FAKULTÄT FÜR !INFORMATIK

Entwurf und Entwicklung einer dynamischen Prüfkomponente für den Pseudonymisierungsgrad zum Zweck des Peer-to-Peer Austauschs medizinischer Daten in Forschung und Lehre

MAGISTERARBEIT

zur Erlangung des akademischen Grades

Magister

im Rahmen des Studiums

Informatikmanagement

eingereicht von

Pujan Shadlau, BSc.
Matrikelnummer 9725126

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Betreuer/Betreuerin: Univ.-Prof. Dipl.-Ing. Dr. techn. Thomas Grechenig

Wien, 20.07.2009 _____ _____
(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Entwurf und Entwicklung
einer dynamischen Prüfkomponente
für den Pseudonymisierungsgrad zum Zweck
des Peer-to-Peer Austauschs medizinischer Daten
in Forschung und Lehre

MAGISTERARBEIT

zur Erlangung des akademischen Grades

Magister

im Rahmen des Studiums

Informatikmanagement

eingereicht von

Pujan Shadlau, BSc.
9725126

ausgeführt am
Institut für Rechnergestützte Automation
Forschungsgruppe Industrial Software
der Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Betreuer: Univ.-Prof. Dipl.-Ing. Dr. techn. Thomas Grechenig
Mitwirkung: Mario Bernhart, Florian Fankhauser

Wien, 20.07.2009

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benützt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 20.07.2009 ----------

Name

Danksagung

Zu Beginn der Arbeit möchte ich eine Danksagung an alle Personen aussprechen, die mir während des Verfassens der Arbeit mit Rat und Tat in menschlicher sowie professioneller Sicht zur Seite gestanden sind.

Zu aller Erst gilt mein Dank meinen Eltern, die in schweren Zeiten einen schweren Schritt gewagt haben und ihre Heimat aufgegeben haben, um mir ein sorgloses Leben und eine Ausbildung ermöglichen zu können. Ihnen alleine habe ich es zu verdanken heute diese Danksagung verfassen zu können, in einer Magisterarbeit, der ein Studium vorausgegangen ist, welches durch ihre Unterstützung erst möglich geworden ist.

Besonderer Dank gilt auch meinem Betreuer Dipl. Ing. Florian Fankhauser, der mir immer mit professionellem Rat zur Seite stand, immer eine Lösung parat hatte und auch nie die Geduld mit mir verloren hat. Seiner Erfahrung und seiner Ambition habe ich es zu verdanken, dass die schriftliche Aufarbeitung des Themas in dieser heutigen Form existiert, inhaltlich sowie strukturell.

Ich bedanke mich außerdem bei meinen jahrelangen Freunden, meiner Freundin und weiteren Familienmitgliedern, die mich immer motivierend unterstützt haben und auch immer ein offenes Ohr hatten, egal zu welcher Zeit. Besonders Beatrix wusste mich immer wieder zu motivieren und mir das Ziel vor Augen zu führen, auch wenn es manchmal sehr verschwommen am Horizont erschien.

Last but not least bedanke ich mich selbstverständlich beim Institut für rechnergestützte Automation, Forschungsgruppe Industrial Software an der TU Wien - insbesondere bei Dr. Thomas Grechenig und Dipl. Ing. Mario Bernhart, die diese Arbeit ermöglicht haben.

Herzlichen Dank nochmals an alle!

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kurzfassung

Persönliche Daten bedürfen seit jeher besonderer sorgfältiger Behandlung. Da diese Daten im Regelfall sensibler Natur sind, ist es heutzutage mehr denn je ein essentielles Streben diese im elektronischen Datenverkehr zu anonymisieren bzw. zu pseudonymisieren, sodass sie keiner bestimmten Person zugeordnet werden können.

Zunächst werden vergleichend unterschiedliche aktuelle Anonymisierungs- und Pseudonymisierungsmethoden analysiert. Danach werden anhand eines Fallbeispiels, das an der Technischen Universität Wien umgesetzt wird, sowohl Aspekte, die für die Lehre wichtig sind, als auch Anforderungen, die für die Forschung relevant sind, analysiert und diskutiert. Darauf aufbauend wird ein konkreter Vorschlag für eine Prüfkomponente im Rahmen dieses Fallbeispieles gegeben. Aufgabe dieser Komponente ist die Überprüfung und die Bereitstellung einer Entscheidungsbasis für das Versenden von Daten, um sicherzustellen, dass diese keiner Person eindeutig zugeordnet werden können. Die Prüfkomponente stellt also ein gewisses Pseudonymisierungsniveau der Daten bezüglich ihrer Diversität und Anonymität sicher.

Die vorliegende Arbeit führt die unterschiedlichen aktuellen Methoden an, die es momentan im Bereich der Anonymisierung und Pseudonymisierung gibt. Dabei konnte eindeutig aufgezeigt werden, dass diese Methoden nicht ausreichen, um die Speicherung und Weitergabe sensibler Daten nach heutigen technischen Möglichkeiten, maximal sicher zu gestalten.

Abstract

Personal Data, concerning unique individuals, requires accurate treatment ever since. To avoid the disclosure of these unique individuals through the exploit of personal data during electronic exchange, certain steps can be taken, including anonymization- and pseudonymization-techniques.

In the first instance we focus on the comparison of current anonymization- and pseudonymization-techniques and then continue with the proposal for a verifying-component which is going to be realized within the scope of a case-study at the Technical University of Vienna. Purpose of the component is the delivery of a decision-base for the peer-to-peer exchange of medical data so no unique individuals can be identified during the exploit of this data. Therefore the component assures a certain standard of pseudonymity for the medical data in regards of diversity and anonymity.

Finally a proposal for the verifying-component is made in the last section of this scientific work. Research showed that current anonymization and pseudonymization methods are insufficient in terms of security and attacks on them, so that a maximum level of security cannot be guaranteed at this point of time.

1. Einführung

Nationale Datenschutzgesetze sind zum Schutz der Privatsphäre in vielerlei Hinsicht nicht mehr zeitgemäß. Sie sind im Grunde an der Vorstellung einer Datenverarbeitung (von personenbezogenen Daten) in Rechenzentren orientiert. In Zeiten des Internet, E-Commerce und computervermittelter Kommunikation - Phänomene der Globalisierung - sind nationale Gesetzgebungen unzureichend und müssen durch technische, juristische und organisatorische Maßnahmen zum Schutz der Privatsphäre und der Rechte von Individuen wie Organisationen ergänzt werden [51]. Vom heutigen Standpunkt aus kann behauptet werden ([66] [67] [68]), dass internationale und europäische Rechtsinstitutionen sich mit der Thematik des Schützens von persönlichen Daten auseinandersetzen und dies ein relevanter Punkt für sie ist . Um dieses Ziel durchzusetzen gibt es mehrere Möglichkeiten.
Die Anonymisierung und Pseudonymisierung von Daten für den elektronischen Gebrauch (Speicherung und Austausch) tragen zum gegenwärtigen Zeitpunkt u.a. einen großen Teil dazu bei, Interessen von Individuen hinsichtlich Privatsphäre und Verschluss von vertraulichen Informationen zu wahren. Da in einem breiten Spektrum der Dienstleistungsbranche die Verwendung persönlicher Daten in digitaler Form (Speicherung und Austausch) notwendig geworden ist, gewinnen diese Methoden immer weiter an Bedeutung. Beispielsweise ist im Finanzsektor (e-banking) und im Gesundheitsbereich (alleine in Österreich sind 8,4 Millionen eCards aktiv – Stand März 2009 [15]) die alleinige Verwendung von analogen und lokalen Daten kaum noch vorstellbar, weil Prozesse bereits stark vernetzt sind (Internet) und in Echtzeit abgewickelt werden müssen. Castells geht noch einen Schritt weiter: die Netzwerklogik, verkörpert durch das Internet, ist nun auf jeden Tätigkeitsbereich anwendbar, auf jeden Zusammenhang, an jedem Ort, der elektronisch angeschlossen werden kann [14].

Ein weiteres Beispiel für die Verwendung von personenbezogenen Daten in digitaler Form liefert die in Österreich im Einsatz befindliche Bürgerkarte [73]. Die Bürgerkarte ist ein Schlüssel der den Zugang zu diversen E-Government Angeboten (zum Beispiel Finanz-Online [77]) und zu Web-Diensten der Wirtschaft ermöglich. Die Bürgerkartenfunktion kann auf der e-card [15] oder der Bankomatkarte [75] aktiviert werden – auch Studentenausweise und Dienstausweise können mit der Bürgerkartenfunktion versehen werden.
In Hinblick auf Datenschutz und Datensicherheit werden auf der Bürgerkarten-Homepage die auf der Karte gespeicherten Daten erwähnt, welche abhängig vom Einsatzzweck sind. Unter anderem wird zusätzlich zum Vornamen und zum Nachnamen auch die Stammzahl (stark verschlüsselte Ableitung des Eintrages im Zentralen Melderegister [76]) gespeichert.

Die Stabstelle IKT–Strategie des österreichischen Bundes beschreibt in ihrer XML Definition der Personenbindung [74], letztere als integralen Bestandteil des Konzeptes Bürgerkarte. Sie ist eine von der Behörde signierte Struktur, welche ein eindeutiges Identifikationsmerkmal einer Person (zum Beispiel eine Registernummer) einem oder mehreren Zertifkaten dieser Person zuordnet. Die Personenbindung dient der eindeutigen, automatisierbaren Identifikation einer Person, wenn sie im Zuge eines Verfahrens an eine öffentliche Behörde herantritt. Eine solche Behörde (oder Amt) bildet abhängig von Name und Geburtsdatum das bereichspezifische Personenkennzeichen. Mit Hilfe dieses Kennzeichens können von der Behörde zusätzliche Serviceleistungen (zum Beispiel die Einsicht in laufende Verfahren) angeboten werden.

Um Leistung in Hinblick auf Komfort anzubieten, genügt es nicht nur Prozesse, die personenbezogene Daten in digitaler Form heranziehen, schnell und unkompliziert abzubilden. Die Verwendung dieser Daten erfordert zusätzlich die Einhaltung von Richtlinien, um Individuen zu schützen. Diese Richtlinien sind in Österreich im Datenschutzgesetz [5] festgehalten und können anhand unterschiedlicher Konzepte (u.a. Anonymisierung, Pseudonymisierung) realisiert werden.

Mittels unterschiedlicher Ansätze, die in der Arbeit erwähnt werden (Homogeneity- und Background-Knowledge Attacken), ist es möglich trotz Einhaltung von Richtlinien und Verwendung von Anonymisierungsmethoden an geschützte Informationen zu gelangen. Um solchen Attacken entgegenzuwirken wurde eine Vielzahl an Konzepten entworfen, welche die gegenwärtigen Methoden erweitern, modifizieren und zuverlässiger machen sollen. Diese Konzepte beinhalten u.a. Vorschläge von Forschungskreisen (TMF-Datenschutzkonzept, Kapitel 3.1.1 und 3.1.2) für pseudonymisierte Datenbanken und unterschiedliche Methoden der k-Anonymisierung (Kapitel 3.3).

Im Projekt SPICS Soul werden ebenfalls medizinische Daten ausgetauscht. Um eine Entscheidungsbasis für die Weitergabe solcher Daten geben zu können, wird in dieser Arbeit ein Entwurf für eine Prüfkomponente vorgestellt, welche aufgrund globaler und lokaler Daten einen Vorschlag liefert. Die Einhaltung der dezentralen Architektur spielt hier eine große Rolle und wird bei der Empfehlung berücksichtigt. Jede Forschungseinrichtung soll ihre eigene Datenbank und ihre eigene Komponente haben und Daten unabhängig von anderen Einrichtungen speichern.

Der Vorschlag basiert auf den in der Arbeit vorgestellten Methoden der Pseudonymisierung (Kapitel 3.1), k-Anonymisierung (Kapitel 3.3) und der l-Diversität (Kapitel 3.5) und soll gegebenenfalls Defizite aufzeigen welche mit den vorgestellten Mitteln nicht bewältigt werden können. Fortführend werden die Einsatzzwecke Forschung und Lehre behandelt, da es auch in diesem Zusammenhang ausschlaggebend ist welche Methoden man zum Einsatz in Erwägung zieht.

Für die Durchführung der Arbeit wurde Fachliteratur aus den Bereichen der medizinischen Forschungsarbeit, Mathematik, Statistik, Algorithmik, Anonymisierung und Pseudonymisierung verwendet. Relevante Gesetze und Richtlinien wurden ebenfalls zitiert.

Die Arbeit gliedert sich wie folgt:

Zuerst werden die Grundlagen dargestellt, um das Umfeld der Anonymisierung und Pseudonymisierung zu verstehen. Es werden Terminologien vorgestellt und Begriffe erklärt, weiters soll ein Einblick in die bestehenden, gängigen Methoden und Konzepte der Pseudonymisierung und Anonymisierung gegeben werden. Fortführend erfolgt die nähere Betrachtung dieser Begriffe im Kontext des Datenaustauschs, insbesondere im medizinischen Umfeld. Hierzu sind selbstverständlich auch die Beleuchtung der momentanen Gesetzeslage in Österreich und Europa sowie die Vorstellung korrespondierender Richtlinien mit Bezug zu diesem Thema notwendig. Auch die Hervorhebung der *l*-Diversität als Weiterentwicklung der *k*-Anonymität spielt eine entscheidende Rolle.

Nachdem die notwendigen Grundlagen definiert worden sind, folgt die Vorstellung von SPICS Soul und die Anforderungsanalyse dieses Projektes. Hier werden die vorgestellten Mechanismen und Modelle der einführenden Kapitel auf die Bedürfnisse in SPICS Soul adaptiert. Im Zuge dessen erfolgen auch eine Erörterung der Problemstellung und ein daraus resultierender konkreter Vorschlag für die Komponente. Das Modell welches hier ausgearbeitet wird, ist die Voraussetzung für die Reflexion, das Conclusio und einem Ausblick, dem letzten Teil der Arbeit.

2. Grundlagen

Um die Konzepte und Möglichkeiten, welche in dieser Arbeit vorgestellt werden, im korrekten Kontext verstehen zu können, ist eine vorhergehende Erklärung der Begrifflichkeiten im Umfeld der Anonymisierung und Pseudonymisierung unabdinglich.
Wenn nicht anders angegeben, ist mit dem Begriff Daten/Datensatz personenbezogene Information gemeint, mit Hilfe derer man eine Person eindeutig identifizieren kann. Diese Daten werden in Tabellenform gespeichert und bestehen aus Spalten (Attributen) und Zeilen (Tupel). Eine Zeile enthält jeweils einen Datensatz welcher sich wiederum aus mehreren Spalten zusammensetzt. Die Spalten enthalten personenbezogene Daten, wie beispielsweise den Namen, das Geburtsdatum oder eine Krankheit der jeweiligen Person und können somit verschiedene Werte annehmen.
Tupel einer Tabelle sind nicht zwingend einmalig, hingegen können Attribute jeweils nur einmal pro Tabelle vorkommen [3].

2.1. IT-Sicherheit

Sicherheit besteht bei informationstechnischen (IT-) Systemen darin, dass Schutzziele wie Vertraulichkeit, Integrität, Verfügbarkeit und Verbindlichkeit trotz intelligenter Angreifer durchgesetzt werden [61]. Insbesondere bei offenen Kommunikationssystemen kann nicht davon ausgegangen werden, dass sich Beteiligte vollständig vertrauen. Bei der Analyse von deren Sicherheit sind prinzipiell alle Beteiligten potentiell auch als Angreifer zu betrachten [1]. Bruce Schneier [60] geht davon aus, dass jedes IT-System, welches in unserer Realität existiert, eine komplizierte Serie von Verbindungen untereinander ist. Diese Systeme müssen in ihren Komponenten und Verbindungen von Sicherheit durchzogen und geprägt sein.

Seit den frühen achtziger Jahren findet sich eine Dreiteilung der Bedrohungen und deren korrespondierenden Schutzziele [1], [1]:

- Unbefugter Informationsgewinn (Bedrohung), Verlust der Vertraulichkeit, Vertraulichkeit ist das Schutzziel: Patientendaten (Untersuchungsdaten, Diagnosen, Behandlungsdaten) sollen Unbefugten (Angreifern, seien dies nun andere Patienten, Mitarbeiter des Netzbetreibers über dessen Netz sie übertragen werden oder Außenstehende) nicht zur Kenntnis gelangen.

