netze ist es, die Erkenntnisse aus den Naturwissenschaften auf Computerarchitekturen und -algorithmen zu übertragen.

Das Lernen aus Fehlern und die Verallgemeinerung des Gelernten auf neue Stichproben stellt das Grundprinzip neuronaler Netze dar. Die Lernmenge enthält bekannte Eingaben und Ausgaben. Die Lernmethoden basieren auf Erkenntnissen über biologische Nervensysteme. Die kleinste Lerneinheit bildet das Neuron. Das Neuron arbeitet nach dem EVA-Schema eines Prozessors (Eingabe, Verarbeitung, Ausgabe). Ein neuronales Netz verbindet Neuronen numerisch, indem jedes Neuron aus mehreren Eingaben eine einzige Ausgabe an mehrere andere Neuronen berechnet. Die Verbindungen zwischen den einzelnen Neuronen sind unterschiedlich stark und bekommen daher eine unterschiedliche Gewichtung. Die Gewichte zwischen den Neuronen werden so lange angepasst, bis der Unterschied zwischen dem berechneten und dem tatsächlichen Wert minimal wird. Beim eben beschriebenen Vorgehen handelt es sich um überwachtes Lernen, da in der Lernmenge neben den Eingaben die korrekten Ausgaben bekannt sind. Lernt das System hingegen unüberwacht, ist nur eine Lernmenge von Eingaben gegeben. Einstufige neuronale Netze bestehen aus Eingaben und Ausgaben, die über eine einzige Gewichtsstufe verbunden sind.

Mehrstufige neuronale Netze bestehen aus Eingabe- und Ausgabeneuronen und verborgenen Neuronen, sie sind durch mehr als eine Gewichtsstufe miteinander verbunden. **Abbildung 7** zeigt die schematische Darstellung eines neuronalen Netzes.

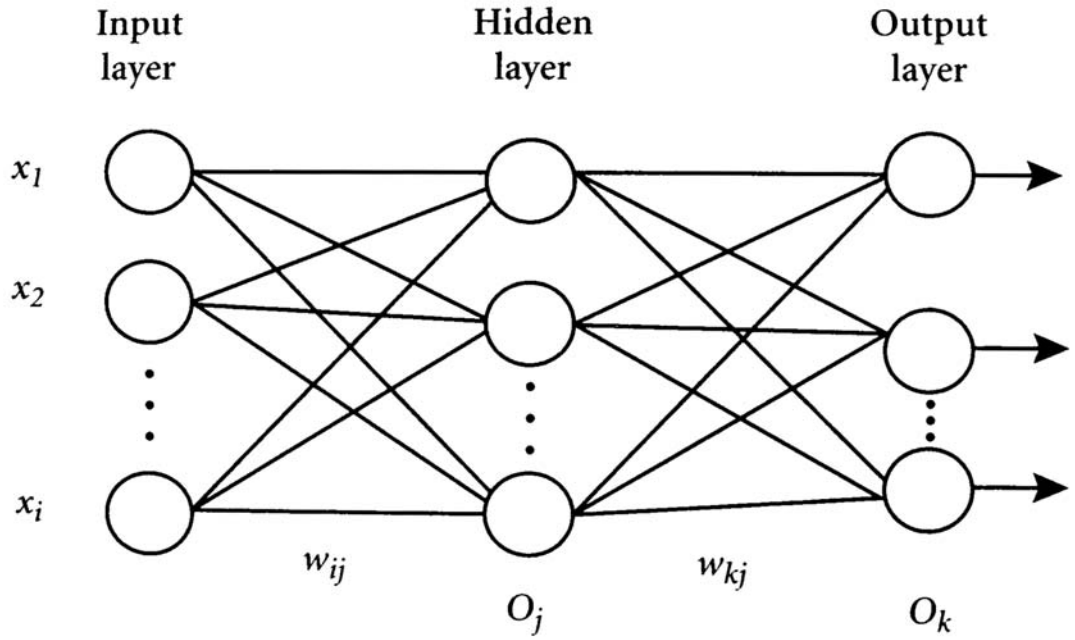


Abbildung 7: Mehrstufiges neuronales Netz

(Quelle: Han und Kamber, 2001, S.304)

Neuronale Netze eignen sich vor allem für die Klassifikation und Vorhersage von Werten und Optimierungsaufgaben. Betriebliche Anwendungen sind beispielsweise die Optimierung von Direct Mailing-Aktionen oder die Bonitätsbeurteilung von Kunden.[1]

[1] vgl. Lusti, 1999, S.306ff.

6.4.2 Entscheidungsbaumverfahren

Entscheidungsbäume dienen der Zuordnung von Objekten zu bekannten Klassen und werden seit den 60er Jahren eingesetzt. Mit Hilfe von Entscheidungsbäumen entstehen Klassifikationsregeln, die einen Zusammenhang zwischen mehreren unabhängigen Variablen und einem abhängigen Merkmal (der Klassenzugehörigkeit) herstellen. Die Entscheidungsbaumverfahren zählen zu den überwachten Verfahren, da die Klassenzugehörigkeit im Trainingsdatensatz in der Lernphase bekannt ist.

Zur Bestimmung von Entscheidungsbäumen wurden viele verschiedene Algorithmen entwickelt, das Durchführungsprinzip ist allen ähnlich: In einem Trainingsdatensatz, in dem die interessierende Klassenzugehörigkeit bekannt ist, wird ein Entscheidungsbaum aufgebaut, der die Klassenstruktur der Trainingsdaten bestmöglich repräsentieren soll. Zunächst wird, ausgehend von der gesamten Trainingsdatenmenge, ein Knoten eingefügt, der die Daten in zwei (oder mehr) Gruppen teilt. Die Auswahl des Knotens orientiert sich am maximalen Informationsgewinn. Dann wird die interessierende Teilmenge der Testdaten in weitere Teilmengen unterteilt, denen neue Knoten des Baumes zugewiesen werden. Auf diese Weise verzweigt sich der Baum solange weiter, bis ein bestimmtes Abbruchkriterium erfüllt ist und die zuletzt ermittelten Knoten zu Endknoten werden.

Zur Überprüfung der Güte des ermittelten Baumes wird er auf Testdaten angewendet, in denen das abhängige Merkmal ebenfalls bekannt ist. Daraus wird ersichtlich, wie häufig die Klassifikationsregeln des ermittelten Baumes zum richtigen Ergebnis führen bzw. wie hoch die Fehlerrate ist.[1]

Die Algorithmen zur Bestimmung des Baumes unterscheiden sich dadurch, wie die Knoten bestimmt werden und wie das Abbruchkriterium festgelegt ist. Schon bei einer relativ geringen Anzahl von Dimensionen können sehr komplexe Entscheidungsbäume entstehen, die schwer interpretierbar sind. Ziel sollte aber ein möglichst einfacher Baum sein, der eine Entscheidungshilfe darstellt und trotz

[1] vgl. Säuberlich, 2000, S.79f.

geringer Komplexität und einer akzeptablen Fehlerrate befriedigende Resultate liefert.[1] Ein Beispiel für einen sehr einfachen Entscheidungsbaum zur Klassifikation von Kunden zeigt **Abbildung 8**.

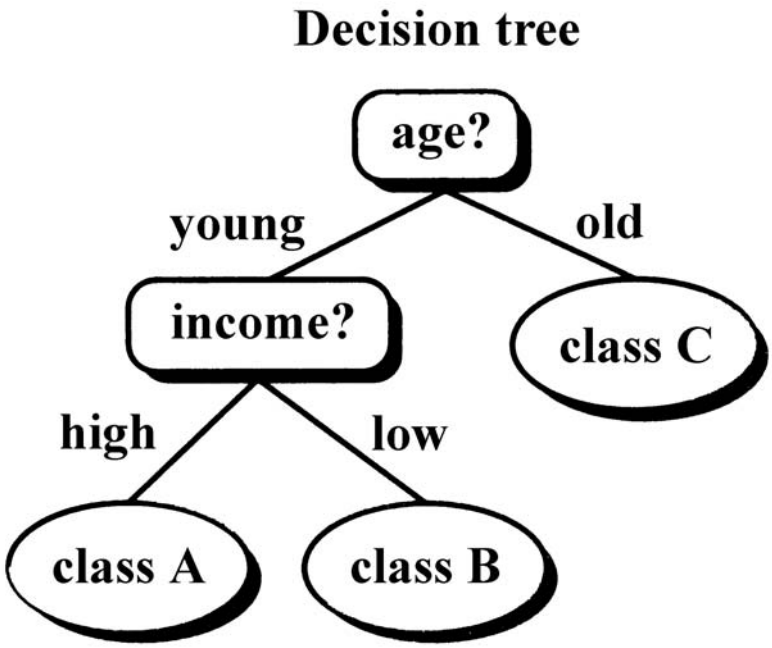


Abbildung 8: Beispiel für einen einfachen Entscheidungsbaum

(Quelle: Han und Kamber, 2001, S.158)

6.4.3 Clusteranalyse

Clusteranalysen dienen der Gruppierung von nominal und metrisch skalierten Daten, wobei eine Vielzahl von Algorithmen für diesen Zweck entwickelt wurde. Ziel ist, in sich möglichst ähnliche Gruppen zu bilden (Homogenitätsprinzip), die Gruppen untereinander sollen aber möglichst verschieden sein (Heterogenitätsprinzip).

Die Verfahren laufen grundsätzlich in zwei Schritten ab: Im ersten Schritt wird jeweils paarweise die Nähe (Proximitätsmaß) zwischen zwei Datensätzen bestimmt.

[1] vgl. Küppers, 1999, S.56f.

Bei metrisch skalierten Daten werden sowohl der relative als auch der absolute Abstand als Messgrößen herangezogen. Bei nominal skalierten Daten wird geprüft, ob zwei Datensätze gleich sind. Anhand dieser Berechnungen wird eine sogenannte Distanzmatrix gebildet, in der die jeweiligen Abstände eingetragen sind. Im zweiten Schritt erfolgt mit Hilfe der Distanzmatrix die Gruppenbildung, wobei dafür verschiedene Varianten zur Verfügung stehen:

- Hierarchische Methoden: Beim Top-Down-Ansatz (divisiv) werden die Daten zunächst in einer einzigen Gruppe zusammengefasst, die dann schrittweise in kleinere Gruppen aufgeteilt wird. Beim Bottom-Up-Ansatz (agglomerativ) wird zunächst jeder Datensatz als eigene Gruppe betrachtet und vereinigt die jeweils ähnlichsten Gruppen, bis ein Abbruchkriterium erreicht ist. Agglomerative hierarchische Methoden werden häufig eingesetzt, weil sie leicht zu implementieren sind und die Rechenzeit günstiger ausfällt als bei den anderen Methoden.
- Partitionierende Methoden: Sie verfolgen den gegenteiligen Ansatz der hierarchischen Methoden. Ausgehend von vorgegebenen Gruppenzugehörigkeiten werden die Datensätze so lange verschoben, bis die Kriterien an Homogenität innerhalb und Heterogenität zwischen den Gruppen erfüllt sind.
- Überlappende Methoden: Ein Datensatz kann mehreren Gruppen zugeordnet werden, wobei für die Zugehörigkeit ein Maß berechnet wird. Bei großen Überschneidungen bereitet die Interpretation der Ergebnisse allerdings meist Probleme.[1]

Abbildung 9 zeigt das Ergebnis einer Clusteranalyse zur Segmentierung von Kunden nach bevorzugten Fahrzeugtypen.

[1] vgl. Küppers, 1999, S.70f.