- Unbefugte Modifikation von Informationen (Bedrohung), Verlust der Integrität, Integrität ist das Schutzziel: Werden unbefugt oder unbemerkt Daten verändert, kann dies zum Beispiel im Falle einer Medikamentendosis tödliche Folgen haben.
- Unbefugte Beeinträchtigung der Funktionalität (Bedrohung), Verlust der Verfügbarkeit, Verfügbarkeit ist das Schutzziel: Der Verlust der Verfügbarkeit einer Patientenakte oder Behandlungsdaten kann zu kritischen Zeitpunkten, in denen diese unabdinglich für die weitere Behandlung sind, ebenfalls tödlich sein.

In [52] wird kritisiert, dass die Bedrohung *Verlust der Verbindlichkeit* in [1] nicht erwähnt wird. Das Schutzziel der Verbindlichkeit wird in den kanadischen IT-Sicherheitsevaluationskriterien [53] [54] als zusätzliches Viertes eingeführt. Da in SPICS-Soul die Zurechenbarkeit der Ärzte eine große Rolle spielt, da sie schlussendlich entscheiden ob Daten weitergegeben werden dürfen oder nicht, wird diese Bedrohung hier ebenfalls erwähnt:

- Unzulässige Unverbindlichkeit (Bedrohung), d.h. Verlust der Zurechenbarkeit, Zurechenbarkeit ist das Schutzziel: Wenn für Vorgänge in IT-Systemen, etwa für den Versand von Diagnosen oder Abrechnungen, nicht die jeweils Verantwortlichen auszumachen sind, kann dies zu verantwortungslosem Handeln führen. Außerdem können die Folgen eines Fehlers für die Geschädigten noch verschlimmert werden, weil möglicherweise unklar bleibt, an wen sie sich mit den Schadenersatzansprüchen zu halten haben.

2.2. Risiko

Der weit verbreitete mathematische Ansatz [16], der auf den Modellen der quantitativen Risikokalkulation beruht [17], definiert Risiko (R) formal als das Produkt aus dem Ausmaß eines negativen Ereignisses (A) und der Eintrittswahrscheinlichkeit (E) [18]:

$$R = A \times E$$

Formel 1: formale Definition von Risiko [18]

Risiko wird in dieser Arbeit im Kontext der Eintrittswahrscheinlichkeit eines erfolgreichen Angriffes auf pseudonymisierte und anonymisierte Daten erwähnt. Das Risiko eines solchen Angriffes soll mit Hilfe der vorgestellten Methoden möglichst gering gehalten werden. Ein solches Risiko entsteht

durch Hintergrundwissen oder unzureichende Methoden der *k*-Anonymisierung, aber auch durch die Verwendung von Zuordnungslisten bei der Pseudonymisierung, welche die Pseudonyme und deren Zuordnung zu individuellen Personen enthalten.

Sei $T = \{t_1, t_2, ..., t_n\}$ eine Tabelle welche aus den Datensätzen (Tupeln) $t_1, t_2, ..., t_n$ besteht. Die Spalten werden mit $A_1, A_2, ..., A_m$ bezeichnet und bilden zusammen mit den Tupeln die erwähnte Tabelle T, welche wiederum eine Untermenge einer größeren Population Ω bildet. A bezeichnet den Satz aller Attribute $\{A_1, A_2, ..., A_m\}$ der Tabelle T. Die Schreibweise $t_i[A_i]$ wird verwendet, um den Wert eines Attributes A_i des Tupel t zu bezeichnen.
Ist $C = \{C_1, C_2, \ldots, C_m\}$ eine Teilmenge von A, wird die Notation $t[C]$ für das Tupel $(t[C_1], \ldots, t[C_p])$ verwendet, welches die Projektion von t auf die Attribute in C ist.
S bezeichnet das Set aller sensiblen Attribute. Ein sensibles Attribut wäre eine persönliche Meinung oder die Krankheit eines Patienten, welche einen direkten Nachteil, zum Beispiel im Berufsleben, bei Bekanntwerden verursachen kann. Insofern muss die Relation zwischen den Patienten und den Krankheiten unter denen sie leiden geheim gehalten werden. So ist beispielsweise zu vermeiden, dass jeder sofort sehen kann dass ein Patient an Krebs leidet. Andererseits ist es möglich dass das Krankenhaus publik macht, dass es in dieser Einrichtung Krebspatienten gibt. Alle nicht-sensiblen Attribute werden mit N bezeichnet [3].

Zur Veranschaulichung ist im Folgenden eine Tabelle T (Tabelle 1) mit 12 Tupeln und 4 Attributen $(A = \{Plz, Alter, Nationalität und Diagnose\})$ dargestellt, die Daten enthält wie sie in einer medizinischen Akte vorkommen können (alle Namen und Daten in dieser Arbeit sind frei erfunden):

	Non-sensible Attribute			*Sensible Attribute*
	PLZ	**Alter**	**Nationalität**	**Diagnose**
1	1030	28	USA	Alkoholismus
2	1010	29	Österreich	Alkoholismus
3	1011	21	Japan	Manisch Depressiv
4	1030	23	Österreich	Manisch Depressiv
5	1110	50	Indien	Drogenabhängig
6	1110	55	USA	Alkoholismus
7	1150	47	Österreich	Manisch Depressiv
8	1140	49	Österreich	Manisch Depressiv
9	1030	31	Österreich	Drogenabhängig
10	1030	37	Indien	Drogenabhängig
11	1010	36	Japan	Drogenabhängig
12	1012	35	Österreich	Drogenabhängig

Tabelle 1: personenbezogene, medizinische Daten verschiedener Personen in einer Tabelle

2.3. Identifikationsmerkmale

Als Identifikationsmerkmale bezeichnet man Attribute mit Hilfe derer man Personen identifizieren kann. Je nach Art der Attribute unterscheidet man zwischen primären und sekundären Identifikationsmerkmalen. Treten sekundäre Identifikationsmerkmale alleine auf, reichen diese in dieser Form für eine Identifikation nicht aus. Eine genaue Definition der Begriffe primäre und sekundäre Identifikationsmerkmale ist im folgenden Abschnitt gegeben.

2.3.1. Primäre Identifikationsmerkmale

Als primäre Identifikationsmerkmale einer Person werden jene Attribute bezeichnet, mit Hilfe derer man eine Person eindeutig identifizieren kann[3]. Um den direkten Personenbezug unwiederherstellbar zu machen, ist es notwendig solche Attribute zu verschlüsseln oder zu entfernen. Eine Entfernung dieser Attribute reicht zur Gewährleistung vollkommener Anonymität in vielen Fällen (Kapitel 3.2.1) aber nicht aus. Solche Attribute treten in vielen Fällen (Taufurkunde, Staatsbürgerschaftsnachweis,...) zusammen mit sensiblen Daten laut untenstehender Definition auf. Primäre, identifizierende Attribute sind zum Beispiel der Vor- und Nachname oder die Versicherungsnummer einer Person.

Laut §4 Abs. 2 DSG2000 sind sensible Daten wie folgt definiert:

§ 4 Im Sinne der folgenden Bestimmungen dieses Bundesgesetzes bedeuten die Begriffe:
2. "sensible Daten" ("besonders schutzwürdige Daten"): Daten natürlicher Personen über ihre rassische und ethnische Herkunft, politische Meinung, Gewerkschaftszugehörigkeit, religiöse oder philosophische Überzeugung, Gesundheit oder ihr Sexualleben; [5]

2.3.2. Sekundäre Identifikationsmerkmale

Sekundäre Identifikationsmerkmale sind solche Attribute, die durch alleiniges Auftreten nicht ausreichen, um eine Person eindeutig identifizieren zu können [3]. Eine Kombination sekundärer Identifikationsmerkmale kann jedoch zu einer eindeutigen Identifikation einer Person führen.
Ein Fall in den USA [55] zeigte auf, wie die Kombination von sekundären Identifikationsmerkmalen zur nahezu eindeutigen Identifikation von 87% der dortigen Bevölkerung führen kann. Diese Identifikation war durch die Verlinkung der Attribute *Geschlecht*, *Geburtsdatum* und *Postleitzahl* mit den Wahlkreis-Daten im Staat Massechusettes, welche die 3 oben genannten Attribute und zusätzlich den

Namen enthielten, möglich. Diese so genannte *Linking Attack* konnte die medizinische Akte des Gouverneurs von Massechusetts eindeutig identifizieren [37].

Ein Set von nicht-sensiblen Attributen $\{Q_1,\ldots,Q_m\}$ einer Tabelle wird als *Quasi-Identifier* [20] bezeichnet, wenn diese Attribute mit anderen externen Daten dazu verlinkt werden können, um mindestens eine Person in der Population Ω eindeutig zu identifizieren. Das Set aller *Quasi-Identifiers* wird als *QI* notiert.

Laut Datenschutzgesetz in der geltenden Fassung [5] sind nicht sensible Daten in §4 Abs. 1 wie folgt definiert:

***§ 4** Im Sinne der folgenden Bestimmungen dieses Bundesgesetzes bedeuten die Begriffe:*

1. "Daten" ("personenbezogene Daten"): Angaben über Betroffene (Z 3), deren Identität bestimmt oder bestimmbar ist; "nur indirekt personenbezogen" sind Daten für einen Auftraggeber (Z 4), Dienstleister (Z 5) oder Empfänger einer Übermittlung (Z 12) dann, wenn der Personenbezug der Daten derart ist, dass dieser Auftraggeber, Dienstleister oder Übermittlungsempfänger die Identität des Betroffenen mit rechtlich zulässigen Mitteln nicht bestimmen kann; [5]

2.4. Anonymität

Nach [1] bezeichnet Anonymität den Status der Unidentifizierbarkeit innerhalb einer gewissen Anzahl von Subjekten oder Personen, dem *Anonymity-Set* [21]. Bei Anonymität besteht ein Bezug nur auf eine Gesamtmenge von Individuen.

Um die Anonymität eines Subjektes zu gewährleisten, muss eine ausreichende Anzahl an anderen Subjekten mit potenziell gleichen Attributen vorhanden sein.

	Non-sensible Attribute		
	PLZ	**Alter**	**Nationalität**
1	1***	28	*
2	1***	29	*
3	1***	21	*
4	1***	23	*
5	1***	50	*
6	1***	55	*
7	1***	47	*
8	1***	49	*
9	1***	31	*
10	1***	37	*
11	1***	36	*
12	1***	35	*

Tabelle 2: Anonymisierte Tabelle 1

Das *Anonymity-Set* ist das Set aller möglichen Personen. Dieses Set der möglichen verdächtigen Personen ist abhängig vom Wissen des Angreifers, der Interesse an den Daten hat. Es ist zum Beispiel ein Unterschied ob die Angreifer die Personen und deren sensiblen Attribute kennen und ein gewisses Hintergrundwissen (z.B. demographische Informationen) vorhanden ist, oder ob der Angreifer nicht über dieses Wissen verfügt und er quasi blind raten muss. Je nachdem wie viele Informationen bekannt sind, spricht man von verschiedenen Stufen von Anonymität [1]. Das *Anonymity-Set* ändert sich im Laufe der Zeit und ist deshalb dynamisch. Man kann also festhalten, dass Anonymität sehr kontextabhängig ist (Anzahl der Attribute, Diversität der Attribute, Population, Zeitraum,…) [1]. In Tabelle 2 ist zu sehen, dass sekundäre Identifikationsmerkmale weggelassen wurden und eine eindeutige Identifizierung ohne zusätzliche externe Daten nicht möglich ist. Da die Anonymisierung meist nicht ausreicht, um den Personenbezug unwiederherstellbar zu entfernen (hier werden nur primäre Identifikationsmerkmale entfernt – in [1] wird gezeigt, dass das Verlinken sekundärer Identifikationsmerkmale ebenfalls zu einer eindeutigen Identifikation führen kann), wird im nächsten Kapitel das Konzept der Pseudonymisierung vorgestellt. Die Anonymisierung wird demnach im medizinisch-wissenschaftlichen Bereich bzw. in der Lehre eine größere Akzeptanz finden als im klinischen Alltag, wo eine (befugte) Rückidentifikation (welche durch das Weglassen der primären Attribute erheblich erschwert oder auch unmöglich wird) durchaus gewünscht sein kann.

2.5. Pseudonymität

Pseudonym Systems wurden erstmals 1985 von Chaum [22] vorgestellt und sollen den Benutzern effektives aber gleichzeitig anonymes Arbeiten in Kooperation mit anderen Organisationen durch den Einsatz von Pseudonymen [1] ermöglichen. Verschiedene Möglichkeiten von *Pseudonym Systems* werden von Chaum und Evertse [23], Damgård [24] sowie Chen [25] vorgestellt. Für genauere Definition dieser wird an dieser Stelle auf deren Publikationen verwiesen, da sie für die weitere Arbeit nicht erforderlich ist.

Nach [1] ist Pseudonymität der Status der Verwendung von Pseudonymen (Kennzeichen) an Stelle von primären Identifikationsmerkmalen oder realen Namen. Diese Pseudonyme werden nach gewissen Regeln bzw. Algorithmen gebildet.

Um die Pseudonymität [1] eines Subjektes zu gewährleisten sind primäre Identifikationsmerkmale, die eine eindeutige Identifizierung ermöglichen, zu entfernen und durch Pseudonyme zu ersetzen –

somit wird die Identität des Subjektes bei der Pseudonymisierung nicht Preis gegeben. Pseudonyme könnten beispielsweise fortlaufende Nummern oder Buchstaben sein oder auch Kombinationen aus diesen, die nach Pseudonymisierungsregeln (einem Pseudonymisierungsalgorithmus bzw. einem Pseudonymisierungsmodell) aus den primären Attributen (einem oder mehreren) erstellt werden. Um die bevollmächtige, valide Depseudonymisierung (Rückschluss auf die Originaldaten) zu gewährleisten, wird eine entsprechende Zuordnungstabelle benötigt. In solch einer Tabelle sind die identifizierenden Attribute zusammen abgespeichert und können zu einer Rückidentifizierung verwendet werden [19].

„Zweck der Pseudonymisierung - im Unterschied zur Anonymisierung - ist zum ersten, Daten zu einem Fall aus verschiedenen Quellen oder von verschiedenen Zeitpunkten zusammenführen zu können, und zum zweiten, die "Fährte zum Fall" für besondere Anlässe offen zu halten.“ [1]

Eine solche offen gehaltene Fährte zum Fall (Depseudonymisierung) ist beispielsweise im medizinischen Bereich zur Verständigung eines betroffenen Patienten wünschenswert. Im Projekt SPICS kann die Pseudonymisierung der Daten eine wichtige Grundlage für den Austausch dieser zwecks medizinischer Forschung bilden da die Daten an verschiedenen Stellen (Forschungseinrichtungen) gespeichert sind. Eine Rückidentifizierung (Depseudonymisierung) ist im Gegensatz zur Anonymisierung durchaus gewollt (im Falle der Notwendigkeit zur Verständigung von Patienten) oder unbefugt (mit Hilfe von Inferenzen) möglich [1]. Infolge dessen ist rechtlich betrachtet keine Äquivalenz zwischen Anonymisierung und Pseudonymisierung zu erkennen, da bei Letzterer auf jeden Fall eine Einwilligung des Probanden notwendig ist. Diese Einwilligung entfällt bei der Anonymisierung, da sensible Daten (primäre Identifikationsmerkmale) vollständig entfernt werden. Eine solche Einwilligung erfolgt im Regelfall schriftlich und muss, entsprechend dem österreichischen Datenschutzgesetz (Kapitel 2.6.1) den Verwendungszweck der Daten festlegen.

2.6. Gesetzliche Grundlagen, Richt- und Leitlinien

Das Datenschutzgesetz 2000 (DSG 2000), BGBl. I Nr. 165/1999 idF. BGBl. I Nr. 13/2005, ist das geltende österreichische Datenschutzgesetz, und damit die wichtigste Rechtsvorschrift zum Datenschutz in Österreich. [5]
Es beinhaltet folgende für die Arbeit relevante Abschnitte, die in diesem Teil vorgestellt werden, um eine Basis für den Entwurf der Prüfkomponente bereit zu stellen. Alle zitierten Paragraphen wurden, so wie sie im DSG2000 aufgeführt sind, übernommen. Im Anschluss folgt die Vorstellung zweier

Richtlinien –MAGDA-LENA und HIPAA. MAGDA-LENA ist eine Richtlinie für ein logisches österreichisches Gesundheitsdatennetz, HIPAA ist eine staatliche Richtlinie welche in den USA Anwendung findet. Beide Richtlinien zeigen Möglichkeiten für technische und organisatorische Rahmenbedingungen wie sie auch in SPICS Soul auftreten können.

2.6.1. Datenschutzgesetz von 2000

Das österreichische Datenschutzgesetz in seiner geltenden Fassung [5] beinhaltet Richtlinien, die für die Verwendung von personenbezogenen Daten beachtet werden müssen. Weiters erfolgt per Definition die Zulässigkeit der Verwendung dieser Daten und es regelt das Grundrecht auf Datenschutz für Jedermann.

2.6.1.1. Grundrecht auf Datenschutz

Dieser erste Paragraph erläutert, das Recht auf Geheimhaltung personenbezogener Daten:

§ 1 *(1) Jedermann hat, insbesondere auch im Hinblick auf die Achtung seines Privat- und Familienlebens, Anspruch auf Geheimhaltung der ihn betreffenden personenbezogenen Daten, soweit ein schutzwürdiges Interesse daran besteht. Das Bestehen eines solchen Interesses ist ausgeschlossen, wenn Daten infolge ihrer allgemeinen Verfügbarkeit oder wegen ihrer mangelnden Rückführbarkeit auf den Betroffenen einem Geheimhaltungsanspruch nicht zugänglich sind.*

Weiters wird definiert, dass Jedermann das Recht auf Auskunft darüber hat, welche Daten über ihn verarbeitet werden, woher diese stammen, wozu sie verwendet werden und an wen sie übermittelt werden:

(3) *Jedermann hat, soweit ihn betreffende personenbezogene Daten zur automationsunterstützten Verarbeitung oder zur Verarbeitung in manuell, d.h. ohne Automationsunterstützung geführten Dateien bestimmt sind, nach Maßgabe gesetzlicher Bestimmungen*

> *1. das Recht auf Auskunft darüber, wer welche Daten über ihn verarbeitet, woher die Daten stammen, und wozu sie verwendet werden, insbesondere auch, an wen sie übermittelt werden;*

2. das Recht auf Richtigstellung unrichtiger Daten und das Recht auf Löschung unzulässigerweise verarbeiteter Daten.

2.6.1.2. Verwendung von Daten

Personenbezogene und nicht personenbezogene Daten werden zum Zweck des Austauschs und der Weitergabe in verschiedenen Projekten in verschiedenen Bereichen und in unterschiedlichen Zusammen hängen verwendet. Für diese Arbeit besonders relevant ist die Verwendung für medizinisch-wissenschaftliche Forschungszwecke. Da für den Austausch der Daten auch das Speichern dieser notwendig ist, muss beachtet werden inwiefern die Verwendung dieser Daten durch Gesetze geregelt ist, um die Interessen der Patienten zu wahren.

Laut §4 Z8 DSG2000 wird die Verwendung von Daten wie folgt definiert:
"Verwenden von Daten": jede Art der Handhabung von Daten einer Datenanwendung, also sowohl das Verarbeiten (Z 9) als auch das Übermitteln (Z 12) von Daten;
Somit lassen sich folgende, für SPICS Soul relevante, Zusätze hervorheben da im Laufe der Kooperation von medizinisch-wissenschaftliche Forschungseinrichtungen Aktionen wie die Abfrage und die Ausgabe, sowie das Überlassen und Übermitteln von Daten auftreten.

§4 Z9 DSG2000 Verarbeiten von Daten:
"Verarbeiten von Daten": das Ermitteln, Erfassen, Speichern, Aufbewahren, Ordnen, Vergleichen, Verändern, Verknüpfen, Vervielfältigen, Abfragen, Ausgeben, Benützen, Überlassen (Z 11), Sperren, Löschen, Vernichten oder jede andere Art der Handhabung von Daten einer Datenanwendung durch den Auftraggeber oder Dienstleister mit Ausnahme des Übermittelns (Z 12) von Daten;

§4 Z11 DSG2000 Überlassen von Daten:
die Weitergabe von Daten vom Auftraggeber an einen Dienstleister;

§4 Z12 DSG2000 Übermitteln von Daten:
die Weitergabe von Daten einer Datenanwendung an andere Empfänger als den Betroffenen, den Auftraggeber oder einen Dienstleister, insbesondere auch das Veröffentlichen solcher Daten; darüber hinaus auch die Verwendung von Daten für ein anderes Aufgabengebiet des Auftraggebers;

In Abschnitt 2 §6 DSG2000 werden die Grundsätze für die Verwendung von Daten festgelegt:

*§ **6** (1) Daten dürfen nur*

1. nach Treu und Glauben und auf rechtmäßige Weise verwendet werden;

2. für festgelegte, eindeutige und rechtmäßige Zwecke ermittelt und nicht in einer mit diesen Zwecken unvereinbaren Weise weiterverwendet werden; die Weiterverwendung für wissenschaftliche oder statistische Zwecke ist nach Maßgabe der §§ 46 und 47 zulässig;

3. soweit sie für den Zweck der Datenanwendung wesentlich sind, verwendet werden und über diesen Zweck nicht hinausgehen;

4. so verwendet werden, dass sie im Hinblick auf den Verwendungszweck im Ergebnis sachlich richtig und, wenn nötig, auf den neuesten Stand gebracht sind;

5. solange in personenbezogener Form aufbewahrt werden, als dies für die Erreichung der Zwecke, für die sie ermittelt wurden, erforderlich ist; eine längere Aufbewahrungsdauer kann sich aus besonderen gesetzlichen, insbesondere archivrechtlichen Vorschriften ergeben.

Abschnitt 8, der besondere Verwendungszwecke von Daten definiert, beinhaltet §46 welcher die Verwendung für wissenschaftliche Forschung, wie sie in SPICS Soul auch auftritt, und Statistik regelt:

*§ **46** (1) Für Zwecke wissenschaftlicher oder statistischer Untersuchungen, die keine personenbezogenen Ergebnisse zum Ziel haben, darf der Auftraggeber der Untersuchung alle Daten verwenden, die*

> *1. öffentlich zugänglich sind oder*
>
> *2. der Auftraggeber für andere Untersuchungen oder auch andere Zwecke zulässigerweise ermittelt hat oder*
>
> *3. für den Auftraggeber nur indirekt personenbezogen sind.*
>
> *Andere Daten dürfen nur unter den Voraussetzungen des Abs. 2 Z 1 bis 3 verwendet werden.*

Im Fall von SPICS Soul sei hier noch erwähnt, dass aufgrund der verschiedenen Attacken, die es auf *k*-Anonymität gibt, ungewollter Zugriff auf personenbezogene Daten trotzdem möglich ist weshalb weitere Maßnahmen (*l*-Diversität) Anwendung finden müssen. Details dazu sind in Kapitel 3.5 beschrieben.

(2) Bei Datenanwendungen für Zwecke wissenschaftlicher Forschung und Statistik, die nicht unter Abs. 1 fallen, dürfen Daten, die nicht öffentlich zugänglich sind, nur

1. gemäß besonderen gesetzlichen Vorschriften oder

2. mit Zustimmung des Betroffenen oder

3. mit Genehmigung der Datenschutzkommission gemäß Abs. 3 verwendet werden.

(3) Eine Genehmigung der Datenschutzkommission für die Verwendung von Daten für Zwecke der wissenschaftlichen Forschung oder Statistik ist zu erteilen, wenn

1. die Einholung der Zustimmung der Betroffenen mangels ihrer Erreichbarkeit unmöglich ist oder sonst einen unverhältnismäßigen Aufwand bedeutet und

2. ein öffentliches Interesse an der beantragten Verwendung besteht und

3. die fachliche Eignung des Antragstellers glaubhaft gemacht wird.

Sollen sensible Daten übermittelt werden, muss ein wichtiges öffentliches Interesse an der Untersuchung vorliegen; weiters muss gewährleistet sein, dass die Daten beim Empfänger nur von Personen verwendet werden, die hinsichtlich des Gegenstandes der Untersuchung einer gesetzlichen Verschwiegenheitspflicht unterliegen oder deren diesbezügliche Verlässlichkeit sonst glaubhaft ist. Die Datenschutzkommission kann die Genehmigung an die Erfüllung von Bedingungen und Auflagen knüpfen, soweit dies zur Wahrung der schutzwürdigen Interessen der Betroffenen, insbesondere bei der Verwendung sensibler Daten, notwendig ist.

(4) Rechtliche Beschränkungen der Zulässigkeit der Benützung von Daten aus anderen, insbesondere urheberrechtlichen Gründen bleiben unberührt.

(5) Auch in jenen Fällen, in welchen gemäß den vorstehenden Bestimmungen die Verwendung von Daten für Zwecke der wissenschaftlichen Forschung oder Statistik in personenbezogener Form zulässig ist, ist der direkte Personenbezug unverzüglich zu verschlüsseln, wenn in einzelnen Phasen der wissenschaftlichen oder statistischen Arbeit mit nur indirekt personenbezogenen Daten das Auslangen gefunden werden kann. Sofern gesetzlich nicht ausdrücklich anderes vorgesehen ist, ist der Per-

sonenbezug der Daten gänzlich zu beseitigen, sobald er für die wissenschaftliche oder statistische Arbeit nicht mehr notwendig ist.

2.6.1.3. Zulässigkeit der Verwendung der Daten

§ 7 (1) Daten dürfen nur verarbeitet werden, soweit Zweck und Inhalt der Datenanwendung von den gesetzlichen Zuständigkeiten oder rechtlichen Befugnissen des jeweiligen Auftraggebers gedeckt sind und die schutzwürdigen Geheimhaltungsinteressen der Betroffenen nicht verletzen.

(2) Daten dürfen nur übermittelt werden, wenn

1. sie aus einer gemäß Abs. 1 zulässigen Datenanwendung stammen und

2. der Empfänger dem Übermittelnden seine ausreichende gesetzliche Zuständigkeit oder rechtliche Befugnis - soweit diese nicht außer Zweifel steht - im Hinblick auf den Übermittlungszweck glaubhaft gemacht hat und

3. durch Zweck und Inhalt der Übermittlung die schutzwürdigen Geheimhaltungsinteressen des Betroffenen nicht verletzt werden.

(3) Die Zulässigkeit einer Datenverwendung setzt voraus, dass die dadurch verursachten Eingriffe in das Grundrecht auf Datenschutz nur im erforderlichen Ausmaß und mit den gelindesten zur Verfügung stehenden Mitteln erfolgen und dass die Grundsätze des § 6 eingehalten werden

Die Definition über die Zulässigkeit der Verwendung der Daten regelt somit, dass keine unbefugten Institutionen Daten verarbeiten und die schutzwürdigen Geheimhaltungsinteressen der Betroffenen verletzen. Somit müssen im Forschungsverbund von SPICS Soul alle Einrichtungen diese Voraussetzung erfüllen, um am Austausch der Daten und den Forschungsprojekten teilnehmen zu können.

2.6.2. MAGDA-LENA

Magda-Lena (Medizinisch-Administrativer Gesundheitsdatenaustausch – Logisches und Elektronisches Netzwerk Austria) [9] stellt eine Richtlinie dar, die technische und organisatorische Rahmenbedingungen für ein logisches österreichisches Gesundheitsdatennetz beschreibt. Diese Richtlinie wurde von der STRING- (Standards und Richtlinien für den Informatikeinsatz im österreichischen Gesundheitswesen) Kommission [10] beim Bundesministerium für soziale Sicherheit und Generationen erarbeitet. Das es sich hierbei aber um eine Richtlinie [26] handelt, ist diese nicht rechtliche bin-

dend bzw. keine verbindliche Vorschrift, sondern konkretisiert als Orientierungsmarke Handlungsempfehlungen [27].

MAGDA-LENA ist als die Verbindung von Einrichtungen (Leistungsanbieter, Leistungserbringer, Administration, Kostenträger,…) des Gesundheits- und Sozialwesens zum Zweck des elektronischen Datenaustausches, direkt oder indirekt personenbezogener, multimedialer Informationen zu verstehen. Daten ohne Patientenbezug fallen nicht unter die MAGDA-LENA Rahmenbedingungen.
MAGDA-LENA ist keine Norm, sondern enthält Richtlinien, um Schnittstellenprobleme zwischen verschiedenen Leistungsanbietern zu vermeiden. Für die Teilnahme an diesem logischen Gesundheitsdatennetz ist die Einhaltung von Richtlinien verbindlich, u.a. die Einhaltung des Datenschutzgesetzes [9].

Allgemeine Ziele von MAGDA-LENA sind u.a. [9]:

- der elektronische Austausch von Gesundheitsdaten zum Zweck der effizienten Behandlung von Patienten
- Zugang zu den für die aktuelle Versorgung der Patienten notwendigen Informationen für jeden Teilnehmer
- Verminderung von Redundanzen von identischen Informationen
- Verbesserung der Versorgungsqualität, Erhöhung der Effizienz der Patientenbetreuung bei gleichzeitig daraus resultierender Kostenreduktion

Die Stellung des Patienten wird explizit erwähnt. Alle Rahmenbedingungen in MAGDA-LENA haben die Wahrung der Interessen des Patienten in Bezug auf Geheimhaltung und Weitergabe der Daten nur auf Basis bestehender gesetzlichen Regelungen zum Ziel. Der Patient kann seine Person betreffende Daten jederzeit einsehen. Wenn möglich soll der Austausch der Daten immer in indirekt personenbezogener Form erfolgen, sodass kein Bezug zu einer individuellen Person hergestellt werden kann.

Bereits vorhandene internationale oder nationale Standards der Medizinischen Informatik sind im Rahmen von MAGDA-LENA als Grundlage für Standardnachrichten für die Kommunikation zwischen Partnern im Gesundheitswesen unbedingt zu verwenden. Existieren solche Standards noch nicht, sieht MAGDA-LENA die Entwicklung derer nach einem Vorgehensmodell vor, welches im Folgenden in groben Punkten beschrieben wird [9].

1. Spezifikation des Anwendungsbereiches (Beschreibung des Anwendungsbereiches, der betroffenen Parteien und die wichtigsten Kommunikationspartner bzw. Fachgebiete)
2. Benutzeranforderungen und Anwendungsfälle (Anforderungsanalyse für den elektronischen Informationsaustausch)
3. Kommunikationsrollen und unterstützte Dienste (Durchführung eines Abstraktionsprozesses von den Benutzeranforderungen und den einzelnen Anwendungsfällen ausgehend)
4. Domäneninformationsmodell (Darstellung eines statischen Modells des Anwendungsbereiches, das die Klassen, deren Attribute sowie die Beziehungen zwischen den Klassen erfasst)
5. Allgemeine Nachrichtenbeschreibungen (Beschreibung der kommunikationsspezifischen Sichten des Informationsmodells und Darstellung des Informationsgehalts und der semantischen Struktur einer Nachricht)
6. Allgemeine hierarchische Nachrichtenbeschreibung (Umwandlung der allg. Nachrichtenbeschreibungen aus Punkt 5 in hierarchische Strukturen zum Zweck der Implementierung).
7. Implementierbare Nachrichtenspezifikationen (ausgehend von den allgemeinen hierarchischen Nachrichtenbeschreibungen werden die konkreten Implementierungsvorschriften für eine bestimmte Transfersyntax festgelegt)

Fortführend werden u.a. Identifikationsvariablen (Kommunikations-Teilnehmer ID, Rollen der Teilnehmer, ID der Dokumente, Patienten-IDs), die MAGDA-LENA Sicherheitspolitik (Authentifizierung und Autorisierung, Organisationskontrolle), Richtlinien für Netzbetreiber und Netzteilnehmer und Maßnahmen für die Einhaltung der MAGDA-LENA Richtlinien durch alle Teilnehmer und Provider vorgestellt. In Bezug auf Datenschutz und Datensicherheit werden außerdem Empfehlungen bezüglich Verschlüsselungsprotokolle, Verschlüsselungsalgorithmen, elektronischer Signatur und Passwortsystemen ausgesprochen. Dabei wird k-Anonymität (Kapitel 3.2) ebenfalls vorgeschrieben.