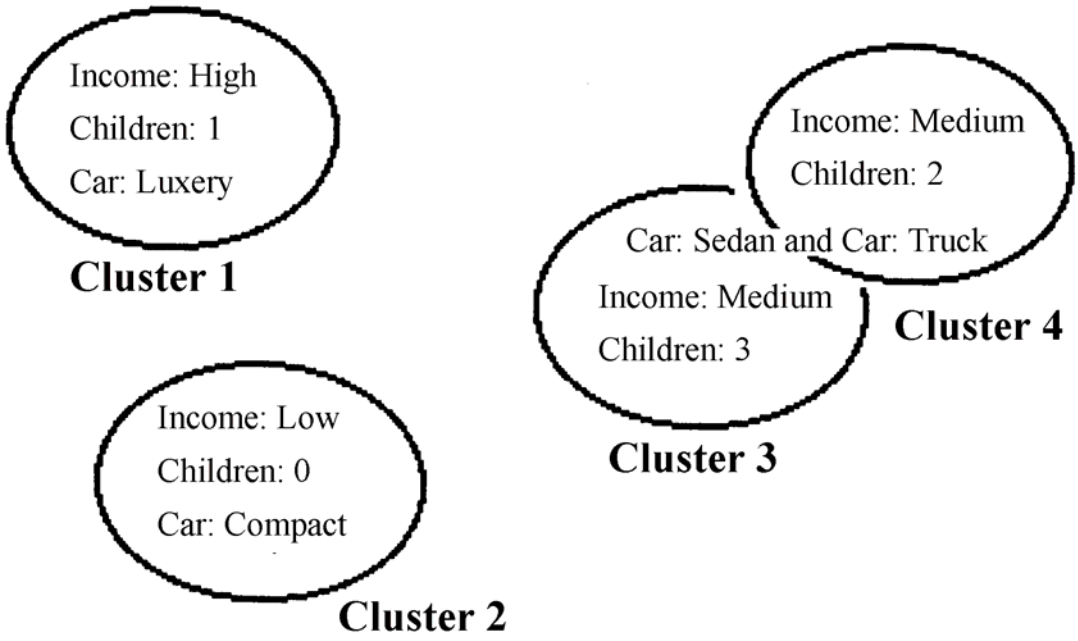


Abbildung 9: Ergebnis einer Clusteranalyse

(Quelle: Groth, 1998, S.8)

6.4.4 Bayes-Verfahren

Mit dem Bayes-Verfahren werden Objekte zu unterschiedlichen Klassen gruppiert. Im Gegensatz zur Clusteranalyse ist nicht die Ähnlichkeit der Objekte entscheidend, sondern die Zugehörigkeitswahrscheinlichkeit eines Objekts zu einer bestimmten Klasse, die aus dem gesamten Datenbestand errechnet wird. Die Zugehörigkeitswahrscheinlichkeit gibt an, mit welcher Wahrscheinlichkeit ein Objekt sich in der jeweiligen Klasse befindet. Jedes Objekt wird dort zugeordnet, wo die maximale Zugehörigkeitswahrscheinlichkeit berechnet wurde.

Die Zugehörigkeitswahrscheinlichkeit wird nach dem Theorem von Bayes als bedingte Wahrscheinlichkeit berechnet:[1]

$$P_z = P(k|E) = \frac{P(E|k) * P(k)}{\sum_k P(E|k) * P(k)}$$

P(k|E) Wahrscheinlichkeit, dass ein Objekt mit den Eigenschaften E aus der Klasse k kommt

P(k) Wahrscheinlichkeit, dass ein Objekt aus der Klasse k kommt, ohne Berücksichtigung seiner Eigenschaften

P(E|k) Wahrscheinlichkeit, dass die Eigenschaften E des Objekts in der Klasse k vorkommen

Die Objekte müssen nicht fest zu einer Klasse zugeordnet werden, sondern können auch mit unterschiedlichen Wahrscheinlichkeiten verschiedenen Klassen zugeordnet werden.

Grundsätzlich sind zwei Verfahrenstypen zu unterscheiden, die nach dem Bayes-Theorem arbeiten. Im ersten Fall ist das gesamte Datenmaterial vollständig vorhanden und für die Software verfügbar. Im zweiten Fall werden laufend neu hinzukommende Daten berücksichtigt, dadurch kann es zur Bildung neuer Klassen kommen bzw. zu einer anderen Zuordnung der Objekte, da die Zugehörigkeitswahrscheinlichkeiten laufend neu berechnet werden. Bayes-Verfahren erfordern einen hohen Arbeits- und Zeitaufwand, führen aber meist zu zuverlässigen Ergebnissen.[2]

[1] vgl. Mertens, P., Wieczorrek, H.W.: Data X Strategien: Data Warehouse, Data Mining und perationale Systeme für die Praxis. Berlin: Springer Verlag 2000, S.218

[2] vgl. Mertens, Wieczorrek , 2000, S.217f.

6.4.5 Fallbasiertes Schließen

Fallbasiertes Schließen oder Case-Based Reasoning (CBR) verfolgt das Ziel, Probleme auf der Grundlage analoger Fälle aus der Vergangenheit zu lösen. Eine Falldatenbank wird nach ähnlichen Fällen durchsucht, um eine Lösung zu finden. Findet sich kein ähnlicher Fall, muss ein neuer Lösungsansatz entwickelt werden, der bei erfolgreicher Durchführung in die Datenbank aufgenommen wird. Einige Ansätze unterteilen Fälle nicht nur in „Problem“ und „Lösung“, sondern in eine Kette von Informationseinheiten (sog. snippets). Schritt für Schritt wird ein Fall mit Informationseinheiten angereichert und gilt dann als gelöst, wenn genügend übereinstimmende Informationseinheiten vorhanden sind. Die Systeme verarbeiten in der Regel sowohl nominal als auch metrisch skalierte Daten. Die Ähnlichkeit der Probleme wird häufig mit Clusteranalysen oder dem Nearest-Neighbour-Ansatz geprüft.

Das fallbasierte Schließen beruht auf der Annahme, dass kleine Änderungen in der Problemstellung auch nur zu kleinen Änderungen bei der Problemlösung führen. Bei komplexen, nichtlinearen Systemen kann allerdings nicht davon ausgegangen werden, dass dieses starke Kausalitätsprinzip gilt, d.h. für unstrukturierte Probleme sind CBR-Ansätze nicht geeignet. Die Leistungsfähigkeit des fallbasierten Schließens hängt stark davon ab, wie die Fälle in der Datenbank gespeichert sind und wie die Ähnlichkeit gemessen wird. Ein grundlegendes Problem liegt darin, dass die Systeme kaum Hinweise auf die Qualität der vorgeschlagenen Lösung geben, die Bewertung der Lösungsvorschläge bleibt dem Anwender überlassen.

Sehr häufig werden CBR-Ansätze in Call Centers bzw. Help-Desk-Anwendungen eingesetzt, um internen und externen Kunden bei der Lösung von Problemen mit Produkten und Dienstleistungen zu helfen.[1]

[1] vgl. Küppers, 1999, S.59f.

6.5 Data Mining Software

Auf dem Markt ist eine Vielzahl an Data-Mining-Produkten verfügbar und es ist nicht möglich, im Rahmen dieser Arbeit einen umfassenden Überblick darzustellen. Die Programme unterscheiden sich hinsichtlich Umfang der Funktionalitäten, hinterlegten Algorithmen, verarbeitbarer Datenmenge, Benutzerfreundlichkeit, im Preis etc. Einige Anwendungen sind auch frei verfügbar z.B. unter www.kdnuggets.com/software/index.html.

Die folgenden Informationen stammen aus einem Übersichtsartikel von Haughton et al. und geben beispielhaft einen Einblick in die Kosten für fünf Programme und deren Eigenschaften. Eine grobe Einteilung kann in drei Gruppen erfolgen: „Stand alone“-Lösungen, „Add-On“-Programme und Business Intelligence Lösungen.

„Stand-alone“-Programme sind einzelne Programme, die für Data Mining eingesetzt werden. Das Programm GhostMiner kostet zwischen $2.500 und $30.000 in der Anschaffung, die jährlichen Wartungsgebühren sind nicht angegeben. Die Lizenzkosten für die Software Quadstone betragen zwischen $200.000 und $1.000.000, hinzu kommen Wartungsgebühren, die von der Anzahl der User und der Anzahl der analysierten Kunden abhängt. Beide Programme sind mit der Möglichkeit für statistische Standardauswertungen (Mittelwert, Standardabweichung, Minimum, Maximum, Anzahl fehlender Werte) ausgestattet. Quadstone verfügt über die überlegenen Visualisierungswerkzeuge und kann eine weitaus höhere Anzahl unabhängiger Variablen verarbeiten als GhostMiner. Quadstone bietet außerdem mehr Möglichkeiten zur Datenmodellierung, was angesichts des Preisunterschieds nicht verwundert.

Eine „Add-on“-Variante stellt der XLMiner dar, der als Plattform MS-Excel nutzt. Eine Vollversion kostet $899 und ist damit preislich sehr attraktiv. Für mit Excel vertraute User ist die Bedienungsfreundlichkeit sehr hoch. Für statistische Auswertungen und Grafiken kann Excel genutzt werden. Der XLMiner eignet sich gut für die Berechnung von Assoziationsanalysen und Entscheidungsbäumen. Für große Datensätze ist der XLMiner nicht geeignet und stellt damit eher für kleinere Unternehmen eine interessante Alternative dar.

Große Business-Intelligence-Lösungen werden beispielsweise von SAS und SPSS angeboten. Die Lizenzkosten für den SAS Enterprise Miner liegen zwischen $119.000 und $281.000, für SPSS-Clementine liegen keine Zahlen vor. Beide verfügen über umfangreiche statistische Auswertungs- und Visualisierungs-möglichkeiten. Auch die Data Mining-Anwendungen sind in diesen beiden Packages am komplettesten.

Zusammenfassend wird dem SAS-Enterprise Miner das beste Zeugnis hinsichtlich Umfang verfügbarer Funktionen und Benutzerfreundlichkeit ausgestellt, nur bezüglich der Visualisierungsmöglichkeiten ist die Software SPSS-Clementine überlegen.[1]

[1] Houghton, D. et al. A Review of Software Packages for Data Mining. The American Statistician **57** (4) 2003: S.290-309

7 Datennutzung im Krankenversicherungsbereich

Während im allgemeinen Teil der Diplomarbeit die unterschiedlichen Strategien und Methoden zur Nutzung von Daten im Rahmen von OLAP, KDD und Data Mining beschrieben werden, soll dieses Kapitel wichtige Anwendungsgebiete im Krankenversicherungsbereich beleuchten, aber auch die damit verbundenen Probleme aufzeigen.

Die überwiegende Anzahl der methodischen Ansätze stammen aus den USA. Das Krankenversicherungswesen der USA ist sehr komplex und durch mehrere geschichtliche Entwicklungen geprägt. Ein wesentliches Merkmal ist die stärkere Verzahnung von Leistungserbringung und Bezahlung im Gesundheitswesen. Neben den staatlichen Versicherungen ist der Anteil privater Versicherungen sehr hoch, daher ist der Versicherungsmarkt durch starke Konkurrenz geprägt. Vor allem aus ökonomischen Gesichtspunkten wurden daher Instrumente wie Case Management oder Disease Management entwickelt. Um diesen Hintergrund näher zu betrachten, wird zunächst die wichtigste Versicherungsart in den USA – die Managed-Care-Versicherung dargestellt. Mehr als 70% der Amerikaner mit Versicherungsschutz sind Mitglied einer MC-Versicherung.[1]

7.1 Managed Care in den USA

7.1.1 Definition von Managed Care

Managed Care ist ein vielschichtiger Begriff und kann nicht eindeutig definiert werden. Grundsätzlich ist zwischen institutionellen und funktionellen Aspekten zu unterscheiden, das heißt zwischen Managed Care als Prozess und Managed Care als

[1] vgl. Lehmann, H.: Managed Care: Kosten senken mit alternativen Krankenversicherungsformen? Zürich: Rüegger 2003, S.29

Organisation.[1]

Rognehaugh definiert Managed Care aus Prozesssicht wie folgt:
„Any method of healthcare delivery designed to reduce unnecessary utilization of services, and provide for cost containment while insuring that high quality of care or performance is maintained“. [2]

Für Managed Care als Organisation betrachtet gibt Rognehaugh folgende Definition:
„Arrangements made by payers to promote cost-effective healthcare through establishing selective relationships with healthcare providers, developing, coordinated or integrated delivery systems, and conducting medical management activities“.[3]

Die Organisationen können hinsichtlich mehrerer Merkmale unterschiedlich ausgestaltet sein, was zum Teil in ihrer langen Geschichte begründet ist aber auch in den wenigen gesetzlichen Bestimmungen zur Gründung einer MC-Organisation.

7.1.2 Entwicklung von Managed Care

Die ersten Verträge, die zwischen Arbeitgebern und Ärzten abgeschlossen wurden, in denen sich Ärzte verpflichteten, eine bestimmte Anzahl Personen gegen eine im Voraus festgelegte Summe medizinisch zu versorgen, gehen auf das Jahr 1849 zurück. Die Verbreitung großer Prepaid Group Practices führte zu Auseinandersetzungen. Viele Ärzte waren gegen die Einführung dieser Versorgungsform. In vielen Bundesstaaten wurde die freie Arztwahl vorgeschrieben, was die meisten Formen von Managed Care verunmöglichte. Durch die steigenden Gesundheitskosten gewannen die MC-Versicherungen wieder an Attraktivität bei

[1] vgl. Gaertner, T., Jelastopulu, E., Niehoff, J.-U.: Managed Care in den USA. In: Medizinische Dienste der Krankenversicherung (Hrsg.). Managed Care – Eine Perspektive für die GKV? Stuttgart: Georg Thieme 2000, S.5

[2] Rognehaugh, 1998, zit. nach Lehmann, 2003, S.28

[3] Rognehaugh, 1998, zit. nach Lehmann, 2003, S.28

Regierungen und Behörden. Die Regulierungen wurden nach und nach gelockert. 1973 wurden mit dem HMO-Act alle Anti-MC-Gesetze aufgehoben. Die Regierung stellte Kapital für neue Health Maintenance Organizations (HMO) zur Verfügung. Große Firmen sollten ihren Angestellten eine HMO zur Wahl anbieten. In den 80er Jahren nahm die Verbreitung der MC-Versicherungen schließlich sehr stark zu.[1]

7.1.3 Formen von Managed Care

Die MC-Versicherungen werden in drei Formen eingeteilt:

1. Preferred Provider Organizations (PPO)
2. Independent Practice Associations (IPA)
3. Health Maintenance Organizations (HMO)

Zur Unterscheidung der einzelnen Formen sind folgende Fragen relevant:

- Welche Leistungserbringer sind beteiligt?
- Wer wählt die Leistungserbringer aus?
- Wie werden die Leistungserbringer bezahlt?
- Wie groß ist die Kostenbeteiligung?
- Welche Rolle spielt der Versicherer?
- Wie werden die Leistungen beschränkt?

Es ist zu berücksichtigen, dass die Begriffe Managed-Care-Versicherung und Health Maintenance Organization häufig synonym verwendet werden. Vor allem der Begriff HMO wird häufig als Überbegriff für alle Managed-Care-Formen genutzt.

7.1.3.1 Preferred Provider Organization

PPOs bauen sich ein Netz mit den von ihnen ausgewählten Anbietern auf. Diese Netze umfassen neben Ärzten häufig auch Spitäler, Apotheken, Labors,

[1] vgl. Lehmann, 2003, S.28

Physiotherapeuten und weitere Anbieter medizinischer Dienstleistungen. PPOs versuchen möglichst effiziente und kostengünstige Anbieter in ihr Netz aufzunehmen, um so einen Spareffekt erzielen zu können. Die Einsparungseffekte werden den Versicherten in Form niedriger Prämien weitergegeben. Die Versicherten können auch Leistungen außerhalb des Netzes in Anspruch nehmen, was einen wichtigen Unterschied zu den HMOs darstellt. Finanzielle Begünstigungen haben die Versicherten aber nur dann, wenn sie sich innerhalb des Netzes mit medizinischen Leistungen versorgen. Typisch für PPOs ist, dass die Kostenbeteilungen bei Anbietern innerhalb des Netzes deutlich geringer ausfallen als außerhalb des Netzes. Zur Kostenkontrolle verlangen einige PPOs, dass die Versicherten vor der Inanspruchnahme bestimmter Behandlungen die Erlaubnis dafür einholen.[1] PPOs besitzen keine eigene Versicherungslizenz und dürfen nicht selbständig am Versicherungsmarkt agieren.[2]

7.1.3.2 Independent Practice Associations

Intependent Practice Associations existieren bereits länger als PPOs. Sie bieten ebenfalls ein Netzwerk von medizinischen Leistungserbringern, v.a. Hausärzten an. Innerhalb des Netzes sind die Kostenbeteiligungen wesentlich günstiger als außerhalb. Die beteiligten Ärzte üben im Gegensatz zu den PPOs eine Gatekeeper-Funktion aus. Sie entscheiden nach dem ersten Besuch des Patienten über den weiteren Behandlungsverlauf, sie bestimmen, welche Untersuchungen erforderlich sind etc. Die Ärzte werden entweder nach Einzelleistungen vergütet oder mittels Capitation. Bei der Honorierung mittels Capitation bekommt der Arzt eine pauschale pro-Kopf-Abgeltung und übernimmt dafür die Gesundheitsversorgung der Versicherten für einen bestimmten Zeitraum. Die Höhe der Pauschalen kann mit Merkmalen wie dem Geschlecht, Alter oder anderen Variablen variieren. [3]

Ursprünglich wurden einheitliche Versicherungsprämien eingehoben, die aufgrund

[1] vgl. Lehmann, 2003, S.29f.

[2] vgl. Gaertner, Jelastopulu, Niehoff, 2000, S.15

[3] vgl. Lehmann, 2003, S.30

der durchschnittlichen Pro-Kopf-Ausgaben einer Region (=community rating) ermittelt wurde. Bedingt durch den Wettbewerbsdruck werden den Arbeitgebern inzwischen Tarifangebote geboten, die sich am Risiko bzw. der bisherigen Leistungsinanspruchnahme der Beschäftigten orientieren (=experience rating). Für die Versicherungsprämie erhalten die Versicherten ein definiertes Leistungspaket. Die Arbeitgeber beteiligen sich an den Prämien mit Zuschüssen bis zu 100%.[1]

7.1.3.3 Health Maintenance Organizations

Health Maintenance Organizations stellen die älteste Form von Managed Care dar. Das wichtigste Merkmal von HMOs ist die Integration von Leistungserbringung und Versicherung. Grundsätzlich ist zwischen dem Group Model und dem Staff Model zu unterscheiden. Beim Staff Model sind die Leistungserbringer Angestellte der HMO, im Group Model schließen die HMOs mit einer oder mehreren Gruppenpraxen Verträge ab, die Eigentümer der Praxen sind die Ärzte. In den Gruppenpraxen sind mehrere Fachrichtungen vertreten. Sehr große HMOs betreiben auch eigene Krankenhäuser, die ausschließlich ihren Mitgliedern vorbehalten sind.[2] Die Honorierung erfolgt über einen fixen Lohn oder Pro-Kopf-Pauschalen, das Einkommen der Ärzte ist damit von den erbrachten Einzelleistungen abgelöst. Die HMOs überwachen die Leistungen und Behandlungsweisen der Ärzte sehr genau. Kommt die HMO zu dem Schluss, dass zu verschwenderisch mit Ressourcen umgegangen wird, werden Gespräche geführt oder es kommt sogar zum Ausschluss aus der HMO.[3]

Systematische Analysen oder empirische Studien zur Managed-Care-Praxis liegen kaum vor oder haben den Schwerpunkt auf einer ganz bestimmten Form. Es gibt keine staatlichen Richtlinien hinsichtlich der gesetzlich erlaubten Begrenzung der Leistungen und des Leistungsumfangs. Die Akkreditierungsverfahren für HMOs sind

[1] vgl. Gaertner, Jelastopulu, Niehoff, 2000, S.14

[2] vgl. Gaertner, Jelastopulu, Niehoff, 2000, S.14

[3] vgl. Lehmann, 2003, S.30

nicht streng gesetzlich geregelt.[1] Gaertner et al. listen folgende Nachteile bzw. Gefahren der Managed-Care-Praxis in den USA auf:[2]

- Marktwirtschaftliche Ausrichtung der medizinischen Versorgung (Gewinnmaximierung, Monopolisierung, Verbetrieblichung, Bürokratisierung, Reduktionismus)
- Hohe Verwaltungs- und Transaktionskosten
- Dominanz wirtschaftlicher Interessen bei der freien Vertragsgestaltung
- Systemgestaltung nach ökonomischen Gesichtspunkten
- Fehlen einer Übernahme von Verantwortung gegenüber der Gesellschaft
- Fehlen einer Rechenschaftspflicht gegenüber der Gesellschaft
- Medizinische Unterversorgung großer Teile der Bevölkerung
- Risikoselektion
- Leistungseinschränkungen
- Durch Vergütungsformen bedingte qualitätsmindernde Anreize
- Einschränkung der freien Arztwahl
- Einschränkung der Therapiefreiheit
- Autonomieverlust der Ärzteschaft
- Sanktionierung
- Datenmissbrauch
- Unzufriedenheit von Patienten und Ärzten

Dem gegenüber stehen laut Gaertner et al. folgende Vorteile der Managed-Care-Praxis:[3]

- Potenzial zur Steigerung der Versorgungseffizienz
- Erschließung von Wirtschaftlichkeitsreserven durch Kontrolle, Standardisierung, Einkaufsplanung, Prüfung der Angemessenheit einer Behandlungsentscheidung („peer review“)
- Patientenführung (Koordination des Zugangs zu den einzelnen Versorgungsstufen)
- Umfassendes Informationssystem (Dokumentation und Datentransfer)

[1] vgl. Gaertner, Jelastopulu, Niehoff, 2000, S.17f.

[2] Gaertner, Jelastopulu, Niehoff, 2000, S.19

[3] Gaertner, Jelastopulu, Niehoff, 2000, S.19

Der Überblick über die Health Maintenance Organizations zeigt die Komplexität des gesamten Gesundheitssystems, die marktwirtschaftliche Ausrichtung des US-Gesundheitswesens und die starke Prägung durch Wettbewerb und Kostendruck. Einige der im Folgenden beschriebenen Anwendungen müssen vor diesem Hintergrund kritisch beurteilt werden.

7.2 Disease Management

7.2.1 Definition von Disease Management

Der Begriff Disease Management stammt aus den USA, wo Anfang der 90er Jahre die ersten Programme durchgeführt wurden. Diese wurden von der Pharmaindustrie gesponsert und hauptsächlich zu Marketingzwecken im Gesundheitsmarkt etabliert.[1] Ellrodt gibt folgende Definition für Disease Management:
„Disease Management is a multi-disciplinary approach to care for chronic diseases that coordinates comprehensive care along the disease continuum across healthcare delivery systems".[2]

Allgemein formuliert ist das Ziel von Disease Management die Optimierung der Betreuung von Patienten mit ähnlichem Krankheitsbild. Disease Management wird vor allem bei Krankheitsbildern angewendet, die hohe Kosten verursachen und durch falsche Behandlung und mangelnde Koordination der behandelnden Stellen die Kosten noch zusätzlich erhöht werden. Gleichzeitig verschlechtert sich die Situation für die betroffenen Patienten.

---

[1] vgl. Eichbauer, H., Klaushofer, K.: Disease-Management – Case Management. In: Almer, S., Bencic, W. (Hrsg.). Mittelverwendung versus Mittelverschwendung: Fehl-, Über- und Unterversorgung im Gesundheitswesen. Gesundheitswissenschaften Band 26. Linz: OÖGKK 2004, S.81

[2] Ellrodt, 1997, zit. nach Eichbauer, Klaushofer, 2004, S.81

Wichtig ist das Management der Schnittstellen innerhalb der Versorgungskette. Der Informationsfluss zwischen den beteiligten Institutionen soll optimiert werden, dadurch sollen Doppel- und Falschbehandlungen möglichst ausgeschlossen werden. Es wird versucht, die Behandlung möglichst in ambulanten Einrichtungen durchzuführen, um teure Krankenhausaufenthalte zu vermeiden.[1]

Der Erfolg von Disease Management hängt wesentlich von der Qualität der Datenbanken ab, auf die von den Ärzten zurückgegriffen wird, um standardisierte Behandlungspfade abrufen und auf den jeweiligen Fall anpassen zu können.[2]

7.2.2 Anwendungsbereiche des Disease Management

Die Entwicklung und Implementierung von Disease-Management-Programmen ist mit hohem Aufwand und zunächst auch mit zusätzlichen Kosten verbunden. Daher ist die Anwendung von solchen Programmen nur sinnvoll, wenn folgende Voraussetzungen erfüllt sind:[3]

- Epidemiologische Notwendigkeit: Es handelt sich um eine Volkskrankheit, von der eine entsprechende Anzahl an Personen betroffen ist.
- Potenzial zur Verbesserung der Versorgungslage: Es bestehen Möglichkeiten, die derzeitige Versorgungslage zu verbessern, da Fehl-, Unter- und Überversorgung vorliegen.
- Verfügbarkeit von Behandlungsleitlinien: Das Disease-Management-Programm stützt sich auf Diagnose- und Therapierichtlinien, deren Wirksamkeit nachgewiesen ist und alle relevanten Fachdisziplinen einbindet.
- Hoher finanzieller Aufwand der Behandlung: Die Implementierung eines Programms ist nur dann gerechtfertigt, wenn dadurch tatsächlich Kosten eingespart werden können.