2.6.3. HIPAA

Die *Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* ist die erste staatlich vorgeschriebene Richtlinie in den USA zum Schutz von persönlichen Informationen und Daten im Gesundheitsbereich [12]. Somit ist diese im Gegensatz zu MAGDA-LENA gesetzlich vorgeschrieben und teilnehmende Einrichtungen müssen sich verbindlich an diese Richtlinie halten.

Schlüsselpunkte dieser Richtlinie sind [1]:

- Bereitstellung eines Standards für den Schutz von privaten, individuell identifizierbaren Patientendaten. Sie gewährt dem Patienten das Recht jederzeit Einsicht in seine Patientenakte zu haben und zu erfahren wann und zu welchem Zweck seine Daten weitergegeben wurden.
- Die Privacy Rule beschreibt weiters Umstände unter denen teilnehmende Entitäten anderen teilnehmenden Entitäten Zugang zu und Gebrauch von Patientendaten gewähren. Die Richtlinie soll Forschungsarbeit aber nicht behindern oder erschweren.
- Die Erfüllung der Richtlinie muss von allen teilnehmenden Einrichtungen bis spätestens 14.April 2003 erfolgt gewesen sein.

Als teilnehmende Einrichtungen werden in [1] Leistungserbringer und Verrechnungsstellen im Gesundheitswesen sowie Krankenkassen, die Patientendaten elektronisch austauschen, erwähnt. Andere Organisationen innerhalb des Gesundheitswesens, die ebenfalls Daten sammeln, verwalten und austauschen können, müssen aber nicht von der Richtlinie betroffen sein. Die Richtlinie ist nicht explizit für die Anwendung in der Forschung gedacht – sie kann jedoch Forschungseinrichtungen insoweit betreffen als dass sie den Zugang zu Informationen für Forschungszwecke reguliert.

Der Schutz der Daten soll, im Gegensatz zu MAGDA-LENA jedoch nicht durch k-Anonymität realisiert werden, sondern durch folgende zwei Möglichkeiten gewährleistet sein [1]:

- Als erste Möglichkeit wird das Entfernen von allen 18 Attributen genannt, die eine Identifikation einer Person, derer Verwandter oder Mitarbeiter ermöglicht. Teil dieser 18 Attribute sind u.a. Name, Telefonnummer, Bilder vom Gesicht der Betroffenen usw. Eine vollständige Liste der Attribute ist in Abbildung 1 zu sehen
- Als zweite Möglichkeit wird erwähnt, dass es ausreichend zum Schutz der Daten sei wenn eine Person, die geeignetes statistisches Hintergrundwissen besitzt (Experte, Statistiker), der Meinung ist, dass ein sehr geringes Risiko besteht in Hinsicht auf Identifikation einer Person (mit Hilfe der vorhandenen Daten oder mit Hilfe anderer Daten). Erteilt diese Person die Erlaubnis, können Daten ausgetauscht werden. Es wird nicht näher erwähnt wie eine solche Person bestimmt werden kann, es wird auch keine Empfehlung abgegeben.

Dem Patient wird darüber hinaus aber die Möglichkeit eingeräumt, Einrichtungen zur Weitergabe und Veröffentlichung seiner Patientendaten zu autorisieren.

1. Names.
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP Code, and their equivalent geographical codes, except for the initial three digits of a ZIP Code if, according to the current publicly available data from the Bureau of the Census:
   a. The geographic unit formed by combining all ZIP Codes with the same three initial digits contains more than 20,000 people.
   b. The initial three digits of a ZIP Code for all such geographic units containing 20,000 or fewer people are changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
4. Telephone numbers.
5. Facsimile numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification.

Abbildung 1: Liste aller Attribute die lt. HIPAA bei Datenweitergabe entfernt werden müssen [1]

In Hinsicht einiger Attribute ist in [1] ebenfalls die Rede von Weglassen von Teilen der Information bei Geburtstagen (alle Daten bis auf das Jahr) oder Postleitzahlen (Weglassen der initialen 3 Ziffern ab einer bestimmten Anzahl von Personen in einem bestimmten geographischen Bereich – in Amerika sind die Postleitzahlen 5-stellig) wie es bei einer Generalisierung im Zuge der k-Anonymität (Kapitel 3.3) üblich ist

3. Ausgewählte Methoden zur k-Anonymität und Pseudonymisierung im medizinischen Bereich

Ein Datensatz kann mit Hilfe verschiedener Methoden anonymisiert oder pseudonymisiert werden. Eine Anonymisierung mit Hilfe der k-Anonymisierung birgt potenzielle Sicherheitslücken, die mit Hilfe von Hintergrundwissen (Kapitel 3.2.1.2) ausgenutzt werden können. Die Pseudonymisierung baut auf dem Prinzip des vollständigen Ersetzens der sensiblen Daten durch Pseudonyme auf [1]. In solch einem Fall ist im medizinischen Bereich eine Rückidentifizierung (Verständigung der korrekten Patienten) gewünscht, welche wiederum ein zusätzliches potentielles Sicherheitsrisiko (Referenzliste für die Rückidentifizierung [1]) birgt. Im SPICS Soul System (Kapitel 4) werden Daten ausgetauscht, die sensible Attribute enthalten können. Daher ist es speziell im Fall dieses Projektes, wünschenswert und im Interesse der Patienten, ihre Anonymität zu bewahren. In den nächsten Abschnitten liegt daher der Fokus auf der Schwierigkeit der Umsetzung dieses Interesses und auch auf den verschiedenen, bis dato vorhandenen Methoden und Systemen die es für die Pseudonymisierung und Anonymisierung gibt. Die verschiedenen vorgestellten Methoden unterscheiden sich in Bezug der Qualität der Anonymisierung bzw. Pseudonymisierung und auch in Bezug auf die Geschwindigkeit. Die Bewertung und Unterscheidung erfolgt aufgrund dieser Attribute.

3.1.Pseudonymisierung

Im Folgenden werden fünf verschiedene Modelle der Pseudonymisierung vorgestellt, welche sich wiederum in zwei Hauptgruppen á drei bzw. zwei Modelle trennen lassen. Die Unterscheidung ist hier im Bereich der Speicherdauer der Daten zu beobachten, die bei drei Modellen nur zwecks einmaliger Nutzung zu vorher bestimmten medizinischen Forschungszwecken und bei den übrigen zwei Modellen zwecks Langzeitspeicherung erfolgt [1].

3.1.1. Modelle für die einmalige Nutzung der Daten

Bei der Einmalnutzung werden Daten nur ein einziges Mal verwendet. Dies kann bei statistischen Auswertungen (Kapitel 3.1.1.2) oder Auswertungen von *Follow-Up* Daten (Kapitel 3.1.1.1) der Fall sein. *Follow-Up* Daten sind Daten, die aus einer vorangegangenen Datenverarbeitung entstanden sind oder gewonnen wurden. Ist eine Rückidentifizierung notwendig (Kapitel 3.1.1.2), muss die Einwilligung des Patienten eingeholt werden, da dies mit alleiniger Anonymisierung, welche keine Rückidentifizierung erlaubt, nicht möglich ist. [1]

Das Modell für eine einzelne Datenquelle bzw. der Einmalnutzung der Daten bildet das typische Szenario für einen Anonymisierungsdienst, wie er zum Beispiel bei einer einfachen statistischen Auswertung von Patientendaten verwendet wird. Aufgrund des Einsatzes der Anonymisierung ist hier die Einwilligung des Patienten nicht notwendig.[1]

3.1.1.1. Mehrere überlappende Datenquellen mit Einmalnutzung

Im Modell für mehrere überlappende Datenquellen mit Einmalnutzung (Abbildung 2) erfolgt die Datenzusammenführung aus verschiedenen Quellen, wie zum Beispiel bei der Auswertung von *Follow-Up-Daten* (Daten, die aus einer vorangegangenen Datenverarbeitung entstanden sind oder gewonnen wurden). Der gewünschte Verwendungszweck der Einmalnutzung ist mit Hilfe von Einweg-Pseudonymen zu erreichen, wobei hier die Voraussetzung dafür die Verfügbarkeit eines einheitlichen, eindeutigen *PID* (Patientenidentifikator) der verschiedenen Datenquellen ist. Die Pseudonymisierung erfolgt dann über die kryptografische Einwegverschlüsselung des *PID*, welche wiederum von einer unabhängigen Stelle durchgeführt wird.

Sollte es sich um einen größeren Forschungsverbund handeln, kann auch ein zentraler *TTP-Dienst* (Trusted Third Party) für diese Aufgabe eingerichtet und in die Netzarchitektur eingebunden werden. Der Dienst sollte aber im Sinne der Gewaltentrennung die medizinischen Daten gar nicht erst zu Gesicht bekommen. Erreicht wird das durch das Prinzip der asymmetrischen Verschlüsselung bei der die medizinischen Daten bei den Datenquellen mit dem öffentlichen Schlüssel des Empfängers verschlüsselt und in dieser Form durch den Pseudonymisierungsdienst durchgereicht werden. Dieses Modell wird in der folgenden Abbildung 2 veranschaulicht (*MDAT* = medizinische Daten, *IDAT* = Identitätsdaten, *PSN* = Pseudonym, *PID* = Patienten-Identifikator). Die Daten zur Sekundärnutzung beinhalten nur noch das Pseudonym welches durch einen Pseudonymisierungsdienst mittels kryptografischer Methoden erstellt wurde. Für SPICS Soul ist ein solcher Dienst ebenfalls denkbar, da auch in diesem Fallbeispiel eine Gewaltentrennung stattfinden soll.

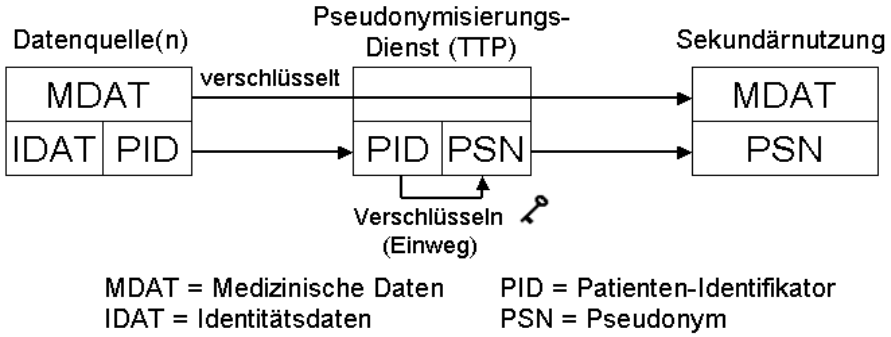

Abbildung 2: Pseudonymisierung für die einmalige Sekundärnutzung [1]

Die Speicherung der Identitätsdaten und der *PID* bei der Datenquelle ist unproblematisch, da der Pseudonymisierungsdienst die Zuordnung zwischen *PID* und *PSN* nicht speichert und die Verschlüsselung ebenfalls nicht umkehren kann.[1]

3.1.1.2. Einmalnutzung mit möglicher Rückidentifizierung

In einigen Fällen von Pseudonymisierung besteht der Bedarf der Rückidentifizierung von Patienten. Mögliche Gründe sind zum Beispiel die Rückmeldung von Diagnosen, die Rekrutierung von Fällen und die Einholung der Bewilligung für weitere Studien, z.B. Allergietests oder eine Unstimmigkeit welche aus nachträglich entdeckten, fehlerhaften Daten und Diagnosen entstanden ist und geklärt werden soll.

Im einfachsten Modell der Pseudonymisierung mit Möglichkeit zur Rückidentifizierung [1] führt ein zentraler Dienst eine Referenzliste für die Zuordnung von Identitätsdaten und Pseudonymen. Eine solche Liste, wie in nachfolgender Abbildung 3 zu sehen ist, ist jedoch ein potenzielles und sensibles Ziel in dem System. Sie führt Buch über den Zusammenhang zwischen Pseudonymen und Identitätsdaten. In SPICS Soul sind zentrale Konzepte aus sicherheitstechnischen- und kostengründen unerwünscht, weshalb dieses Modell keine Anwendung finden wird.

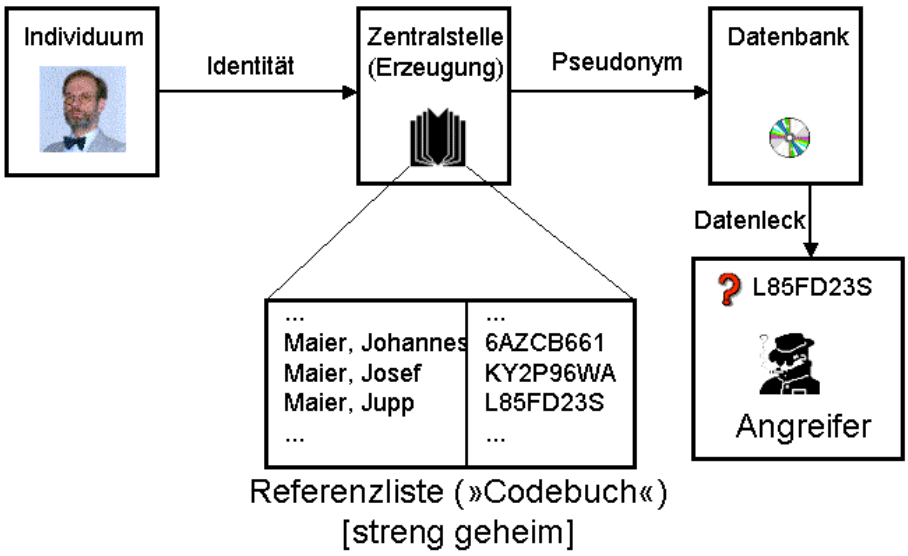

Abbildung 3: Einmalnutzung mit möglicher Rückidentifizierung [1]

Eine sicherere und somit bessere Möglichkeit bietet hingegen ein Dienst wie der im vorhergehenden Absatz (Kapitel 3.1.1.1) erwähnte, der zusätzlich mit einer Komponente zur Umkehr der Verschlüsselung ausgestattet ist. Dann sollte aber (im Gegensatz zu der Methode in Kapitel 3.1.1.1) kein allgemein verfügbarer *PID* verwendet werden, sondern ein speziell für das Projekt erzeugter Patientenidentifikator. Somit ist eine gewisse Diversität gegeben da in verschiedenen Projekten verschiedene PIDs verwendet werden, die extra für das Projekt erzeugt werden. Die Herstellung eines Zusammenhanges wird dadurch erschwert. Der *PID* wird in diesem Modell von einer weiteren unabhängigen Instanz, dem *PID-Dienst* zur Verfügung gestellt. Dieser speichert dann die Referenzliste welche auch bei einem unerlaubten Zugriff keine verwertbaren Informationen preis gibt – der Zusammenhang zwischen PID und Pseudonym bleibt nämlich verborgen. Die informationelle Gewaltentrennung ist hier durch ein zweistufiges Pseudonymisierungsverfahren (PID und PSN) realisiert worden, siehe folgende Abbildung 4:

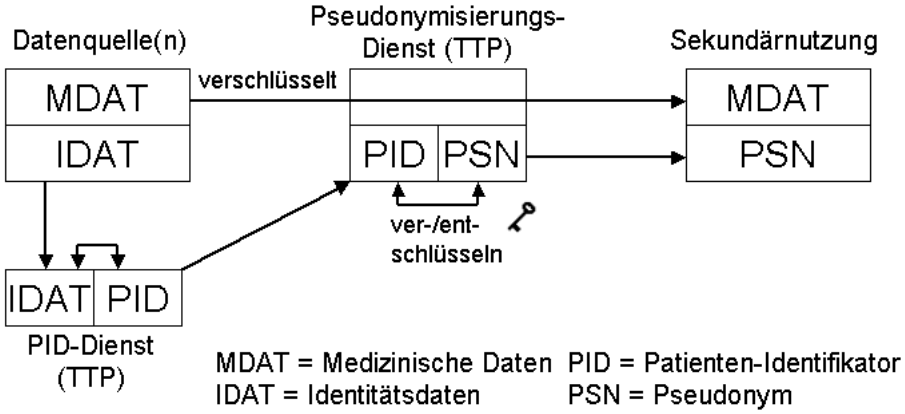

Abbildung 4: Einmalnutzung mit möglicher Rückidentifizierung [1]

3.1.2. Modelle für die Langzeitspeicherung der Daten

Ist nicht nur die einmalige Nutzung der Daten, sondern die längerfristige Speicherung der Daten erwünscht und notwendig, so wird der Schutz der Persönlichkeitsrechte der Patienten mit einem höheren Niveau von Anforderungen an einen solchen Dienst erreicht. Pommerening et. al stellen in ihrer Arbeit [1] zwei Dienste vor die diese Anforderungen erfüllen.