[1] vgl. Lehmann, 2003, S.38

[2] vgl. Lehmann, 2003, S.38

[3] vgl. Eichbauer, Klaushofer, 2004, S.82

Abbildung 10 zeigt zusammenfassend die Elemente eines Disease-Management-Programms.

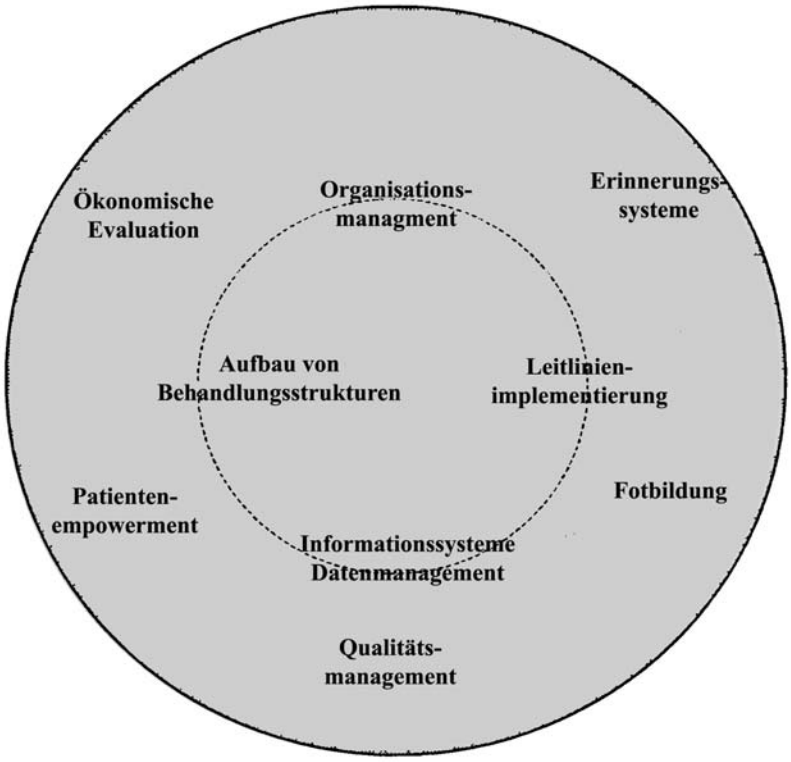


Abbildung 10: Elemente des Disease Management

(Quelle: Pieber, Seereiner, 2004, S.92)

7.2.3 Disease Management und Data Mining

Disease-Management-Programme verfolgen einen präventiven Ansatz, die Bereitstellung von Daten für die Behandlung und ein entsprechendes Schnittstellenmanagement im Behandlungsverlauf spielen dabei eine wichtige Rolle.

Um den präventiven Ansatz noch stärker zu verfolgen, gehen einige Krankenversicherungen nun einen Schritt weiter. Anhand historischer Behandlungsverläufe wird mit Data-Mining-Methoden analysiert, welche Faktoren

zu einer Verschlechterung des Krankheitszustandes führen oder Krankenhausaufenthalte erforderlich machen. Mittels prädiktiver Modelle wird eine Risikobewertung durchgeführt, die demografische Daten und den bisherigen Behandlungsverlauf berücksichtigt. Letztlich sollen „high risk-high costs"-Patienten identifiziert und in spezielle Programme übernommen werden. In den Programmen wird versucht, die Patienten im Umgang mit ihrer Krankheit zu schulen und hinsichtlich ihres Lebensstils zu beraten und damit ihren Gesundheitszustand zu verbessern.[1]

Zu diesem Zweck haben einige Firmen eigene Call Center eingerichtet, ein persönlicher Berater ruft wöchentlich beim betreffenden Patienten an und gibt entsprechende Anweisungen bzw. beantwortet Fragen.

Auf der Homepage des Unternehmens SAS wird von der Implementierung eines Disease-Management-Programms für die Health Management Corporation (HMC) berichtet. Ziele des Programms sind die Optimierung der Behandlung von Menschen mit chronischen Erkrankungen (Diabetes, Asthma, Herzerkrankungen) und die gleichzeitige Einsparung von Kosten. Der Hintergrund für das große Interesse an derartigen Programmen ist die Erkenntnis, dass ein relativ geringer Prozentsatz an Patienten mit chronischen Erkrankungen einen hohen Anteil der entstehenden Kosten verursacht.

Mit der Software-Lösung wird versucht, Patienten zu identifizieren, bei denen ein hohes Risiko besteht, dass sich ihr Gesundheitszustand verschlechtert und dadurch kostspielige Interventionen zu erwarten sind. Die Patienten mit dem höchsten Risiko werden eingeladen, sich an einem Programm zu beteiligen. Der Verschlechterung des Gesundheitszustandes soll durch persönliche Beratung entgegengewirkt werden.

Die Beratung erfolgt über ein Call Center, das 24 Stunden erreichbar ist. Jedem Patienten wird ein persönlicher Berater zur Seite gestellt, der mindestens einmal pro Woche mit dem Patienten Kontakt aufnimmt. Durch das Programm sollen einerseits das Gesundheitsverhalten und damit die Lebensqualität der Menschen positiv

[1] vgl. Gillespie, G.: Data Mining: Solving Care, Cost Capers. Health Data Management **12** (11) 2004: S.52-60

beeinflusst werden und andererseits hohe Kosten vermieden werden. Es wird beispielsweise darauf hingearbeitet, die Ernährung zu ändern, das Gewicht zu senken oder mit dem Rauchen aufzuhören.

Zur Risikobeurteilung werden die Daten von 370.000 Menschen ausgewertet, monatlich erfolgt eine neuerliche Auswertung. Dabei kommen neuronale Netze zur Anwendung. Mehr als 100 Variablen fließen in die Berechnungen ein.
Berücksichtigt werden beispielsweise folgende Faktoren:

- Anzahl der Arztbesuche
- Häufigkeit der Krankenhausaufenthalte
- Medizinische Befunde
- Inanspruchnahme von Heilmitteln
- Soziodemografische Variablen wie Alter und Geschlecht

Das individuelle Risikoprofil entscheidet über Aufnahme und Intensität des Programms. HMC gibt an, dass die Genauigkeit der prädiktiven Modelle in den vergangenen drei Jahren verdoppelt werden konnte. Die erzielten Einsparungen durch das Disease-Management-Programm werden mit 11% beziffert.[1]

7.3 Case Management

Unter Case Management wird die Betreuung von kostenintensiven und mehrere Behandlungsbereiche betreffenden Einzelfällen verstanden. Disease Management hingegen konzentriert sich auf die Verbesserung der Betreuung von Patienten mit dem gleichen Krankheitsbild. Die Übergänge sind allerdings oft fließend und nicht klar voneinander abzugrenzen. Case Management kann auch als Teil des Disease Management verstanden werden.

In der Regel erfolgt Case Management durch eine eigene Instanz, die für die Überwachung und Steuerung dieser Fälle zuständig ist. Ziel ist die Begleitung der

[1] http://www.sas.com/success/hmc.html, 04.08.2004

einzelnen Fälle während des gesamten Behandlungsablaufs in der Behandlungskette, damit die Behandlung medizinisch und ökonomisch optimiert werden kann. Häufig findet Case Management im Krankenhaus statt, da sich hier die kostenintensiven Fälle konzentrieren. Case Management erfolgt aber auch sehr häufig durch eine externe Stelle, z.B. durch die Krankenversicherung. Sind die Leistungserbringer nicht in das Case Management eingebunden, empfinden die Behandler eine starke Kontrolle, was sich negativ auf die Kooperationsbereitschaft auswirken kann. Voraussetzung für Case Management ist die Bereithaltung aller Behandlungsdaten, die auf dem aktuellsten Stand sein müssen.[1]

7.4 Evidenzbasierte Medizin und Behandlungsleitlinien

Evidenzbasierte Medizin und die Entwicklung von Behandlungsleitlinien stehen in engem Zusammenhang mit Case Management und Disease Management. Evidenzbasierte Medizin stellt den Anspruch, die bestmöglichen Diagnose- und Behandlungsverfahren anzuwenden, die wissenschaftlich abgesichert sind. Die Erkenntnisse der evidenzbasierten Medizin sollen die Grundlage für die Erstellung von Behandlungsleitlinien bilden.

Die Behandlungsleitlinien können in zwei Gruppen eingeteilt werden. *Protocols* sind verpflichtende Verfahrensvorschriften. *Guidelines* sind Behandlungsleitlinien, die empfehlenden Charakter haben und von denen auch abgewichen werden kann. Managed-Care-Organisationen versuchen durch die Implementierung von Behandlungsleitlinien im Rahmen von Disease Management und Case Management, die Behandlung von Patienten mit ähnlichem Krankheitsbild zu vereinheitlichen. Voraussetzung für die Implementierung von Behandlungsleitlinien ist die Bereitstellung von Behandlungspfaden in Datenbanken, die für die behandelnden Ärzte leicht abrufbar sind.[2]

Im Rahmen der Versorgungsforschung können mit Hilfe der Versicherungsdaten die

[1] vgl. Lehmann, 2003, S.39

[2] vgl. Lehmann, 2003, S.39

Behandlungsleitlinien unter Alltagsbedingungen (zumindest teilweise) evaluiert werden. Besonders wichtig sind die laufende Aktualisierung der Systeme und die Integration der neuesten medizinischen Erkenntnisse. Gillespie warnt vor der Praxis, Data Mining einzusetzen und Leitlinien zu entwickeln, die überwiegend auf diesen Erkenntnissen beruhen. Das kritische Hinterfragen der Ergebnisse, die Einbeziehung von Experten und die Integration abgesicherter wissenschaftlicher Erkenntnisse sind unumgänglich bei der Entwicklung und Implementierung von Leitlinien.[1]

7.5 Versorgungsforschung

Versorgungsforschung hat die Aufgabe, die Anwendung ineffektiver diagnostischer, therapeutischer und rehabilitativer Behandlungsmaßnahmen zu verhindern. Ziel ist einerseits die Verhinderung von Fehlallokationen medizinischer Leistungen. Andererseits soll die Wirtschaftlichkeit und Effizienz der medizinischen Versorgung gesteigert werden. Es sollen jene Methoden zur Anwendung kommen, die ihre Effektivität nachgewiesen haben. Versorgungsforschung stellt den Anspruch, die medizinische Versorgung unter Alltagsbedingungen zu untersuchen, im Gegensatz zur Forschung im Rahmen von Arzneimittelzulassungen durch die Hersteller. Mehrere Ebenen müssen dabei berücksichtigt werden:

- Strukturqualität (Qualifikation der Behandler und der Institution)
- Prozessqualität (Anwendung medizinischer Kenntnisse und die Dokumentation der durchgeführten medizinischen Intervention)
- Ergebnisqualität (Behandlungsergebnis, das sich durch den Gesundheitszustand und die Zufriedenheit der behandelten Person feststellen lässt)

Die Auswertung von Daten und Transparenz in den Daten stellen den Ausgangspunkt für qualitätssichernde Maßnahmen dar, die Dokumentation alleine stellt keine Sicherung der Qualität dar.