Eine Rechtfertigung für die Langzeitspeicherung stellt der Aufbau von Forschungsnetzen für vorher nicht notwendig bestimmte Fragestellungen dar.

Für den Aufbau solcher Forschungsinfrastrukturen ist die vereinfachte Sicht, bei der nur zwischen Behandlungs- und Forschungskontext unterschieden wird, nicht mehr ausreichend. Es sind komplexere Strukturen notwendig, bei denen bis zu 4 voneinander abgrenzbare Bereiche definiert werden: [1]

- Behandlungszusammenhang
- Die lokale Sammlung von Forschungsdaten
- Zentrale Datenpools für ein Forschungsnetz
- die Nutzung von Daten aus diesen Pools als Datenbasis für konkrete Auswertungen oder zur Rekrutierung von Fällen für neue Studien

3.1.2.1. Die pseudonymisierte Forschungsdatenbank

Im Modell der pseudonymisierten Forschungsdatenbank ist der Datenfluss derselbe wie der im vorhergehenden Kapitel 3.1.1.2 vorgestellte, jedoch werden die Daten auf der Nutzungsseite für

eventuell noch gar nicht definierte zukünftige Forschungszwecke gespeichert bzw. gesammelt. Der Konflikt entsteht hier zwischen den Grundrechten auf Schutz der Persönlichkeit und dem Recht der Forschungsfreiheit. Es entsteht ein Handlungsspielraum, der rechtliche Abwägungen und Kompromisse erfordert.
Das Werkzeug der Einwilligungserklärung stößt hier auf seine Grenzen, da es nur auf vorher definierte Datenverwendung und begrenzte Aufbewahrungsdauer angewandt werden kann. Eine Lösung wäre dann die Abstufung der Einwilligungserklärung, die dem Patienten Wahlmöglichkeiten anbietet. Ein solcher Weg wird durch das sogenannte Modell B des generischen TMF-Datenschutzkonzeptes [1] beschrieben.

3.1.2.2. Die zentrale klinische Datenbank

Das Modell der zentralen klinischen Datenbank unterstützt folgende Aspekte:

- Langzeitbeobachtung chronisch kranker Patienten
- das gemeinsame Datenmanagement verschiedener klinischer Studien
- die individuelle Rückmeldung von Forschungsergebnissen an den Patienten über den behandelnden Arzt ohne die Notwendigkeit, erst einen Depseudonymisierungsprozess anzustoßen

Im Zentrum dieses Modells steht eine Datenbank auf welche die teilnehmenden Ärzte Zugriff haben. Diese sind auch für die Qualität der Daten verantwortlich. Die Datenbank enthält im Gegensatz zu der pseudonymisierten Forschungsdatenbank (Kapitel 3.1.2.1) keine Identitätsdaten sondern nur einen *PID* als Referenz. Dieser *PID* ist nur in der Datenbank und der Patientenliste des *PID-Dienstes* sichtbar und somit selbst schon ein echtes Pseudonym. Eine Patientenliste löst im Falle eines Zugriffs die Referenz auf und der Zugriff wird über ein ad hoc generiertes *Token* (TempID – temporäre ID) möglich. Zusätzliche Quellen wie zum Beispiel Laborproben werden über zusätzliche Referenzen (andere IDs) erschlossen und bilden zusammen ein mehrstufiges Pseudonymisierungsverfahren. Wenn Daten aus der zentralen Datenbank für ein Forschungsprojekt benötigt werden, wird niemals ein Direktzugriff gewährt, sondern es wird ein geeigneter Auszug aus der Datenbank mit einem weiteren ad hoc erzeugten Pseudonym exportiert.

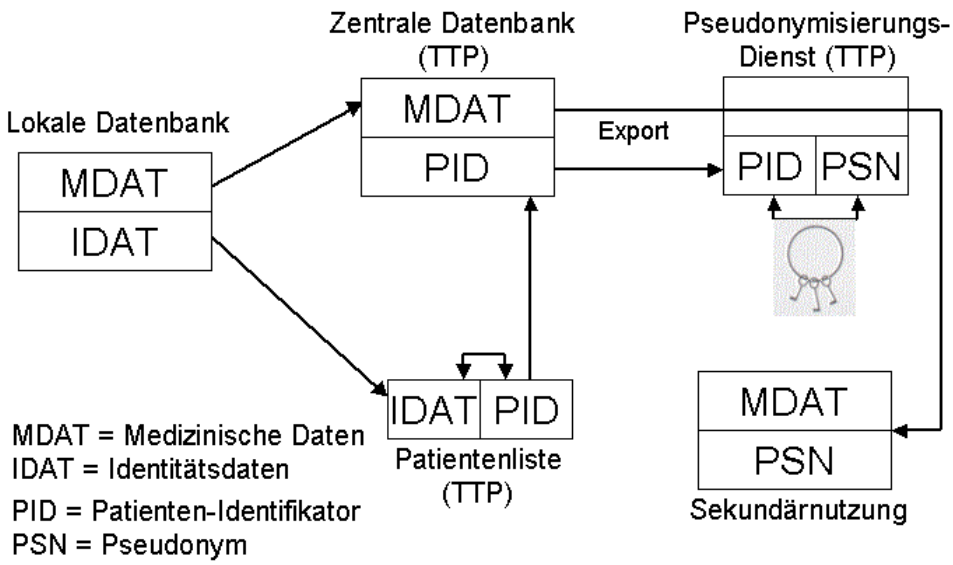

Abbildung 5: Die zentrale klinische Datenbank [1]

Dieses sogenannte Modell A des generischen TMF-Datenschutzkonzeptes erfordert die Implementierung von komplexen Kommunikationsbeziehungen, welche aber durch die Automatisierung vieler Vorgänge, zumindest vor den Teilnehmern, verborgen bleibt [1].

3.2. *k*-Anonymität

Im ersten Abschnitt (Kapitel 2.4 und 2.5) wurden die Begriffe Anonymität und Pseudonymität näher erläutert. Das Konzept der *k*-Anonymität erweitert das Prinzip der Anonymisierung. In einem *k*-anonymen Datensatz ist jedes einzelne Tupel (in einer Projektion über identifizierende Attribute betrachtet) mindestens *k*-mal vorhanden. Hierbei steht *k* üblicherweise für eine ganze Zahl zwischen 2 und 10 [3].

Eine Tabelle entspricht den Regeln bzw. Anforderungen der *k*-Anonymität wenn jedes Tupel t in dieser Tabelle T, in Bezug auf die *Quasi-Identifiers (QI)*, mindestens zu $k-1$ anderen Tupeln t_{i1}, $t_{i2},\ldots,t_{ik-1}$ dieser Tabelle ununterscheidbar ist [3]. C kann ein beliebiges Element aus der Menge der sekundären Identifikationsmerkmale QI sein. Formal geschrieben:

$$t[C] = t_{i1}[C] = t_{i2}[C] = \ldots = t_{ik-1}[C] \text{ für alle } C \in QI$$

Formel 2: k-Anonymität [3]

Eine solche Tabelle wird dann „*k*-anonyme Tabelle“ genannt und mit *T** bezeichnet. Also gibt es für jede Kombination der Werte der Quasi-Identifiers, mindestens *k* Tupel, die ebenfalls diese Werte aufweisen. Dies soll die Sicherheit garantieren, dass Individuen nicht eindeutig identifiziert werden können. Als Beispiel einer *k*-anonymen Tabelle ist untenstehende Tabelle 3, die fiktive Patientendaten eines Krankenhauses beinhaltet, zu sehen.

	Non-sensible Attribute			*sensible Attribute*
	PLZ	**Alter**	**Nationalität**	**Diagnose**
1	10**	< 30	*	manisch depressiv
2	10**	< 30	*	manisch depressiv
3	10**	< 30	*	tablettenabhängig
4	10**	< 30	*	tablettenabhängig
5	11**	≥ 40	*	Krebs
6	11**	≥ 40	*	manisch depressiv
7	11**	≥ 40	*	tablettenabhängig
8	11**	≥ 40	*	tablettenabhängig
9	10**	3*	*	Krebs
10	10**	3*	*	Krebs
11	10**	3*	*	Krebs
12	10**	3*	*	Krebs

Tabelle 3: 4-anonyme Tabelle *T (Tabelle 1) mit Patientendaten**

Zu sehen ist, dass die Tabelle keine primären Identifikationsmerkmale wie die Sozialversicherungsnummer oder den Namen der Patienten enthält. In dieser Tabelle sind die Attribute in 2 Gruppen unterteilt: nicht-sensible Attribute (Postleitzahl, Alter, Nationalität) und sensible Attribute (Diagnose). Es ist deutlich zu erkennen, dass in dieser 4-anonymen Tabelle jedes Tupel dieselben Werte für die Quasi-Identifiers (PLZ, Alter, Nationalität) wie mindestens 3 andere Tupel aufweist. Aufgrund des simplen Grundkonzepts hat sich die *k*-Anonymität zu einer gebräuchlichen und angewandten Variante der Anonymisierung im Kontext des Datenaustauschs entwickelt. Trotzdem garantiert sie keinen vollständigen Schutz von Individuen. Dies wird im nächsten Abschnitt durch 2 Beispiel-Attacken auf die *k*-Anonymität verdeutlicht. Um gegen solche Attacken standhaft zu bleiben muss ein weiteres Werkzeug, nämlich die *l*-Diversität (Kapitel 3.5), eingeführt werden.

3.2.1. Ausgewählte Attacken auf *k*-Anonymität

In diesem Abschnitt werden 2 verschiedene Attacken auf die *k*-Anonymität gezeigt. Es wird erläutert wie es möglich ist, unbefugt an Daten aus einem k-anonymen Datensatz zu gelangen [1]. Der Vollständigkeit halber wird an dieser Stelle ebenfalls erwähnt, dass die Veränderung der Reihenfolge für

die Anonymität essentiell ist. Können anonymisierte Daten in real-time abgefragt werden oder werden diese zu einem Zeitpunkt kurz nach Diagnosestellung veröffentlicht, kann man die, für einen Angreifer interessanten Datensätze, auf einen sehr kleinen Auszug reduzieren – vorausgesetzt die Daten sind nach Einlieferungsdaten oder Diagnosedatum sortiert. Dies muss bei den Attacken und den entsprechenden Gegenmaßnahmen, ebenfalls mit einbezogen werden

3.2.1.1. Homogeneity Attack

In [1] wird folgendes Beispiel angeführt: Angenommen Alice und Bob sind Nachbarn, die kein besonders gutes persönliches Verhältnis zueinander pflegen. Eines Tages wird Bob krank und muss mit der Rettung ins Krankenhaus gebracht werden. Das macht Alice neugierig und sie möchte herausfinden an welcher Krankheit Bob leidet.
Anhand der 4-anonymen Patienten-Tabelle des Krankenhauses welche öffentlich ausgehängt ist, kann Alice davon ausgehen, dass 1 Eintrag davon Bob zuzuteilen sein muss. Da sie ja seine Nachbarin ist, weiß sie, dass er ein 31-jähriger männlicher Österreicher ist und die Postleitzahl seines Wohnortes 1030 lautet. Mit diesen Informationen ausgerüstet kann Alice nun die Einträge der Tabelle, welche für Bob in Frage kommen würden, schon auf die Sätze 9, 10, 11 und 12 einschränken. Da alle Patienten dieser 4 Datensätze an Krebs leiden kann Alice davon ausgehen, dass Bob auch an Krebs leidet.
Laut [1] ist eine solche Situation in der Realität nicht ungewöhnlich. Angenommen es existiert ein Daten-Set mit 60.000 eindeutigen Tupeln, in welchen die sensiblen Attribute 3 verschiedene, eindeutige Werte annehmen können, welche nicht mit non-sensiblen Attributen korreliert werden.
Eine 5-anonymisierung dieser Tabelle enthält dann 12.000 Gruppen und eine von 81 Gruppen wird dann im Durchschnitt keine Diversität besitzen. Das bedeutet, dass alle Werte der sensiblen Attribute innerhalb dieser Gruppe gleich sind. Darauf aufbauend kann man sagen, dass 148 Gruppen ohne Diversität vorhanden sein können. Also wären in diesem Fall sensible Informationen von 740 Personen durch eine Homogeneity-Attack gefährdet [1].

Dieses Beispiel führt vor Augen, dass ein solches Daten-Set zusätzlich zur *k*-Anonymität auch Diversität in den sensiblen Attributen der Daten bieten sollte. Im nächsten Beispiel wird gezeigt, dass ein Widersacher auch mittels Hintergrundwissen zu sensiblen Informationen gelangen kann.

3.2.1.2. Background-Knowledge-Attack

In diesem Beispiel [1] hat Alice eine Brieffreundin namens Umeko, welche in dasselbe Krankenhaus wie Bob eingeliefert wird. Ihre Daten scheinen also ebenfalls in der Patienten-Tabelle auf. Alice weiß, dass Umeko eine 21-jährige weibliche Japanerin ist und momentan in einem Ort mit der Postleitzahl 1030 lebt, weil sie ihr die Briefe immer dorthin schickt.
Auf Basis dieser Informationen kann Alice nun darauf schließen, dass Umekos Daten entweder in Tupel 1, 2, 3 oder 4 enthalten sein müssen. Ohne zusätzliche Informationen weiß Alice nun nicht ob Umeko manisch depressiv oder tabellenabhängig ist. Laut [1] ist nun aber bekannt, dass Japaner sehr selten an Herz-Problemen leiden, was sie zu dem Schluss führt, dass Umeko mit hoher Wahrscheinlichkeit aufgrund einer Tablettenabhängigkeit in das Krankenhaus eingeliefert wurde.

Mit Hilfe dieser beiden Beispiele wird gezeigt, dass mittels einer Homogeneity- oder Background-Knowledge-Attack sensible Daten trotz *k*-Anonymisierung aus einer Tabelle herausgezogen werden können. Da beide dieser Attacken in der Realität relativ plausibel sind, benötigt man für den wirksamen Schutz von sensiblen Patientendaten eine Methode, die Hintergrundwissen und Diversität für die Definition von Privatsphäre mit einbezieht. Eine solche Definition wird mit Hilfe der *l*-Diversität (Kapitel 3.5) möglich.

3.3.Methoden der *k*-Anonymität

Es gibt bereits mehrere Anonymisierungsmethoden, die mittels *k*-Anonymisierung realisiert sind, welche im Folgenden vorgestellt werden. Diese können als Grundlage für eine Komponente zur Entscheidungsfindung für die Weitergabe von sensiblen Daten dienen. Sie alle verwenden die Methode der *k*-Anonymisierung als Basis und bauen auf dem Konzept der Generalisierung (auf Zellen oder Tupel-/Attributsebene) [33] [56], der Unterdrückung der Daten (komplettes Löschen von Zellen, Zeilen oder Spalten) [36] [1] und des Hinzufügens/Vertauschens von Werten (Rauschen) [57] [58] auf:

Generalisierung:
Bei einer Generalisierung werden Werte eines Attributes durch einen generalisierten Inhalt ersetzt. Dieser Inhalt ist nicht mehr so spezifisch wie der Vorangehende – zum Beispiel kann das Geburtsdatum „01.01.1980“ durch die Jahreszahl „1980“ ersetzt werden, falls es nur sehr wenige Einträge mit genau diesem Geburtsdatum gibt. Reicht diese Generalisierung nicht aus, ist das Zusammenfassen zu einer Gruppe auch möglich („1980-1982“). Weiters kann mit Platzhaltern („*“) gearbeitet werden, um eine Generalisierung des Geburtsdatums („198*“) zu erreichen

Unterdrückung:
Hierbei werden Werte einzelner Zeilen bzw. Spalten mit Platzhaltern versehen. Ein solcher Platzhalter kann ein Asterisk („*“) sein, mit Hilfe dessen man den Wert zur Gänze unterdrücken kann (Ersetzen der Postleitzahl „PLZ“ durch „*“).

Hinzufügen und Vertauschen von Werten (Rauschen):
Eine dritte Möglichkeit zum Abändern der Daten ist das Vertauschen von Werten untereinander (auf Zellen-, Zeilen- und Spaltenebene). Auch das Hinzufügen von beliebigen Daten führt zu einem erwünschten Rauschen. Da jedoch die Daten in diesem Fall sehr stark abgeändert werden können, wird im Allgemeinen auf eine Kombination von Generalisierung und Unterdrückung zurückgegriffen, um aussagekräftige Ergebnisse mit verwendbaren Daten zu erhalten.

Domain Generalization Hierarchies werden im Folgenden des Öfteren verwendet. In [30] verwendet Sweeney den Begriff Domäne im Kontext mit relationalen Datenbanken. Dieser Begriff beschreibt auch die möglichen Werte eines Attributes welches alle verschiedenen Werte, die auch in der Grunddomäne (Ausgangtabelle) vorkommen, annehmen kann. Um weniger spezifische Werte zu bekommen kann man auch eine Projektion der Domäne P0 auf die Domäne P1 erfolgen (Beschrieben durch den Operator $<_D$) was einer hierarchisch angeordneten Verallgemeinerung der Werte entspricht (PLZ ‚1100’ wird zu ‚110*’ bzw. zu ‚11**’ oder ‚1***’). Durch ein solches Vorgehen lassen sich unterschiedliche sogenannte Domain Generalization Hierarchies (DGH) erstellen. Die DGH ist definiert als eine Menge von Domänen, vollständig geordnet durch die Beziehung $<_D$ [3].

3.3.1. MinGen Algorithmus

k-Anonymität kann unter anderen durch Unterdrückung und Generalisierung von Daten erreicht werden. In der Publikation von Latanya Sweeney [30] wird der MinGen - Algorithmus vorgestellt, der mit Hilfe von minimalen Veränderungen eine gegebene Tabelle in *k*-anonyme Form bringt. Dabei wird hinsichtlich einer gegebenen Metrik die optimale Lösung gefunden.

Algorithmus: [30]
Gegeben seien eine Tabelle $\boldsymbol{PT}$ $(A_x,\ldots,A_y)$, die *Quasi-Identifiers* $\boldsymbol{QI}=\{A1,\ldots,An\} \subseteq \{Ax,\ldots,Ay\}$, „Domain Generalization Hierarchies“ $\boldsymbol{DGH_{Ai}}$ und eine natürliche Zahl $\boldsymbol{k} < |\boldsymbol{PT}|$ (welches Voraussetzung

für die Existenz einer k-minimalen Generalisierung ist und den Grad der Anonymitität angibt). Aus diesen gegebenen Entitäten wird die *k*-minimale Tabelle *MGT* produziert.

Schritte des Algorithmus [30]*:*

1. Ist PT bereits k-anonym? Wenn nicht, weiter bei Schritt 2.1.

2.1. Speichere alle möglichen Generalisierungen von *PT* über *QI* in *allgens*.

2.2. Speichere alle Generalisierungen von *allgens*, die den k-Anonymitätsanforderungen genügen in *protected*.

2.3. Die Generalisierung(en) mit der geringsten Veränderung bzw. besten Lösung (basierend auf der gegebenen Metrik), wird in *MGT* gespeichert.

2.4. Die Funktion *preferred* wählt die Lösung aus.

3. Ausgabe der Tabelle *MGT*

In der folgenden Abbildung 6 ist eine abstrahierte Form der Vorgehensweise des Algorithmus dargestellt:

Input: Private Table **PT**; quasi-identifier QI = $(A_1, \ldots, A_n)$,
k constraint; domain generalization hierarchies
DGH_{Ai}, where i=1,…,n, and *preferred()* specifications.
Output: MGT, a minimal distortion of PT[QI] with respect to k
chosen according to the preference specifications
Assumes: $|PT| \geq k$
Method:

1. **if PT**[QI] satisfies k-anonymity requirement with respect to k **then do**
   1.1. MGT ← { **PT** } // **PT** is the solution
2. **else do**
   2.1. *allgen* ← {T_i : T_i is a generalization of **PT** over QI}
   2.2. *protected* ← {T_i : $T_i \in$ *allgen* ∧ T_i satisfies k-anonymity of k}
   2.3. MGT ← {T_i : $T_i \in$ *protected* ∧ there does not exist $T_z \in$ *protected* such that *Prec*(T_z) > *Prec*(T_i) }
   2.4. MGT ← **preferred**(MGT) // select the preferred solution
3. **return** MGT

Abbildung 6: MinGen Algorithmus [30]

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT *Prec*=1.00

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
person	1965	female	0213*	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT *Prec*=0.90

Tabelle 4: k-minimale Abänderung der Tabelle PT (links) für k=2 [30]

Der Algorithmus ist nicht nur sehr komplex, sondern auch ausgesprochen ineffizient. Fordert man Generalisierung auf Attribut-Level lässt sich die Anzahl der möglichen Generalisierungen, *|allgens|*, mit folgender Formel berechnen:

$$\prod_{i=1}^{n} (|DGH_i| + 1)$$

Formel 3: Formel zur Berechnung der DGH auf Attributs-Ebene [30]

In oben beschriebenem Min-Gen-Algorithmus erfolgt die Generalisierung jedoch auf Zellenebene was zu einem noch ineffizienteren Laufzeitverhalten führt, welches sich für die Anzahl aller möglichen Generalisierungen mit folgender Formel berechnen lässt:

$$\prod_{i=1}^{n} (|DGH_i| + 1)^{|PT|}$$

Formel 4: Formel zur Berechnung der DGH auf Zellen-Ebene [30]

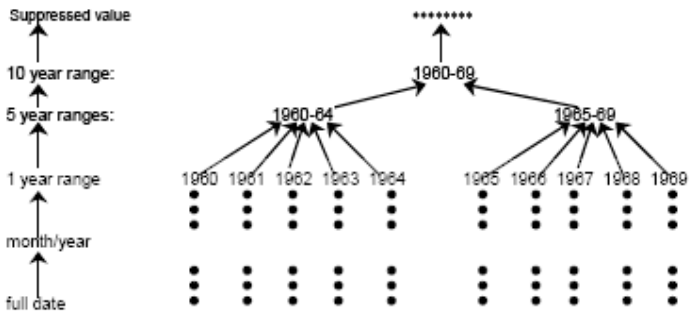

Abbildung 7: Generalisierung des Geburtsdatums [30]

Abbildung 7 zeigt einen Baum, der die Generalisierung des Geburtsdatums darstellt. Angefangen vom unveränderten Datum wird aufgrund der Kriterien der k-Anonymität nach und nach eine Generalisierung durchgeführt indem zuerst der Tag, sollte dies nicht ausreichen, dann der Monat usw. weggelassen werden. In weiteren Schritten werden Jahre zu Ranges zusammengefasst (z.B. 1960-64) – reicht dies immer noch nicht aus, wird das gesamte Attribut weggelassen, also unterdrückt.

Im nächsten Kapitel werden zwei Systeme vorgestellt, die in der Praxis angewendet werden. Beide bauen auf dem Ansatz der *k*-Anonymität durch Generalisierung mittels Unterdrückung auf. Bei diesen beiden Systemen handelt es sich um das (1) Datafly System [31] und das (2) Statistics Netherlands' μ-Argus [32] System. Sweeney verdeutlich in der Arbeit [30], dass das Datafly-System die Daten zu stark generalisiert und μ-Argus sogar versagen kann wenn es um den Schutz der Privatsphäre der Patienten geht. Beide Methoden stammen aus den späten 90er-Jahren und wurden/werden mangels fehlender Methoden und Weiterentwicklungen teilweise immer noch in der Praxis vewendet.

3.3.2. Datafly System

Das Datafly System stellt nicht immer k-minimale Generalisierungen oder *k*-minimale Verformung von Daten zur Verfügung aber die Ergebnisse sind immer den Anforderungen der *k*-Anonymität genügend. Eine Tabelle T_m (Generalisierung einer Tabelle T_I) ist *k*-minimal wenn sie den Anforderungen der *k*-Anonymität genügt und sie kein Ergebnis weiterer Generalisierungen einer k-anonymen Generalisierung der Tabelle T_I ist [31].
Ein Problem des Systems ist, dass der Algorithmus undurchdachte Entscheidungen fällt, wie zum Beispiel die Generalisierung aller Werte, die mit einem Attribut in Verbindung gebracht werden oder die Unterdrückung aller Werte innerhalb eines Tupels. Das bedeutet, dass Generalisierungen auf Tupel- (Zeilen) oder Attributs-Ebene (Spalten) erfolgen und somit Daten zu stark generalisiert werden. Wie in Tabelle 3 verdeutlicht, hätte eine Generalisierung des Attributs „PLZ" mittels Weglassen der letzten beiden Stellen (10**) genügt:

M/W	PLZ
M	1010
M	1020
W	1010
M	1010
W	1040
W	1090
W	1070
M	1080

M/W	PLZ
M	1***
M	1***
W	1***
M	1***
W	1***
W	1***
W	1***
M	1***

Tabelle 5: zu starke Generalisierung der Tabelle

Ein weiteres Problem ist die Heuristik, die höher distinkte Attribute (große Diversität) zur Generalisierung heranzieht. Diese ist sicherlich in Hinblick auf die Laufzeit des Algorithmus effizienter, generalisiert aber unnötigerweise Daten, die ohnehin schon eine hohe Diversität aufweisen [8]. Der Algorithmus liefert also nicht immer die optimalen Resultate – wie oben erwähnt, generalisiert er manchmal zu stark – doch aufgrund der kurzen Berechnungszeit ist er für den Einsatz in der Praxis gut geeignet [31].

Vorgehensweise von Datafly [31]: Der Dateninhaber kennzeichnet spezifische Attribute und Tupel der originalen Private Table (***PT***) die veröffentlicht werden sollen. Außerdem gruppiert er ein Subset von Attributen von ***PT*** in einen oder mehrere *Quasi-Identifier* ($\boldsymbol{QI_i}$) und ordnet den Attributen innerhalb jedes $\boldsymbol{QI_i}$ eine Gewichtung von 0 bis 1 zu, welche die Wahrscheinlichkeit angibt mit der das Attribut in Kombination mit anderen Daten zur Identifikation führen könnte. 0 bedeutet in diesem Fall „sehr unwahrscheinlich“, 1 bedeutet „sehr wahrscheinlich“. Weiters weist der Datenhalter der Tabelle einen minimalen Anonymitätslevel zu, der erreicht werden soll. Dieser Level entspricht dem Wert von *k*. Soll das Ergebnis mindestens 2-anonym sein muss infolgedessen auch k mindestens den Wert 2 annehmen. Zuletzt erfolgt noch eine Gewichtung der Attribute (zwischen 0 und 1) innerhalb der *Quasi-Identifier*, die angibt welche Attribute zuerst (entspricht der Zahl 1) und welche Attribute überhaupt nicht (entspricht der Zahl 0) verändert oder gelöscht werden sollen.
Mit einer solchen Gewichtung ist es möglich *k*-anonyme medizinische Daten zu erzeugen, die den Verwendungszweck berücksichtigen. Sollen beispielsweise Daten für medizinische Forschungsarbeit ausgetauscht werden, die den Zusammenhang zwischen Alter, Geschlecht, Nationalität und Krebsrisiko untersuchen, so ist es möglich durch Zuweisung entsprechender Werte zwischen 0 und 1 diese Attribute vor zu starker Generalisierung zu schützen. Trotzdem kann aber die geforderte *k*-Anonymität durch entsprechend stärkere Abänderung der restlichen Attribute, welche für die Forschungsarbeit nicht unbedingt entscheidend sind, erreicht werden.

Die Steuerung des Anonymisierungsgrads jedes einzelnen Attributes hat hier auch einen angenehmen „Nebeneffekt“: Da sich manche Attribute besonders gut eignen, um in Verbindung mit anderen veröffentlichten Daten Individuen zu identifizieren, kann für solche Attribute eine stärkere Generalisierung durch die Zuweisung einer Zahl nahe bei 1 gefordert werden.

Algorithmus [31]*:*

In dem folgenden gezeigten Beispiel wurden den Attributen aus Gründen der Übersichtlichkeit und des Verständnisses keine Gewichte zugewiesen.

Gegeben seien eine Tabelle ***PT*** *(A_x,…,A_y)*, die *Quasi-Identifiers* ***QI****={A1,…,An}* ⊆ *{Ax,…,Ay}*, „domain generalization hierarchies“ $\boldsymbol{DGH_{Ai}}$ und eine natürliche Zahl $\boldsymbol{k} < |\boldsymbol{PT}|$ (welches Voraussetzung für die Existenz einer *k*-minimalen Generalisierung ist). Aus diesen gegebenen Entitäten wird die *k*-anonyme Tabelle *MGT* produziert.

Schritte des Algorithmus

1. Im ersten Schritt erfolgt die Zusammenfassung der gegebenen Tabelle, in dem die Tupel, welche die gleichen sekundären Identifikationsmerkmale aufweisen, zu einer einzelnen Zeile zusammengefasst werden. Der so entstandenen neuen Tabelle wird eine zusätzliche Spalte mir der Bezeichnung *freq* (Frequenzliste) hinzugefügt, die pro zusammengefasster Zeile die Anzahl der enthaltenen, gleichen Tupel speichert.
2. Sind mehr als *k* Werte der Liste *freq* <= *k*, wird für jedes Attribut die Anzahl der distinkten (diversen) Werte berechnet, die es in der Tabelle einnimmt. Das Attribut mit den meisten verschiedenen Werten wird entsprechend der *domain generalization hierarchies* um eine Stufe generalisiert. Daraufhin wird die Liste *freq* aktualisiert, und der 2. Schritt solange wiederholt, bis weniger als *k* Werte dieser Liste <= *k* sind. Diese Abbruchbedingung soll verhindern, dass aufgrund von wenigen (<= *k*) Ausreißern die Werte der Tabelle zu stark generalisiert werden. Eine zu starke Generalisierung verhindert bei Bedarf der Verwendung der Daten für medizinische Forschungszwecke die korrekte und aussagekräftige Wiederherstellung der Daten und könnte so Ergebnisse verfälschen und die Forschungsarbeit behindern.
3. Löschen der Tupel, die in der Frequenzliste einen Wert <= *k* haben.
4. Herstellung der k-anonymen Form ***MGT*** der Ursprungstabelle ***PT*** mittels der Werte aus *freq* und den zuvor abgeschnittenen medizinischen Daten.