[1] vgl. Gillespie, 2004, S.52-60

Die Versorgungsforschung ermöglicht die Aufdeckung von Über-, Unter- und Fehlversorgung im Gesundheitssystem. Überversorgung ergibt sich zum Teil aus falsch gesetzten Anreizen im Gesundheitssystem, die geltenden Honorarordnungen und das Gewinnstreben der Leistungsanbieter. Es ist anzunehmen, dass hier erhebliche Wirtschaftlichkeitsreserven vorliegen. Fehlversorgung resultiert häufig aus der Verschreibung von Medikamenten, deren therapeutische Wirksamkeit als umstritten gilt. Schätzungen zufolge liegen die Ausgaben für diese Arzneimittelgruppen in Deutschland bei rund 1 Mrd. Euro pro Jahr. Unterversorgung ist vor allem im Bereich der Schmerztherapie festzustellen. Bis sich bessere Medikamente als Standard durchsetzen, vergehen oft Jahre, was ebenfalls eine Unterversorgung bedeutet.[1]

7.6 Betrugsaufdeckung

Der Anteil falscher Abrechnungen ist zwar sehr gering, der entstehende finanzielle Schaden ist aber sehr hoch. Aus diesem Grund setzen in den USA sehr viele staatliche und private Krankenversicherungen vermehrt auf Data-Mining-Technologien, um Betrugsmuster zu entdecken und zu unterbinden. Die betrügerischen Aktivitäten sind sehr unterschiedlich. Häufig werden Tests innerhalb weniger Monate mehrfach verrechnet, die gar nicht notwendig wären bzw. tatsächlich nur einmal durchgeführt wurden. Vielfach werden auch Leistungen in Rechnung gestellt, die gar nicht erbracht wurden.

Mit Hilfe überwachter Data-Mining-Verfahren konnten die Versicherungen einige Betrugsmuster aufdecken und Regeln implementieren, die zur Überprüfung „verdächtiger“ Abrechnungen führen. Zwei Probleme treten dabei auf: Zum einen basieren die Regeln auf historischen Daten und neue Betrugsvarianten werden nicht entdeckt. Zum anderen werden die Leistungen oft lange Zeit vor der Entdeckung von Unregelmäßigkeiten bezahlt. Daher kommen vermehrt unüberwachte Verfahren zur

[1] vgl. Glaeske, G.: Arzneimittelforschung: Basis für mehr Transparenz, Qualität und Patientenschutz – Verordnungsdaten der OÖGKK zu Antidepressiva. In: Bencic, W. (Hrsg.). Versorgung mit Antidepressiva. Gesundheitswissenschaften Band 23. Linz: OÖGKK 2003, S.11ff.

Anwendung, die noch vor der Zahlung Unregelmäßigkeiten aufdecken. Die Verfahren stammen überwiegend aus der Kreditkartenindustrie. Jede Abrechnung wird mit historischen Daten verglichen und das Betrugspotenzial festgestellt. „Verdächtige“ Abrechnungen werden in der Folge von Sachbearbeitern geprüft. Unüberwachte Verfahren ermöglichen schnelleres Handeln und die Aufdeckung neuer Muster. Die erhöhte Wahrscheinlichkeit, „erwischt“ zu werden, hat zusätzlich präventiven Charakter.[1]

Besonders betroffen von Betrugsfällen ist die große staatliche Versicherung Medicare, da sie für 41 Mio. Menschen verantwortlich ist und jährlich mit vielen verschiedenen Leistungserbringen rund eine Mrd. an Abrechnungen abwickelt. Der Anteil von betrügerischen Abrechnungen konnte seit 1996 stark verringert werden. Im Jahr 2002 wurden Abrechnungen in der Höhe von 11,6 Mrd. Dollar als fehlerhaft oder betrügerisch identifiziert. Schätzungen zufolge liegt der tatsächliche Anteil bei bis zu 50 Mrd. Dollar, die auf 1% der Abrechnungen zurückzuführen sind.[2]

7.7 Kundengewinnung und Kundenbindung

Der US-Versicherungsmarkt ist durch starke Konkurrenz geprägt. Die Nutzung von Daten spielt sowohl in der Gewinnung von Kunden als auch in der Bindung der Kunden eine wichtige Rolle.

7.7.1 Kundengewinnung

Die vorhandenen Daten werden von den Versicherungen vielfach genutzt, um Risikoprofile zu erstellen und ein entsprechendes Prämiensystem zu entwickeln. Es wird analysiert, welche Kundengruppen am profitabelsten sind. Mit gezielten

[1] Dorn, C. Data Mining Technology Helps Insurers Detect Health Care Fraud. National Underwriter **108** (39) 2004: S.34-39.

[2] vgl. Brandon, B. For Medicare, signs of fraud difficult to spot. San Jose Mercury News, 23.11.2003

Marketingaktivitäten wird versucht, genau diese Gruppen mit entsprechenden Versicherungsprodukten anzusprechen. Die Adressen bestimmter Haushalte, die angesprochen werden sollen, werden häufig zugekauft. Die Agenturen, die mit diesen Daten handeln, liefern Informationen über Alter und Geschlecht der Personen, die in einem Haushalt leben, das Ausbildungsniveau, das Einkaufsverhalten bis hin zu bevorzugten Medien und politischen Einstellungen. Anstatt Massenmarketing zu betreiben, wollen die Versicherungen ganz gezielt die gewinnträchtigsten Kundensegmente ansprechen, Dazu nutzen sie ihre eigenen Daten und jene von Datenagenturen, um an entsprechende Kontaktadressen zu kommen.[1]

Aus dieser Praxis ergeben sich erhebliche Nachteile. Personen, die einer Gruppe mit höherem Risiko angehören, werden benachteiligt. Es kommt zu einer Verzerrung des Wettbewerbs, da vor allem Versicherungen mit guter Selektionspraxis Vorteile haben anstatt Versicherungen mit gutem Kostenmanagement.[2]

7.7.2 Kundenbindung

Data Mining spielt in den US-Versicherungen auch eine wichtige Rolle, um Kunden zu binden. Data Mining unterstützt CRM-Aktivitäten und Marketing-Aktivitäten.

Data Mining dient dabei dem Zweck, die Bedürfnisse und Wünsche genauer kennen zu lernen und bestimmte Kundensegmente zu bilden. Daraus leiten die Organisationen entsprechende Maßnahmen ab. Beispielsweise erfolgen bestimmte Aussendungen an werdende Mütter mit Informationen über die Schwangerschaft und notwendige Untersuchungen. Automatisch generierte Aussendungen an Familien erinnern an Untersuchungen oder Behandlungen wie z.B. Impfungen.[3]

Das Unternehmen First Health Group nutzt Data Mining, um zu analysieren, in welchen Fällen die Kunden mit Fragen oder Sorgen an sie herantreten. Aufgrund der

[1] vgl. Sturm, 2004, S.100-102

[2] vgl. Lehmann, 2003, S.64

[3] vgl. Sturm, 2004, S.100-102

Ergebnisse entwickelte das Unternehmen Regeln, unter welchen Umständen die Versicherung von sich aus Anrufe tätigt. Dieses proaktive Verhalten wird von den Versicherten sehr positiv aufgenommen, sie fühlen sich dadurch gut betreut und ernst genommen. Die Kosten für die technische Lösung wurden nicht preisgegeben. Die Versicherung gibt aber an, dass sie neben der erhöhten Kundenzufriedenheit auch eine wesentlich effizientere Abwicklung von Anfragen und beträchtliche Zeitersparnisse erreichen konnte.[1]

Wie in anderen Branchen weit verbreitet, werden auch Kundenclubs angeboten (z.B. für die Versicherten über 50 Jahre) oder Versuche unternommen, Cross Selling zu betreiben.

7.8 Probleme der Datennutzung

Im Folgenden sollen einige Überlegungen zu den möglichen Gefahren und Problemen der Nutzung angestellt werden, die es neben den vielen Vorteilen zu bedenken gilt.

Disesase-Management-Programme stellen ein wirkungsvolles Instrument zur Leitung von Patienten durch das Gesundheitssystem dar und gewährleisten in vielen Fällen eine optimale Behandlung. Die Beratung von Risiko-Patienten über Call Center mag von vielen Menschen als positiv erlebt werden, dennoch stellen diese Programme einen Eingriff in die Privatsphäre dar. Es wird sehr stark auf den individuellen Lebensstil eingewirkt und es stellt sich die Frage, ob nicht auch Druck ausgeübt wird, um die Patienten zur Teilnahme an Beratungsprogrammen zu bewegen. Die Qualität und die Zufriedenheit der Patienten werden von Organisation zu Organisation variieren, tritt der Kostenaspekt zu stark in den Vordergrund, ist von „unerwünschten Nebenwirkungen“ für die Patienten auszugehen.

Datenmodelle versuchen die Realität möglichst genau abzubilden, aber sie bleiben

[1] vgl. Briggs, B. (2004). Data Helps Foretell Customer Needs. Health Data Management **12** (2) 2004: S.124-126.

eben immer nur ein Abbild, das nicht genau mit der Wirklichkeit übereinstimmt. Es ist zu bedenken, dass in Datensätzen meist Fehler enthalten sind, die trotz gründlicher Säuberung nicht völlig ausgeräumt werden können. Die Analysen basieren oft auf historischen Daten, wird der medizinische Fortschritt nicht berücksichtigt, besteht die Gefahr falscher Analysen und die Entwicklung von Behandlungsleitlinien, die nicht auf dem aktuellsten Stand sind. Eine weitere nicht unwesentliche Fehlerquelle stellt das Verhalten der Patienten selbst dar. Die Verschreibung von Medikamenten bedeutet noch lange nicht, dass sie tatsächlich eingenommen werden. Strikte Behandlungsleitlinien bedeuten auch eine Einschränkung der Freiheit der Ärzte, die über sehr gute Erfahrungswerte verfügen und im Einzelfall eine abweichende Behandlung vorziehen würden. Data Mining ist kein Allheilmittel zur Verbesserung der Versorgungsqualität und gleichzeitiger Kostensenkung. Die Herausforderung liegt darin, den gezielten Einsatz der Datennutzung durch die geeigneten Begleitmaßnahmen zu ergänzen.

Die Nutzung von Daten erlaubt die sehr genaue Ermittlung von Risikoprofilen. Die privaten Versicherungen sind daher bestrebt, möglichst viele Menschen mit einer günstigen Risikoprognose zu versichern. Die Datennutzung ist dafür zwar nicht kausal verantwortlich, aber stellt einen Aspekt dieser Entwicklung dar. Jedenfalls sind von den 240 Mio. amerikanischen Bürgern 40 Millionen ohne Versicherung und 50 Millionen erheblich unterversichert.[1]

Die Verfügbarkeit von umfassendem Datenmaterial und die intensive Datennutzung bringen die Gefahr des Datenmissbrauchs mit sich. Vielfach schließen die Arbeitgeber Verträge für ihre Mitarbeiter mit den Versicherungen ab. Sie könnten an Auswertungen sehr interessiert sein, um möglichst gesunde Mitarbeiter zu halten. Diese Gedanken stellen nur Spekulationen dar, aber sind nicht ganz von der Hand zu weisen. Dieser Gefahr kann nur durch Sicherheitstechnologien, strenge gesetzliche Bestimmungen und entsprechende Strafen begegnet werden.

[1] vgl. Gaertner, Jelastopulu, Niehoff, 2000, S.15

8 Fallbeispiel Oberösterreichische Gebiets-krankenkasse

Die oberösterreichische Gebietskrankenkasse (OÖGKK) implementierte gemeinsam mit dem SAS-Institute Mitte der 90er Jahre ein Informationssystem mit dem Ziel, die Fülle an vorliegenden Verrechnungsdaten besser nutzbar zu machen und für analytische Zwecke einzusetzen. Herzstück des Systems ist die die FOKO-Auswertungssoftware. Der Name FOKO leitet sich vom Terminus ärztliche Eigen- und Folgekostenanalyse ab. In diesem Kapitel wird die Entstehung und der Aufbau des Systems erläutert und welche Auswertungsmöglichkeiten es grundsätzlich bietet. Anhand der Ausführung eines konkreten Forschungsprojekts am Ende des Kapitels wird das Potenzial von FOKO nochmals verdeutlicht. Die verwendeten schriftlichen Unterlagen für dieses Kapitel sind durch Informationen aus Gesprächen mit Experten der OÖGKK ergänzt.

8.1 Entwicklung von FOKO

Nach einer Ausschreibung erfolgte die Vergabe für die Entwicklung von FOKO an das Unternehmen SAS, 1994 wurde mit der Programmierung begonnen und 1996 ging die erste Version in Betrieb und wird neben der OÖGKK von 12 Sozialversicherungsträgern in Oberösterreich als Standardprodukt betrieben.

Ziel der ersten Ausbaustufe (FOKO I) war die Erstellunge eines ökonomischen Gesamtbildes. Die Vertragspartnerabrechnung stand im Vordergrund, um die Ökonomie zu überwachen und gleichzeitig die Behandlungsqualität zu gewährleisten.

In der zweiten Ausbaustufe (FOKO II) wurden die versichertenbezogenen Daten integriert, was die Auswertung von Sach- und Barleistungen auf Ebene der Versicherten ermöglicht bzw. die Verknüpfung von arzt- und patientenbezogenen

Daten. FOKO dient rein Informations- und nicht Abrechnungszwecken.
Heute ist FOKO die einzige Plattform, die alle Behandlungsdaten aller versicherten Personen über einen bestimmten Zeitraum enthält. Die Verknüpfung mit externen Daten ist zum Teil schon möglich, soll aber noch verstärkt werden.[1]

Die folgende Auflistung zeigt die vielfältigen Anwendungsgebiete von FOKO, die noch näher ausgeführt werden:

- Information für gesundheitspolitische Entscheidungen
- Trendanalysen
- Controlling, Ableitung von Steuerungsmaßnahmen
- Feststellung von Einsparungspotenzial
- Information für Ärzte
- Information für Versicherte
- Forschungsprojekte

8.2 Architektur des Data Warehouse

Die FOKO-Datenbank ist multidimensional aufgebaut, woraus sich viele Vorteile ergeben. Alle im Data Warehouse enthaltenen Daten können verknüpft werden. Die Durchführung von Abfragen erfolgt sehr flexibel und orientiert sich an den Bedürfnissen der Anwender. Der überwiegende Teil der Abfragen benötigt nur sehr kurze Antwortzeiten, neues Wissen kann so rasch generiert werden.

Die Abrechnungsdaten sind auf mehreren Rechnern in uneinheitlicher Form gespeichert. Die Uneinheitlichkeit ergibt sich z.B. aus unterschiedlichen Verträgen mit den Leistungserbringern. Daher ist eine Schnittstelle erforderlich, um die Daten in das Data Warehouse einzuspeisen. Im Data Warehouse kommen Basisdaten, Dimensionen, Fakten und Aggregate zusammen.

Die Auswertungen erfolgen durch den *Aggregate Navigator*. Die Programmierung

[1] vgl. Hofer, 2004, o.S.

dieses komplexen Algorithmus erforderte ein Jahr Programmierarbeit und stellt das Herzstück des FOKO-Systems dar. Die Anwender können auf einer einfachen Benutzeroberfläche die Abfragen formulieren, die vom Aggregate Navigator erledigt werden. Gab es bereits eine ähnliche Abfrage, erhöht sich die Geschwindigkeit des Navigators. Selbst komplexe Abfragen sind meistens in einigen Minuten erledigt. Vor Einführung von FOKO benötigten einfache Abfragen mehrere Stunden, da es sich um sequentielle Daten handelte. **Abbildung 11** zeigt schematisch die Architektur des FOKO-Systems.

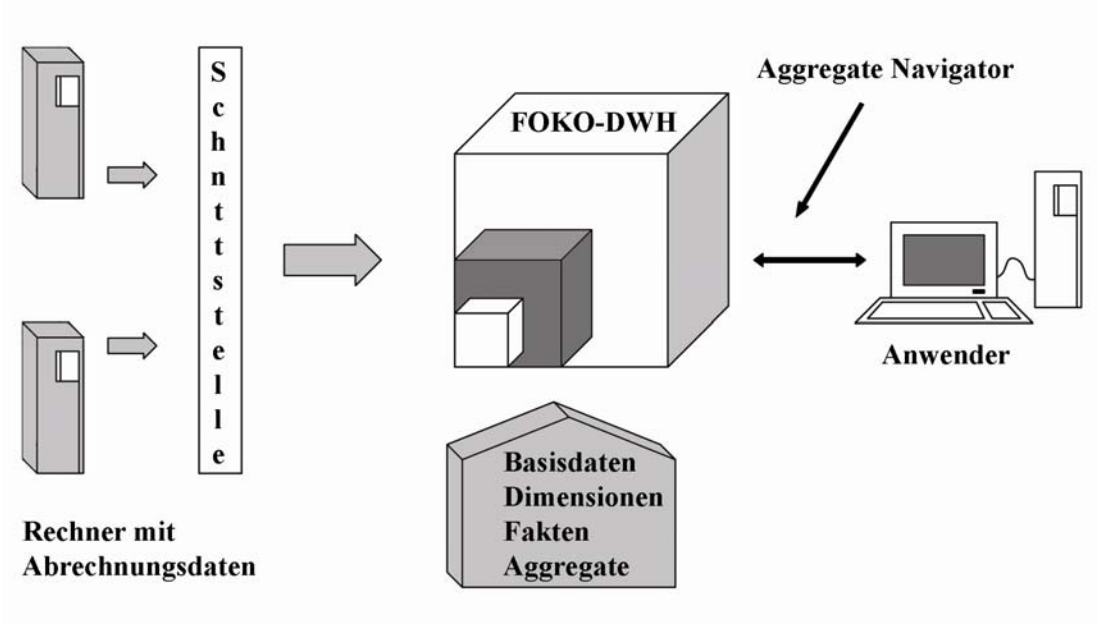


Abbildung 11: Architektur des FOKO-Data Warehouse

(Quelle: Hofer, 2004, o.S.)

In Oberösterreich werden beispielsweise pro Monat rund 1 Mio. Heilmittel verordnet. Es ist daher nicht verwunderlich, dass die Größe des Data Warehouse derzeit bereits eine Größe von 600 Gigabyte umfasst. Es ist davon auszugehen, dass in den nächsten Jahren das Datenvolumen noch wesentlich ansteigen wird. Dieses Datenvolumen erfordert sehr viel kontinuierliche Wartung. Auf „Knopfdruck“ können Daten über einen Zeitraum von acht Quartalen abgerufen werden, nach zwei

Jahren wird der Großteil der Daten auf Bändern gesichert.[1]

8.3 Verfügbare Daten im Data Warehouse

8.3.1 Stammdaten Ärzte

Über die Vertragspartner sind folgende Daten verfügbar:

- Name
- Alter
- Vertragsinformation
- Fachgruppe
- Adresse
- Gemeinde

8.3.2 Stammdaten Versicherte

Der zentrale Schlüssel ist die Versicherungsnummer des Versicherten, folgende Daten sind über die Versicherten verfügbar:

- Identität
- Wohnort (Gemeinde, Bezirk)
- Geschlecht
- Versichertenkategorie (Angestellter, Arbeiter, etc.)
- Einkommen
- Angaben zum Dienstverhältnis (Wirtschaftsklasse, Betriebsgröße, Standort, Beitragsgruppen)

[1] vgl. Hofer, 2004, o.S.

8.3.3 FOKO Leistungsdaten

Die Informationen über Diagnosen, erbrachte Leistungen und Krankenstände gelangen auf drei Wegen zur Kasse:

1. Information auf den Krankenscheinen, die die Abrechnungsgrundlage darstellen
2. Krankmeldungen der Unternehmen
3. Krankenhausdaten (Aufnahme- und Entlassungsdiagnose)

Auf Basis dieser Informationen können die Arztkontakte, die daraus resultierenden Arzt-Eigenkosten und veranlasste Leistungen (Folgekosten) ermittelt werden. Sehr differenzierte Informationen liegen über die Heilmittel und Heilbehelfe vor. Zeiten der Arbeitsunfähigkeit, Krankenhausaufenthalte, Kuraufenthalte und in Ambulatorien erbrachte Leistungen können analysiert werden.

8.4 Datenmängel und Verbesserungspotenzial

In den Gesprächen mit Experten der OÖGKK wurden einige Datenmängel und Möglichkeiten zur Verbesserung der Datenlage besprochen, die im Folgenden ausgeführt werden.

8.4.1 Datenmängel

Im niedergelassenen Bereich erfolgen die Diagnosen nicht nach dem internationalen Klassifikationsschema ICD-10 (=International Classification of Disesases) sondern in Wortdiagnosen. Für eine bestimmte Krankheit liegen daher bis zu 20 Synonyme und Abkürzungen vor. Aus diesem Grund müssen für Auswertungen in vielen Fällen zeitaufwändige Kategorisierungen vorgenommen werden. Für die Krankenversicherung wäre die Klassifizierung nach ICD-10 sehr wünschenswert,

Diskussionen über die Einführung im extramuralen Bereich werden seit Jahren geführt.

Die Krankenscheine stellen die Abrechnungsgrundlage für die erbrachten Leistungen der Ärzte dar. Die Diagnosen oder Angaben können unvollständig sein, vor allem, wenn mit einer erbrachten Einzelleistung, z.B. aufgrund einer Deckelung in der Honorarordnung, kein Entgeltanspruch verbunden ist. Häufig werden auch Probepackungen von Medikamenten abgegeben, die nicht aufscheinen. Das heißt, dass Ärzte tatsächlich mehr Diagnosen stellen und Leistungen erbringen, als auf dem Krankenschein vermerkt sind.

Aufgrund des komplexen Honorierungssystems ist die Ermittlung von den tatsächlichen Kosten meist aufwändig. Zu einer bestimmten pauschalen Grundleistungsvergütung kommen verrechenbare Einzelleistungen hinzu. Einige Tarife sind degressiv gestaffelt oder durch KO-Limitierungen beschränkt. All diese Faktoren sind bei den Auswertungen zu berücksichtigen.

Wenig Information liegt über die Tätigkeit von Wahlärzten und die erbrachten Leistungen in Krankenhausambulanzen vor. Hier kann nur die Integration von externen Daten Abhilfe schaffen, was vereinzelt in Projekten versucht wurde. Für Mitversicherte, die keine eigene Versicherungsnummer haben, ist die Datenlage nicht sehr gut, bzw. können erbrachte Leistungen nicht eindeutig zugeordnet werden. Mit der Vergabe von Versicherungsnummern für Mitversicherte wurde vor längerer Zeit begonnen, mit der Einführung der e-card wird dieser Datenmangel behoben.

Ein grundsätzliches Problem stellt die Tatsache dar, dass sich Patienten häufig nicht an die Anordnungen ihres Arztes halten und beispielsweise Medikamente unregelmäßig oder gar nicht einnehmen. Die Auswertung von Daten erlaubt nur beschränkt Rückschluss auf das tatsächliche Verhalten der Patienten. Für Forschungsvorhaben ist daher oft die Befragung von Patienten sinnvoll, um das Verhalten der Patienten und deren Zufriedenheit mit einer Behandlung beurteilen zu können.

8.4.2 Auswirkungen der e-card

Ende Februar 2005 startete der Probetrieb für die neue e-card, bis Jahresende 2005 sollen 8 Millionen Österreicher mit der Karte ausgestattet werden. Auf dem Datenchip selbst sind nur persönliche Daten des Karteninhabers wie Name, akademischer Grad, Geburtsdatum und Versicherungsnummer gespeichert. Gesundheitsdaten werden auf der Karte nicht gespeichert, um höchstmöglichen Datenschutz zu gewährleisten. Die e-card ersetzt den Krankenschein in Österreich und den Auslandkrankenschein in der EU und einigen weiteren Ländern.[1]

Die e-card bringt vor allem eine große Verwaltungsvereinfachung mit sich. Die Datenqualität für Auswertungen wird gesteigert, da nun alle mitversicherten Personen eine eigene Karte bekommen und erbrachte Leistungen so immer eindeutig einer Person zuordenbar sind.