In Abbildung 8 ist eine abstrahierte Form der Vorgehensweise des Algorithmus dargestellt:

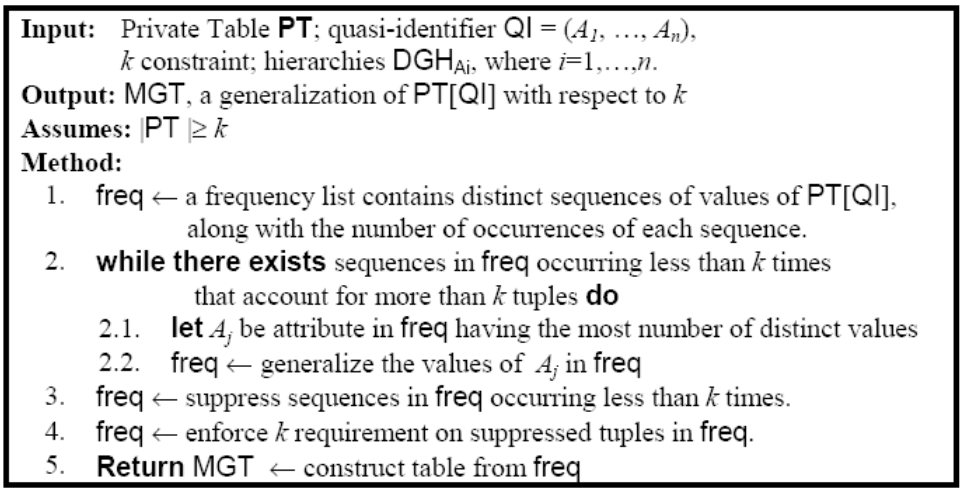

Abbildung 8: Core-Datafly Algorithmus [31]

Beispiel: unten stehende Tabelle 6 links zeigt bereits den Zustand nach dem 1.Schritt des Datafly-Systems. Sie beinhaltet nur sekundäre Identifikationsmerkmale (Nationalität, Geb.Datum, Geschlecht und PLZ) und die zusätzlich, neu hinzugefügte Spalte *occurs*. Da jedes Tupel aufgrund der Ausprägung der sekundären Identifikationsmerkmale eindeutig ist, konnte noch keine Zusammenfassung von Tupeln erfolgen. Spalte *occurs* zeigt, dass die Häufigkeit eines Tupels gleich 1 ist. Da der gewünschte Wert von k=2 noch nicht erreicht ist, wird der Algorithmus weiter ausgeführt. Da mehr als k Tupel eine Häufigkeit von < k aufweisen, wird die Spalte mit den meisten verschiedenen Ausprägungen (Geb.Datum hat 12 verschiedene Werte) generalisiert. Durch diesen Schritt ergeben sich nun mehrere gleiche Tupel, die jeweils zu gleichen Zeilen zusammengezogen werden können und Tabelle 6 (unten) bilden.

Wie zu sehen ist, sind immer noch 2 Zeilen vorhanden (Tupel t7 und t8) bei denen der Wert k=2 noch nicht erreicht ist. Diese beiden Ausreißer (=< k Tupel mit *occurs* < k) werden durch den Algorithmus eliminiert, um die Tabelle nicht zu stark zu generalisieren. Nach dem Löschen der beiden Tupel, dem Hinzufügen des vorher unterdrückten medizinischen Attributes (Diagnose) und dem Auseinanderziehen der vorher zusammengefassten Zeilen, ergibt sich die *2*-anonyme Tabelle *MGT* (Tabelle 7).

Tupel	*Nationalität*	*Geb.Datum*	*Geschlecht*	*PLZ*	*occurs*
t1	Österreich	20.09.1965	männlich	1030	1
t2	Österreich	14.02.1965	männlich	1030	1
t3	Österreich	23.10.1965	weiblich	1120	1
t4	Österreich	24.08.1965	weiblich	1120	1
t5	Österreich	07.11.1964	weiblich	1120	1
t6	Österreich	01.12.1964	weiblich	1120	1
t7	Italien	23.10.1964	männlich	1120	1
t8	Italien	15.03.1965	weiblich	1010	1
t9	Italien	13.08.1965	männlich	1010	1
t10	Italien	05.05.1964	männlich	1010	1
t11	Italien	13.02.1967	männlich	1080	1
t12	Italien	21.03.1967	männlich	1080	1

Tupel	*Nationalität*	*Geb.Datum*	*Geschlecht*	*PLZ*	*occurs*
t1,t2	Italien	1965	männlich	1030	2
t3,t4	Italien	1965	weiblich	1120	2
t5,t6	Italien	1964	weiblich	1120	2
t7	Österreich	1964	männlich	1120	1
t8	Österreich	1965	weiblich	1010	1
t9, t10	Österreich	1964	männlich	1010	2
t11,t12	Österreich	1967	männlich	1080	2

Tabelle 6: Einzelne Schritte des Core-Datafly Algorithmus

Nationalität	*Geb.Datum*	*Geschlecht*	*PLZ*	*Diagnose*
Italien	1965	männlich	1030	Ödipus
Italien	1965	männlich	1030	Alkoholismus
Italien	1965	weibich	1120	Drogenabhängig
Italien	1965	weibich	1120	Platzangst
Italien	1964	weibich	1120	Tablettenabhängig
Italien	1964	weibich	1120	Alkoholismus
Österreich	1964	männlich	1010	Tablettenabhängig
Österreich	1964	männlich	1010	Manisch Depressiv
Österreich	1967	männlich	1120	Burnout Syndrom
Österreich	1967	männlich	1120	Magersucht

Tabelle 7: Tabelle MGT, resultierend aus Datafly, k=2, QI={Race, Birthdate, Gender, ZIP}

3.3.3. μ-ARGUS

μ-ARGUS [32] wurde von Statistics Netherlands im Zuge der Suche nach sicheren Methoden für „Statistic Disclosure Control (SDC) for Microdata“ entwickelt [30]. Ziel ist auch hier, dass keine Individuen aufgrund von Kombination von Daten identifiziert werden können. Der Ausdruck Microda-

ta wird in der Statistik-Literatur [64] verwendet und beschreibt Daten in ihrer rohen, nicht-aggregierten Form.
Momentan gibt es 2 Methoden für SDC bei den Statistics Netherlands welche auch in μ-ARGUS Anwendung finden. Diese beiden Methoden sind [32]:

- *Global Recoding* (Generalisierung auf Attributebene)

  Beim Global Recoding werden mehrere Kategorien einer Variable bzw. Attribute zu einer Kategorie zusammengefasst. Zum Beispiel können die Berufe Statistiker und Mathematiker zu einer Berufs-Kategorie „Statistiker oder Mathematiker" zusammengefasst werden, wenn es nicht ausreichend Einwohner je Berufsgruppe in einem Ort gibt. Dieses Recoding hat dann eine sichere Kombination zusammen mit anderen Attributen als Ergebnis, welche veröffentlich werden kann, Natürlich können auch andere Attribute wie der Wohnort oder der Geburtsjahrgang zusammengefasst werden. Global Recoding wird also auf das globale Datenset und nicht nur auf den unsicheren Part, der Schutz benötigt angewandt. Im Unterschied zur Generalisierung bei Datafly bleiben hier Informationen wie „Mathematiker" oder „Statistiker" vollkommen erhalten, sie werden lediglich zu einer Gruppe zusammengefasst. Somit ist zumindest ersichtlich welche Ausgangsdaten zum Einsatz kamen. Für medizinische Forschungszwecke bedeutet dass, dass man genauere Aussagen hinsichtlich Zusammenhängen treffen kann (z.B. Krankheit hängt mit einer Berufsgruppe zusammen). Im Falle einer Postleitzahl könnte man erkennen dass es sich um 1030 oder 1040 handelt wenn man diese Nummernkreise zusammenfasst. Hingegen ist dies bei der Generalisierung 10** nicht mehr eindeutig.

- *Local Suppression* (Unterdrückung auf Zellenebene)

  Local Suppression beruht auf dem Konzept der Minimum unsafe Combinations (*MINUC*), welches eine zentrale Rolle für die Auswahl der Attribute und Kategorien für die Unterdrückung der Werte spielt. *MINUC*s sind Kombinationen von Daten, die ohne Unterdrückung der enthaltenen Attribute zu einer Identifikation führen können. So können einzelne Attribute wie zum Beispiel der Beruf unterdrückt und nicht veröffentlicht werden. Alle in einer *MINUC* enthaltenen Werte führen (egal in welcher Kombination) bei Unterdrückung einer dieser Werte auf jeden Fall zu einer sicheren Kombination die veröffentlicht werden kann. Das Ergebnis sind *k*-anonyme Daten. Dieses Konzept findet ebenfalls im Datafly-System Verwendung.

Beide SDC-Methoden führen schlussendlich zu Informationsverlust weshalb eine ausbalancierte Kombination der beiden Methoden gefunden werden muss, die diesen Verlust so gering wie möglich hält. μ-ARGUS bietet folgenden Ansatz [32]:

Es wird mit *Global Recoding* der Daten begonnen und solange weitergemacht bis die Anzahl der unsicheren Kombinationen, die durch *Local Suppression* geschützt werden müssen, ausreichend niedrig ist. Die übriggebliebenen unsicheren Kombinationen werden dann durch *Local Suppression* geschützt.
μ-ARGUS hilft bei dieser Vorgehensweise dem Anwender auf einer interaktiven Art und Weise bei der Spezifikation der *Global Recodings*. So kann der Anwender beispielsweise das Ergebnis rückgängig machen wenn er damit nicht zufrieden ist. Bei *Local Suppression* wählt der Algorithmus die zu unterdrückenden Werte optimal und automatisch aus, d.h. die zu unterdrückenden Werte werden auf ein Minimum reduziert, sodass *k*-anonyme Daten das Ergebnis sind. Der wesentliche Unterschied zum Datafly-System besteht in der Möglichkeit des Eingreifens des Anwenders und der Möglichkeit das Ergebnis rückgängig zu machen. Durch diesen Eingriff ist es möglich eine zu starke Generalisierung, wie bei Datafly zu verhindern. Mehr zu diesem Thema kann in [62] nachgelesen werden.

μ-ARGUS ist ein Software-Paket welches in Borland C++ geschrieben wurde. Abbildung 9 veranschaulicht die oben beschriebene funktionale Vorgehensweise des μ-ARGUS Algorithmus. Hierbei werden Micro- und Metadaten in Tabellenform auf unsichere Kombinationen, welche eine Identifizierung von Personen mit sich bringen könnten, untersucht. Wurden solche Kombinationen gefunden, folgen die Schritte *Global Recording* und *Local Supression*, um am Ende des Prozesses sichere Daten hinsichtlich Identifikation von Individuen zu erhalten.
Algorithmus:
In [30] liefert Sweeney nach Reverse Engineering und Neuimplementierung von μ-ARGUS eine genaue Beschreibung der Algorithmus.
Gegeben seien eine Tabelle $\boldsymbol{PT}$ $(A_x,\ldots,A_y)$, die sekundären Identifikationsmerkmale $\boldsymbol{Q_{PT}}=\{A1,\ldots,An\} \subseteq \{Ax,\ldots,Ay\}$, „domain generalization hierarchies" $\boldsymbol{DGH_{Ai}}$ und eine natürliche Zahl $\boldsymbol{k} < |\boldsymbol{PT}|$. Weiters werden die sekundären Identifikationsmerkmale auf drei Gruppen aufgeteilt – Attribute mit den Kennzeichnungen:

- most identifying (*Most*)
- more identifying und (*More*)
- identifying (*Identifying*)

Diese Einteilung hat, ähnlich wie beim bereits vorgestellten Datafly-System, Einfluss auf die Reihenfolge der Generalisierung.

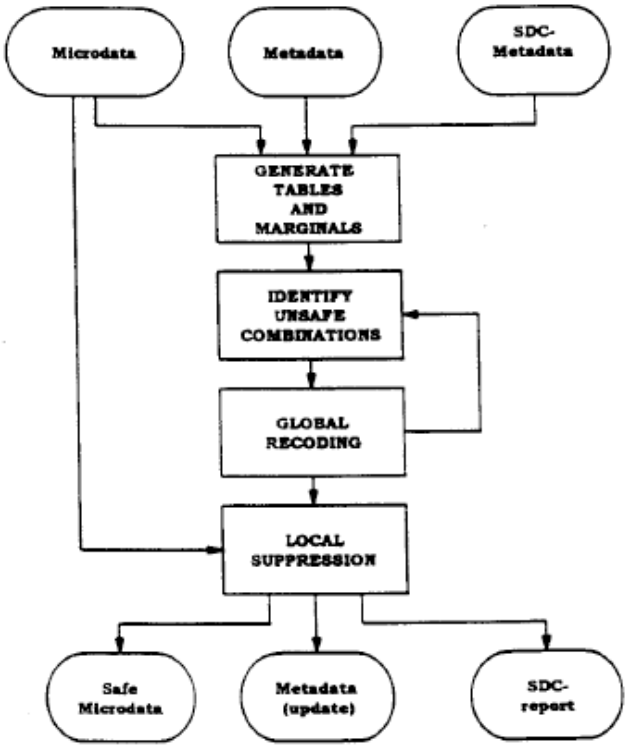

Abbildung 9: Funktionale Vorgehensweise des μ-ARGUS Algorithmus [32]

Schritte des Algorithmus:

1. Im ersten Schritt erfolgt das Anlegen einer Frequenzliste „*freq*“, die pro Attribut die voneinander unterschiedlichen Werte und die Häufigkeit (Anzahl des Auftretens) innerhalb von PT speichert.
2. Generalisierung der Attribute entsprechend der *DGH*, bis jeder Wert mindestens *k*-mal vorkommt.
3. Testen von verschiedenen 2er- und 3er Kombinationen der Attribute auf *k*-Anonymität. Erfüllen die Kombinationen nicht die Voraussetzungen für *k*-Anonymität, werden sie in die Datei *outliers* gespeichert.
4. Der Anwender bestimmt, ob ein Attribut, welches in *outliers* vorkommt, generalisiert werden soll.
5. Wiederholung der Schritte 3 und 4 bis der Anwender keine Attribute mehr zur Generalisierung freigeben möchte.

6. Der Zelleneintrag, der am Öftesten in *outliers* vorkommt, wird unterdrückt. Dieser Schritt wiederholt sich solange bis sich keine Kombinationen mehr in *outliers* befinden.

Laut [30] weist der Algorithmus in Punkt 3 einen Schwachpunkt auf, da hier nicht immer alle 2er- und 3er-Kombinationen getestet werden wodurch es möglich ist, dass Kombinationen, die nicht den Anforderungen der *k*-Anonymität genügen, ebenfalls in „outliers“ gespeichert werden. Diese können auf Wunsch des Anwenders gelöscht werden. Diese Option bringt allerdings die Notwendigkeit mit sich ein komplexes Optimierungsproblem zu lösen welches in [63] beschrieben wird. In Abbildung 10 ist zu sehen welche Kombinationen vom Algorithmus getestet werden:

Combinations Always Tested
Identifying × *More* × *Most*, *Identifying* × *Most* × *Most*, *Most* × *Most* × *Most*, *Identifying* × *More*, *Identifying* × *Most*, *More* × *Most*, *Most* × *Most*
Combinations Tested only if |*Identifying*| > 1
More × *More* × *Most*, *Most* × *Most* × *More*, *More* × *More*

Abbildung 10: Durch den µ-ARGUS Algorithmus getestete Kombinationen

Es können aber auch Kombinationen von 4 oder mehr Attributen existieren, die einmalig sind und ebenfalls nicht vom Algorithmus beachtet werden. Expandiert man den Algorithmus um diese Funktionalität, damit auch solche Kombinationen getestet werden, verliert er schnell an Effizienz. So kann es hier zu Ergebnissen kommen, die der *k*-Anonymität nicht genügen weshalb der Algorithmus noch Verbesserungsbedarf in dieser Hinsicht benötigt und in seiner momentanen Form nicht verwendet werden kann, falls absolute *k*-Anonymität garantiert werden soll.