Aus Sicht des Autors bietet die e-card weitere Nutzungsmöglichkeiten, die bereits Gegenstand von Diskussionen sind und in der Zukunft möglicherweise auch umgesetzt werden. Es ist vorstellbar, auch Daten auf die e-card zu speichern, die lebensrettend sein können, wie z.B. die Blutgruppe. Besonders vielversprechend scheint die Implementierung von Disease-Management-Programmen, die in Kapitel 7.2 beschrieben sind. Dazu müssten institutionelle und technische Rahmenbedingungen geschaffen werden, die mittels e-card den Abruf der Krankengeschichte und des Behandlungsverlaufs ermöglichen. Damit könnte der Behandlungsverlauf überwacht und optimiert werden. Dadurch könnten rechtzeitig notwendige Interventionen gesetzt werden, um die Verschlechterung des Gesundheitszustandes von Patienten und damit verbundene höhere Kosten zu vermeiden. Für ohnehin oft sehr belastete chronisch kranke Menschen würde der Weg durch das komplexe Gesundheitssystem erleichtert.

Die Umsetzung neuer Ansätze bringt zwei große Herausforderungen mit sich. Zum einen muss der Datenschutz gewährleistet sein. Zum anderen müssen Wege gefunden

[1] vgl. http://esv-sva.sozvers.at/esvapps, 10.04.2005

werden, um diese Ansätze in die bestehende Versorgungsstruktur zu integrieren und entsprechende Anreize für deren Umsetzung geschaffen werden.

8.5 FOKO-Anwendungen

8.5.1 Controlling und statistische Auswertungen

FOKO ermöglicht die rasche Durchführung von Standardabfragen, die Darstellung von Zeitreihen und die Generierung von Berichten. Daten können auf einfache Weise in andere Anwendungen, z.B. Microsoft-Programme, importiert werden.

Mit FOKO wird das Finanzcontrolling stark unterstützt. Die Budgetplanung erfolgt auf Jahresbasis. Monatlich werden von den Budgetverantwortlichen Analysen durchgeführt und Hochrechnungen zum Jahresende durchgeführt. In monatlichen Besprechungen werden die Entwicklungen diskutiert und falls erforderlich Maßnahmen eingeleitet. Der Planungshorizont über das aktuelle Jahr hinaus beträgt drei Jahre. Diese Mittelfristplanung ist systematisch wie die Jahresplanung aufgebaut und wird quartalsweise aktualisiert.[1]

Die FOKO-Analysen sind Teil der Gesundheitsberichterstattung. Die Auswertungen stellen eine Entscheidungshilfe für gesundheitspolitische Maßnahmen dar. Mit FOKO kann rasch Zahlenmaterial zu aktuellen Themen und tagespolitischen Anfragen geliefert werden.[2]

[1]vgl. Wesenauer, A.: Controlling als Instrument der Unternehmenssteuerung und Organisationsentwicklung in der OÖGKK. Soziale Sicherheit, März 2004: S.100-106

[2] vgl. Hofer, 2004, o.S.

8.5.2 Vertragspartnerauswertung und -information

Mit FOKO kann das Verordnungsverhalten und die daraus resultierenden Kosten genau analysiert werden. Auf Quartalsbasis werden die Kosten für einzelne Positionen eines Arztes mit den Durchschnittswerten der jeweiligen Fachgruppe in Beziehung gesetzt. Dabei werden eine Mengenkomponente (z.B. verordnet viel aber günstig) und eine Preiskomponente (z.B. verordnet wenig aber teuer) berücksichtigt. Die Altersstruktur der Patienten findet ebenfalls Eingang in die Auswertung (Einteilung in Dekaden). Neben dem Vergleich mit der Fachgruppe können die Durchschnittswerte auf Bezirks-, Landes- und Bundesebene und Vorjahreswerte angezeigt werden.[1]

Das System dient Kontrollzwecken, bei sehr starken Abweichungen wird das Gespräch mit den betroffenen Ärzten gesucht, um die Situation zu analysieren und Verbesserungen einzuleiten. Deutlich mehr Gewicht als die Kontrolle hat aber die Information der Vertragspartner, die auf unterschiedliche Weise kommuniziert wird. Auf Wunsch erhalten die Ärzte quartalsweise einen behandlungsökonomischen Servicebrief mit einem Umfang von ca. 10 Seiten, in dem ihre Leistung im Vergleich zur Fachgruppe dargestellt wird. Etwa ein Drittel der Ärzte nimmt dieses Angebot in Anspruch. Automatisch generierte Serienbriefe machen die Ärzte darauf aufmerksam, wie viel Geld sie bis zum Jahresende durch die Verschreibung bestimmter Generika einsparen könnten. Darüber hinaus wird auf Informationsveranstaltungen über aktuelle Entwicklungen berichtet.[2]

[1] vgl. Hofer, 2004, o.S.

[2] vgl. Hofer, 2004, o.S.

8.5.3 Patienteninformation

Viele Erkenntnisse aus FOKO-Analysen werden in verschiedenen Zeitschriften und Ratgebern für Patienten aufbereitet. Im Jahr 2004 hatten die Krankenversicherungen erstmals den gesetzlichen Auftrag, jeden Versicherten über die für ihn erbrachten Leistungen und daraus resultierenden Kosten im vergangenen Jahr zu informieren.

Dieses Leistungsinformationsblatt wurde an alle Personen über 14 Jahre versandt mit dem Hintergrund, die Gesundheitskosten transparenter zu machen und persönlich zu informieren. Die Kosten für die gesamte Aktion wurden mit sechs Mio. Euro beziffert.[1]

Die Leistungsinformation wurde auf Basis von FOKO-Daten erstellt. Die Maßnahme sorgte für einige Diskussionen und Unmut. Zum Teil schlichen sich Fehler in die Auswertungen ein, was die Einrichtung von Hotlines erforderte, um Anfragen zu beantworten. Die angeführten Leistungen wurden nicht sehr detailliert aufgeschlüsselt, da dies in vielen Fällen zu sehr langen Berichten geführt hätte. Es ist zu bezweifeln, dass sich Patienten Monate nach einer Behandlung noch genau an die erbrachten Leistungen erinnern können, um sie mit der Abrechnungsinformation vergleichen zu können. Es stellt sich die Frage, ob sich schwer kranke Menschen nicht unter Druck gesetzt fühlen, wenn sie mit den hohen Kosten konfrontiert werden. Auf der anderen Seite stehen Menschen, die hohe Beiträge einzahlen und wenig Kosten verursachen und sich eventuell über die hohen Beiträge ärgern. Ob der Informationsgewinn für den einzelnen Versicherten den hohen Aufwand für die Aussendung der Leistungsinformation rechtfertigen kann, ist fraglich.

[1] vgl. Müller, H. Briefe mit falschen Zahlen. Die Presse, 21.08.2004: S.10

8.5.4 Beispiel für ein Forschungsprojekt

Die OÖGKK führt viele wissenschaftliche Untersuchungen durch. Häufig erfolgt eine Kooperation mit Universitäten und anderen Institutionen. Eine aktuelle und besonders umfangreiche Studie beschäftigt sich mit der Versorgung mit Antidepressiva in Oberösterreich, die im Folgenden näher beschrieben wird und das Potenzial von FOKO für Forschungszwecke veranschaulicht.

8.5.4.1 Ziele und Hintergrund des Projekts

Schätzungen der Weltgesundheitsorganisation WHO zufolge erkranken rund 25% der Menschen zumindest einmal in ihrem Leben an einer psychischen Erkrankung. Besonders häufig sind Depressionen und Angsterkrankungen. Die WHO geht von einer starken Zunahme psychischer Erkrankungen aus und damit verbundenen Belastungen der Gesundheitssysteme und Volkswirtschaften.[1]

Im Jahr 1999 wurden von den niedergelassenen Ärzten in Österreich rund 6,2 Mio. Psychopharmaka-Packungen verordnet, wobei rund 45% davon auf Antidepressiva entfallen. Zwischen 1991 und 2000 vervierfachte sich der Absatz von Antidepressiva-Packungen. Die verordneten Antidepressiva verursachten im Jahr 1999 rund 58% der gesamten Kosten für Psychopharmaka.[2]

Vor diesem Hintergrund wurde die Studie mit dem Ziel durchgeführt, mehr über die sozioökonomischen Merkmale der Versicherten mit Antidepressiva-Verordnungen zu erfahren, das Verordnungsverhalten der Ärzte zu analysieren und allgemeine Entwicklungen aufzuzeigen.

---

[1] vgl. The World Health Report 2001. Mental Health: New Understanding, New Hope. Genf: WHO 2001, S.23f.

[2] vgl. Riedel, M., Hofmarcher, M.M.: Verordnungsvarianz und Outcome der Antidepressiva-Versorgung in Oberösterreich. In: Bencic, W. (Hrsg.). Versorgung mit Antidepressiva. Gesundheitswissenschaften Band 23. Linz: OÖGKK 2003, S.29

8.5.4.2 Methodik

Die Auswertung erfolgte über die Quartalsdaten 2000 und 2001 der OÖGKK. Einbezogen wurden 88.108 Personen, die in den Jahren 2000 oder 2001 zumindest eine Packung Antidepressiva erhielten. Davon mussten 3.716 Fälle (oder 4%) aufgrund unvollständiger Angaben ausgeschieden werden. Dies sind vor allem mitversicherte Personen, bei denen keine eigene Versicherungsnummer vorliegt. Es wird explizit darauf hingewiesen, dass sich die Untersuchung auf die Verordnungen von Antidepressiva bezieht und nicht auf Personen mit der Diagnose Depression, da die Diagnosen im niedergelassenen Bereich nur teilweise erfolgen.[1]

8.5.4.3 Darstellung wichtiger Ergebnisse

Insgesamt wurde in den Jahren 2000 oder 2001 einem Anteil von 7% der OÖGKK-Versicherten zumindest eine Packung Antidepressiva verordnet. Davon waren 70% Frauen, insgesamt machen die Frauen einen Anteil von 52% der Versicherten aus. Der Anteil von Antidepressiva-Beziehern nimmt in der Regel mit dem Alter zu. Die Geschlechtsunterschiede können nicht durch die höhere Anzahl älterer Frauen erklärt werden.

22.550 Personen (oder 27%) wurde im Beobachtungszeitraum nur eine einzige Packung verordnet. Der Anteil ist mit 30% bei Männern höher als bei Frauen (25%). Außerdem besteht ein Zusammenhang mit dem Alter, je jünger die Bezieher sind, desto höher ist die Wahrscheinlichkeit, dass nur eine Packung verordnet wird. Von den Kosten her betrachtet fällt diese Gruppe nicht stark ins Gewicht, dennoch deutet dieses Ergebnis auf eine Fehlversorgung hin, da Antidepressiva ihre Wirkung oft erst nach vier bis sechs Wochen entfalten.[2]

Die Auswertung nach Versichertenkategorien zeigt, dass überdurchschnittlich viele Arbeitslose Antidepressiva beziehen. Der Zusammenhang ist allerdings unklar, Arbeitslosigkeit kann eine Folge von Depressionen sein, umgekehrt kann drohende Arbeitslosigkeit oder wiederholte Arbeitslosigkeit Depressionen (mit-)verursachen. Die Betrachtung der Einkommen liefert Hinweise auf Zusammenhänge zwischen

[1] vgl. Riedel, Hofmarcher, 2003, S.33f.

[2] vgl. Riedel, Hofmarcher, 2003, S.36f

sozioökonomischer Faktoren und Depression. Die zweit- und die drittniedrigste Einkommensgruppe weisen höhere Anteile an Antidepressiva-Beziehern auf.[1]

Die Studie brachte ein Stadt-Land-Gefälle ans Licht. In der Stadt liegt der Anteil von Antidepressiva-Beziehern bei 9,3%, bei der Landbevölkerung beträgt dieser Wert 6,6%. Ob dieser Unterschied durch unterschiedliches Nachfrage-Verhalten, bessere Versorgungsstrukturen in der Stadt oder höhere Risikofaktoren in der Stadt bedingt wird, kann anhand der Daten nicht beantwortet werden.[2]

Der kurze Ausschnitt aus den Ergebnissen zeigt, wie detailliert Auswertungen mit FOKO durchgeführt werden können und aktuelle Trends und Probleme dadurch sichtbar werden. Gleichzeitig wird deutlich, dass begleitende Forschung notwendig ist, um die Ursachen für die Ergebnisse näher analysieren zu können und geeignete Maßnahmen abzuleiten.

[1] vgl. Riedel, Hofmarcher, 2003, S.37ff.

[2] vgl. Riedel, Hofmarcher, 2003, S.39

9 Zusammenfassung und Ausblick

Angesichts der kontinuierlich ansteigenden Gesundheitsausgaben hat die Nutzung von Daten im Rahmen betrieblicher Informationssysteme für Krankenversicherungen eine sehr große Bedeutung. Das Spektrum der Anwendungen ist sehr breit und beschränkt sich nicht nur auf die Entdeckung von Einsparungspotenzialen.

Voraussetzung für die systematische Nutzung ist die Etablierung eines Data Warehouse, in dem alle erbrachten Leistungen der Versicherung über einen längeren Zeitraum gespeichert sind und nach unterschiedlichen Gesichtspunkten ausgewertet werden können. Besonders wichtige Auswertungstechniken stellen Online Analytical Processing (OLAP) und Data Mining dar.

Viele der in der vorliegenden Arbeit beschriebenen Anwendungen stammen aus den USA. Das US-Gesundheitswesen ist durch einen hohen Anteil privater Versicherungen und Wettbewerb geprägt. Aus diesem Grund setzen die Versicherungen schon seit längerer Zeit auf die Nutzung von Daten, um Wettbewerbsvorteile zu erreichen.

Besonders große Bedeutung hat Data Mining, um Risikofaktoren zu identifizieren, die sich negativ auf den Gesundheitszustand von Patienten auswirken und in absehbarer Zeit zu erhöhten Kosten führen könnten. Die gewonnenen Erkenntnisse werden in Disease-Management-Programme integriert. Ziel von Disease Management ist die Optimierung des Behandlungsverlaufs und die Beratung der Patienten, um Maßnahmen zu setzen, bevor es zu einer Verschlechterung des Gesundheitszustandes und damit verbundenen höheren Kosten kommt. Die behandelnden Ärzte können über Schnittstellen den Behandlungsverlauf einsehen und auf Behandlungsleitlinien zurückgreifen. Die Beratung erfolgt vielfach über eigens eingerichtete Call Center.

Data Mining wird ähnlich wie in der Kreditkartenindustrie zur Aufdeckung und Bekämpfung von Betrugsmustern eingesetzt. Die Datenanalysen unterstützen die

Versicherungen bei der Festsetzung der Prämien durch die Erstellung von Risikoprofilen und Marketing-Aktivitäten. Die Versicherungen sind bestrebt, möglichst viele Kunden mit günstigem Risikoprofil zu gewinnen und durch datengestützte Maßnahmen zu binden (z.B. spezielle Informationen für Familien, Kundenclubs etc.)

Den großen Vorteilen, Einsparungspotenziale aufzudecken und den Behandlungsablauf zu optimieren, stehen gewichtige Nachteile gegenüber. Die starke Risikoselektion führt dazu, dass nicht die Versicherungen mit dem besten Kostenmanagement Wettbewerbsvorteile haben, sondern die Versicherungen mit optimaler Selektionspraxis. Die intensive Datennutzung bringt die Gefahr von Datenmissbrauch mit sich, die nur durch strenge gesetzliche Bestimmungen und entsprechend hohe Strafen bekämpft werden kann.

Die Oberösterreichische Gebietskrankenkasse setzt seit 1996 die Software FOKO ein, um ihre Daten besser nutzen zu können. Im zugrundeliegenden Data Warehouse können die erbrachten Leistungen durch die Vertragspartner mit den Daten der Versicherten verknüpft werden. Mit dem System können rasch ein ökonomisches Gesamtbild erstellt und Kostenbereiche aufgedeckt werden, die besondere Beachtung verdienen. Die Budgetierung für das laufende Jahr und die mittelfristige Planung der kommenden drei Jahre basieren auf dem System. FOKO dient auch der Kontrolle der Vertragspartner, viel wichtiger sind aber die Information der Ärzte und die Aufklärung der Versicherten. Große Einsparungen konnten durch den verstärkten Einsatz von Generika erzielt werden, die durch gezielte Aufklärung der Ärzte bewirkt wurde. FOKO wird auch intensiv für Forschungszwecke genutzt, z.B. um ein besseres Verständnis über den Behandlungsverlauf bestimmter Erkrankungen zu erlangen oder Fehlversorgungen aufzudecken.

Die neue e-card wird den Krankenschein ersetzen und große Verwaltungsvereinfachungen und eine höhere Datenqualität mit sich bringen. Das verbesserte Schnittstellenmanagement durch die neue e-card bietet grundsätzlich das Potenzial, in Zukunft verstärkt präventive Akzente in der Gesundheitsversorgung zu setzen, z.B. durch die Implementierung von Disease-Mangagement-Programmen. Dabei stellt sich die Herausforderung, die neuen Ansätze in die bestehende

Versorgungsstruktur zu integrieren und den Ansprüchen des Datenschutzes gerecht zu werden.

Die vorliegende Diplomarbeit zeigt, wie vielfältig Krankenversicherungsdaten genutzt werden können, um steuerungsrelevantes Wissen zu generieren, Kosteneinsparungen zu erreichen und qualitätssichernde Maßnahmen zu setzen. Die systematische Nutzung von Daten ist kein Allheilmittel für die Probleme des Gesundheitswesens, kann aber einen wichtigen Beitrag dazu leisten, vor allem chronisch Kranke besser zu betreuen und Behandlungsabläufe zu optimieren. Die Bedeutung präventiver Ansätze wie Disease Management wird in Österreich in den kommenden Jahren mit sehr großer Wahrscheinlichkeit zunehmen. Entscheidend für den Erfolg bei der Umsetzung neuer Ansätze sind die gleichzeitige Berücksichtigung der Kosten und der Versorgungsqualität sowie die Evaluierung der gesetzten Maßnahmen.

Abbildungs- und Tabellenverzeichnis

Abbildungen

Tabellen

Literaturverzeichnis

Anahory, S., Murray, D. (1997). Data Warehouse: Planung, Implementierung und Administration. Bonn: Addison-Wesley-Longman.

Briggs, B. (2004). Data Helps Foretell Customer Needs. Health Data Management **12** (2): S.124-126.

Dorn, C. (2004). Data Mining Technology Helps Insurers Detect Health Care Fraud. National Underwriter **108** (39): S.34-39.

Eichbauer, H., Klaushofer, K. (2004). Disease-Management – Case Management. In: Almer, S., Bencic, W. (Hrsg.). Mittelverwendung versus Mittelverschwendung: Fehl-, Über- und Unterversorgung im Gesundheitswesen. Gesundheitswissenschaften Band 26. Linz: OÖGKK: S.81-86.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U.M. et al. (Hrsg.). Advances in knowledge discovery and data mining. AAAI Press/MIT Press: S.1-37.

Gaertner, T., Jelastopulu, E., Niehoff, J.-U. (2000). Managed Care in den USA. In: Medizinische Dienste der Krankenversicherung (Hrsg.). Managed Care – Eine Perspektive für die GKV? Stuttgart: Georg Thieme: S.5-19.

Gillespie, G. (2004). Data Mining: Solving Care, Cost Capers. Health Data Management **12** (11): S.52-60.

Glaeske, G. (2003). Arzneimittelforschung: Basis für mehr Transparenz, Qualität und Patientenschutz – Verordnungsdaten der OÖGKK zu Antidepressiva. In: Bencic, W. (Hrsg.). Versorgung mit Antidepressiva. Gesundheitswissenschaften Band 23. Linz: OÖGKK: S.11-28.

Grimmer, U., Mucha, H.-J. (1998). Datensegmentierung mittels Clusteranalyse. In: Nakhaeizadeh, G. (Hrsg.): Data Mining: Theoretische Aspekte und Anwendungen. Heidelberg: Physica-Verlag: S.109-142

Groth, R. (1998). Data Mining: A Hands-On Approach for Business Professionals. New Jersey: Prentice Hall.

Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques. San Diego: Academic Press.

Hofer, P. (2004). Kosteneinsparung und Qualitätssicherung in der Behandlungsökonomie durch Folgekostenanalyse mit SAS bei der oberösterreichischen Gebietskrankenkasse. Tagungsbericht: ICV Forum Gesundheitswesen Österreich 2004, 24.09.2004, Wien.

Hönig, T. (1998). Desktop OLAP in Theorie und Praxis. In: Martin, W. (Hrsg.): Data Warehousing: Data Mining – OLAP. Bonn: International Thompson: S.169-188.

Houghton, D. et al. (2003). A Review of Software Packages for Data Mining. The American Statistician **57** (4): S.290-309.

Jarke, M. et al. (2000). Fundamentals of data warehouses. Berlin: Springer.

Kirchner, J. (1998). Online Analytical Processing. In: Martin, W. (Hrsg.): Data Warehousing: Data Mining – OLAP. Bonn: International Thompson. S.147-166.

Knobloch, B. (2000). Der Data-Mining-Ansatz zur Analyse betriebswirtschaftlicher Daten. Bamberg: Otto-Friedrich-Universität.

Krahl, D., Windheuser, U., Zick, F.K. (1998). Data Mining: Einsatz in der Praxis. Bonn: Addison-Wesley-Longman.

Küppers, B. (1999). Data Mining in der Praxis: ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld. Wien: Lang.

Lehmann, H. (2003). Managed Care: Kosten senken mit alternativen Krankenversicherungsformen? Zürich: Rüegger.

Lusti, M. (1999). Data Warehousing und Data Mining: Eine Einführung in entscheidungsunterstützende Systeme. Berlin: Springer Verlag.

Martin, W. (1998). Data Warehouse, Data Mining und OLAP: Von der Datenquelle zum Informationsverbraucher. In: Martin, W. (Hrsg.): Data Warehousing: Data Mining – OLAP. Bonn: International Thompson.

Mertens, P., Wieczorrek, H.W. (2000). Data X Strategien: Data Warehouse, Data Mining und operationale Systeme für die Praxis. Berlin: Springer Verlag.

Müller, H. (2004). Briefe mit falschen Zahlen. Die Presse, 21.08.2004: S.10

Nakhaeizadeh, G., Reinartz, T., Wirth, R. (1998). Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick. In: Nakhaeizadeh, G. (Hrsg.): Data Mining: Theoretische Aspekte und Anwendungen. Heidelberg: Physica Verlag: S.1-33.

Pieber, T., Seereiner, S. (2004). Disease Management am Beispiel Diabetes mellitus. In: Almer, S., Bencic, W. (Hrsg.). Mittelverwendung versus Mittelverschwendung: Fehl-, Über- und Unterversorgung im Gesundheitswesen. Gesundheitswissenschaften Band 26. Linz: OÖGKK: S.87-96.

Riedel, M., Hofmarcher, M.M. (2003). Verordnungsvarianz und Outcome der Antidepressiva-Versorgung in Oberösterreich. In: Bencic, W. (Hrsg.). Versorgung mit Antidepressiva. Gesundheitswissenschaften Band 23. Linz: OÖGKK, S.29-62.

Runkler, T. A. (2000). Information Mining: Methoden, Algorithmen und Anwendungen intelligenter Datenanalyse. Braunschweig/Wiesbaden: Vieweg&Sohn.

Schommer, C. (2003). Anwendung von Data Mining. Aachen: Shaker.

Säuberlich, F. (2000). KDD und Data Mining als Hilfsmittel zur Entscheidungsunterstützung. Wien: Lang.

Sturm, A.C. (2004). Looking for Revenue? Try tapping on your keyboard. Health Care Financial Management **58** (3): S.100-102.

Totok, A. (2000). Modellierung von OLAP- und Data-Warehouse-Systemen. Wiesbaden: Gabler.

Two Crows (1999). Introduction to Data Mining and Knowledge Discovery. Dritte Ausgabe. Potomac: Two Crows Corporation. URL: http://www.twocrows.com

Wesenauer, A. (2004). Controlling als Instrument der Unternehmenssteuerung und Organisationsentwicklung in der OÖGKK. Soziale Sicherheit, März 2004: S.100-106.

WHO (2001). The World Health Report 2001. Mental Health: New Understanding, New Hope. Genf: WHO.