3.3.4. INCOGNITO

ICOGNITO [33] basiert wie Datafly auf globaler Generalisierung auf Attributebene. Für das Problem der *k*-Anonymisierung muss immer die Anzahl der Attribute beachtet werden, weshalb bei INCOGNITO die Domain Generalization Hierarchies als Dimension herangezogen werden. Somit werden die Tabelle *T* und die *DGH*, die mit den sekundären Attributen der Tabelle *T* assoziiert werden als relationales „Star-Schema“ dargestellt. In Abbildung 11 lässt sich das Star-Schema für die Quasi-Identifier <Birthday, Sex, Zipcode> folgendermaßen darstellen:

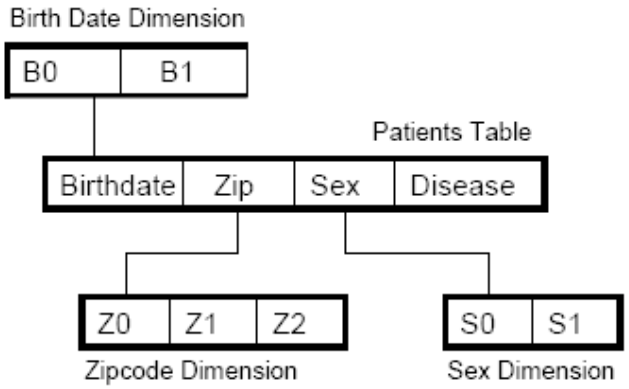

Abbildung 11: Star-Schema inkl. Generalisierungs-Dimensionen für die Quasi Identifier <Birthday, Sex, Zipcode> [12]

Dieses Schema zeigt die Tabelle *T* („*Patients Table*“) und die *DGH*, der sekundären Attribute. In Abbildung 12 ist gut zu sehen, wie sich die verschiedenen Dimensionen der Attribute für das Star-Schema in Abbildung 11 ergeben. Das Attribut *Zipcode* kann beispielsweise in 3 Dimensionen generalisiert werden, bis es sich nicht mehr von den anderen Werten unterscheiden lässt (*k*-anonym) – im Fall von *Birthday* und *Sex* existiert nur eine Dimension der Generalisierung. Um die Tabelle *T* in *k*-anonyme Form zu bringen, werden mit ICOGNITO alle möglichen Generalisierungen der Attribute erzeugt. Dies geschieht indem für jedes Attribut geprüft wird ab welcher Generalisierungsstufe die Tabelle – in Bezug auf das gerade betrachtete Attribut – *k*-anonym wäre. Das kann bei *Birthday* zum Beispiel durch Weglassen der ersten oder letzten beiden Zahlen im Geburtsjahr erfolgen. Anschließend werden immer größere Gruppen von Attributen zusammengefasst, bis am Ende die gesamte Menge der sekundären Identifikationsmerkmale erschöpft ist.

Birthdate	Sex	Zipcode	Disease
1/21/76	Male	53715	Flu
4/13/86	Female	53715	Hepatitis
2/28/76	Male	53703	Brochitis
1/21/76	Male	53703	Broken Arm
4/13/86	Female	53706	Sprained Ankle
2/28/76	Female	53706	Hang Nail

Tabelle 8: Ausgangstabelle *T* mit medizinischen Daten [33]

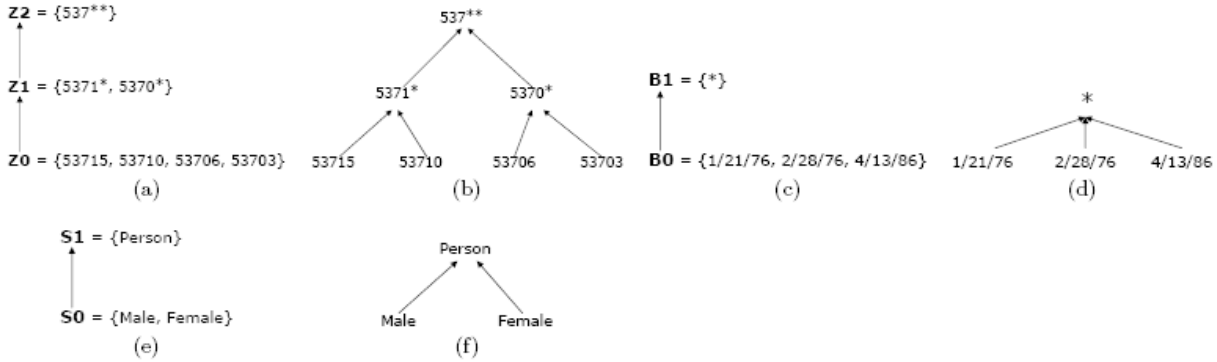

Abbildung 12: Domain Generalization Hierarchies für Zipcode (a,b), Birthday (c,d) und Sex (e,f) [33]

In unten angeführter Abbildung 13 werden verschiedene Generalisierungen geprüft. In der ersten Zeile (Abbildung 13 a) ist das Generalisierungsgitter, welches durch die Verbindung der *DGH* der beiden Attribute *Sex* und *Zipcode* entsteht, zu sehen.

Dieser Graph (Abbildung 13a, links) wird mittels eines modifizierten Breitensuchalgorithmus (Breitensuche durchsucht zuerst die Breite einer Ebene des Baumes [34]) durchsucht. Dabei wird jeder Knoten getestet und überprüft ob die Generalisierung, die er repräsentiert, die Ausgangstabelle in eine *k*-anonyme Tabelle umwandelt. Den Startknoten bildet <S0, Z0> welcher aber hinsichtlich der Tabelle *T* kein zufriedenstellendes Ergebnis liefert. Eine detaillierte Beschreibung des Breitensuchalgorithmus ist in [12] in Kapitel 3.1.1. zu finden.

So entsteht die zweite Abbildung (Abbildung 13a, mitte) wo Knoten <S1, Z0>, der die Bedingungen erfüllt, und Knoten <S0, Z1>, der die Bedingungen nicht erfüllt und entfernt wird, überprüft werden. Der Graph, der nach diesem Schritt entstanden ist, ist in Abbildung 13a, ganz rechts zu sehen. Da <S1, Z0> die Bedingungen erfüllt hat, muss sein Nachfolgeknoten nicht mehr überprüft werden, genauso wie Knoten <S0, Z2>. Mit den verbleibenden 2 Attributskombinationen <Birthday, Sex> und <Birthday, Zipcode> wird ebenso verfahren. Abbildung 13 zeigt die Ergebnisse jeweils ganz rechts in jeder Zeile (a,b und c):

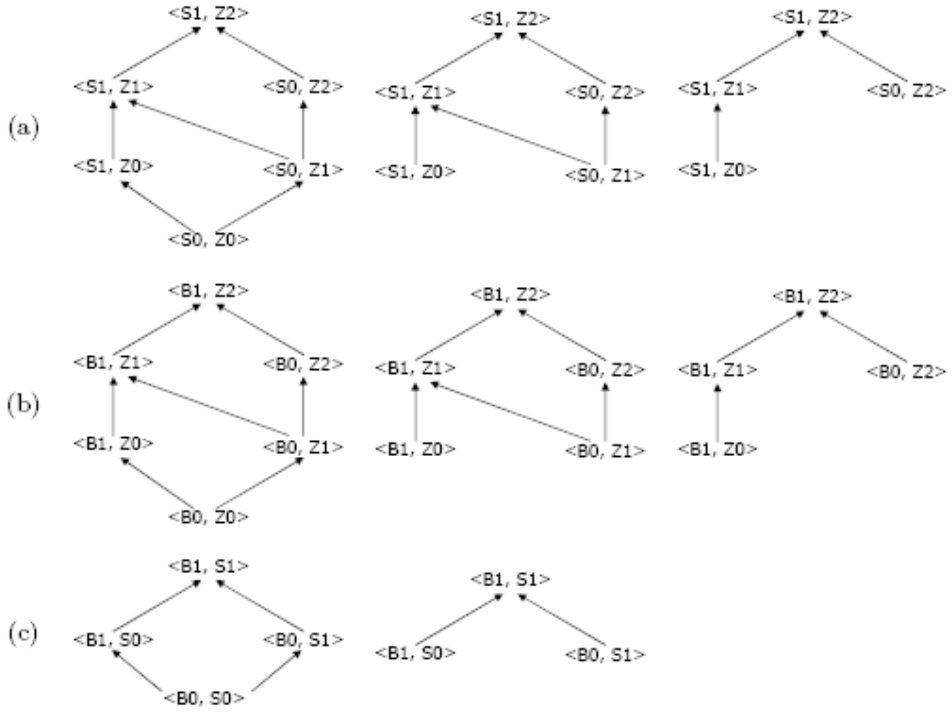

Abbildung 13: Durchsuchen der 2-Attributs-Kombinationen der Tabelle *T* [33]

Nach Zusammenfassen dieser drei verbleibenden Graphen ergibt sich ein 3-Attribut Graph der ebenfalls wieder mittels Breitensuche durchsucht wird. Den ausführlichen Algorithmus dazu stellen LeFevre, DeWitt und Ramakrishnan in [33] vor. Nachdem alle Knoten und Generalisierungen, die zu keiner *k*-anonymen Tabelle führen, entfernt wurden, ergibt sich folgender Graph, der in Abbildung 14 nach der a priori Suche [65] zu sehen ist. Dieser Graph stellt alle möglichen *k*-anonymen Generalisierungen dar, die sich aufgrund der Tabelle *T* und den *DGH* ergeben. In Abbildung 15 darunter ist der Graph zu sehen der ohne vorherigen Durchsuchen der 2-Attributskombinationen durchsucht hätte werden müssen:

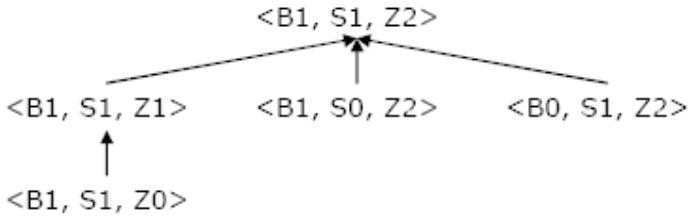

Abbildung 14: 3-Attribut-Grap als Ergebnis der a priori Suche [12]

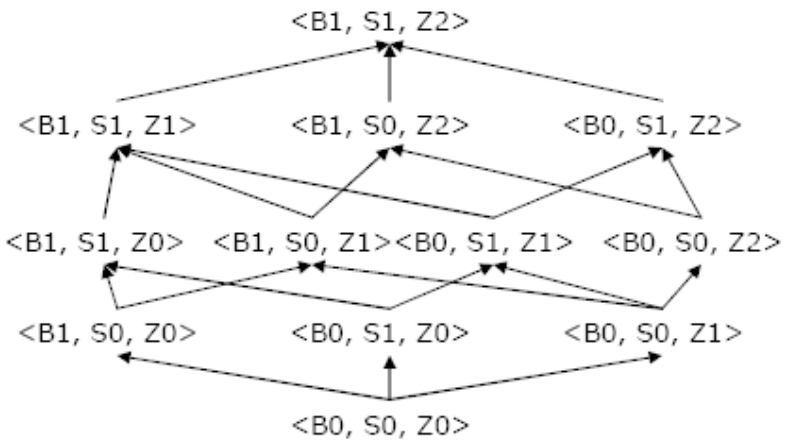

Abbildung 15: 3-Attribut-Graph ohne vorherigem Durchsuchen des 2-Attribut-Graphs [12]

Es ist deutlich zu sehen, dass dies einen hohen Mehraufwand bedeuten würde, da alle möglichen Kombinationen durchsucht werden, was sich wiederum negativ auf die Laufzeit der Berechnung auswirken würde.

Weiters stellen LeFevre, DeWitt und Ramakrishnan 2 optimierte Abwandlungen [33] des INCOGNITO-Algorithmus vor, die ebenfalls einem Performance-Vergleich mit anderen *k*-Anonymisierungs-Algorithmen unterzogen wurden:

Superroots-Incognito: Superroots Incognito ist eine beschleunigte Variante des ursprünglichen INCOGNITO-Algorithmus. Die Beschleunigung erfolgt hier bei der Überprüfung der Knoten bzw. der Bildung der *k*-anonymen Tabelle. Dies wird möglich, da die Berechnungen bei dieser Variante nicht auf Originaldaten, sondern auf bereits berechneten Generalisierungen beruhen.

Cube-Incognito: Auch bei dieser Variante wird versucht Rechenzeit zu sparen. Dies erfolgt hier durch Berechnung einer Generalisierung der Originaltabelle vor dem Start des eigentlichen Algorithmus. Da jedoch eine relativ lange Vorbereitungszeit von Nöten ist, wird der Vorteil der verkürzten Rechenzeit bei großen Tabellen zunichte gemacht. Man kann somit festhalten, dass die Vorteile bei dieser Abwandlung des Algorithmus eher bei kleineren Tabellen zu tragen kommen.

Abbildung 16 zeigt den Vergleich des INCOGNITO-Algorithmus und seinen beiden optimierten Varianten mit andern *k*-Anonymisierungs-Algorithmen [12]. Dieser Benchmark, welcher in [12] publiziert wird, basiert auf einer Datenbank welche 4.591.581 Tupel umfasst und sekundäre Attribute va-

riierender Größe beinhaltet, *k* ist gleich zwei. Das beste Ergebnis hinsichtlich der Rechenzeit liefert Superroots-Incognito, gefolgt vom Basis-Incognito Algorithmus. Platz drei belegt Cube-Incognito und verstärkt hiermit die oben getroffene Aussage, dass er bei großen Datenbanken kontinuierlich langsamer wird. Die Ergebnisse sind insofern aussagekräftig, da bis dato Tests für full-domain *k*-Anonymität nur auf kleinen Datenbanken mit 265 Tupel ausgeführt wurden [36].

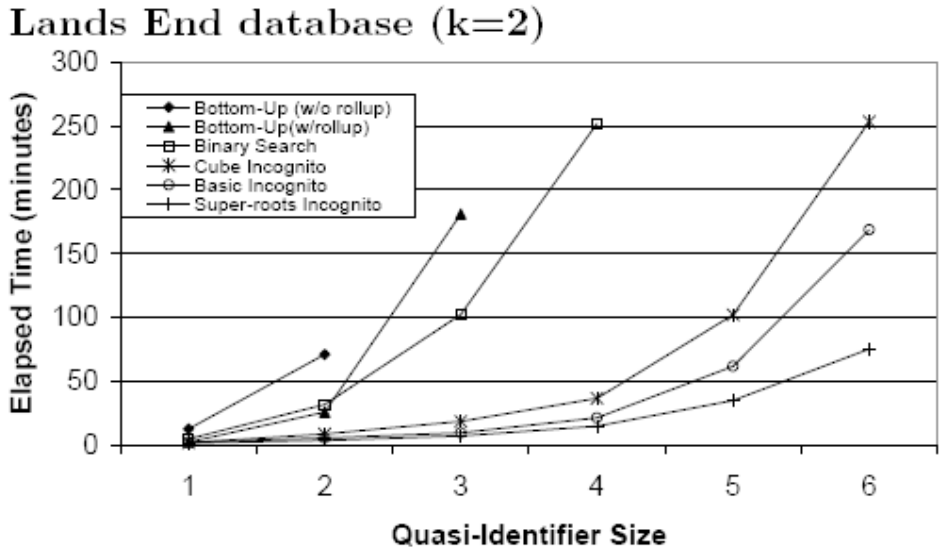

Abbildung 16: Performance-Vergleich verschiedener k-Anonymisierungsalgorithmen [12]

3.3.5. k-OPTIMIZE

Beim k-OPTIMIZE Algorithmus [1] handelt es sich um einen weiteren Optimierungsalgorithmus für die *k*-Anonymisierung. Dieser Algorithmus setzt beim optimalen Datenmanagement an, anstatt sich mit teuren Rechenoperationen des Sortierens zu befassen. Er kann in Bezug auf die gegebene Metrik auch für breite Spektren (getestet wurde in [1] für $k = 5$ bis 1000) von *k* für Input-Daten, die bei anderen Algorithmen für schlechte Ergebnisse sorgen, optimale *k*-anonyme Daten liefern [1].

Eine optimale *k*-Anonymisierung ist jene, die den Ausgangsdatensatz so wenig wie möglich verändert (siehe Kapitel 3.2). Der vorgestellte Algorithmus kann selbst bei kleinen Werten für *k* innerhalb weniger Sekunden/Minuten akzeptable Ergebnisse, die den Anforderungen der *k*-Anonymität genügen, liefern.

Der Ansatz dieses Algorithmus ist ein Anderer als bei den bisher vorgestellten Methoden. Erstens wurde bei anderen Methoden (Kapitel 3.3.1 bis 3.3.4) vorgeschlagen mit den Originaldaten als Input

zu beginnen, die mit greedy-Generalisierungsmethoden [39] in *k*-anonyme Form gebracht werden. k-OPTIMIZE hingegen startet mit einem voll generalisierten Datensatz, und wandelt diesen dann Schritt für Schritt durch Hinzufügen von zusätzlichen Informationen in minimal *k*-anonyme Form um. Obwohl dieser Ansatz etwas willkürlich wirkt, ist er ein wichtiger Ansatz der Methode von Bayardo und Agrawal [1].

Zweitens verwendet der Algorithmus eine Baumsuch-Strategie [40] welche den Suchraum verkleinert und ebenfalls die dynamische Neuanordnung des Suchbaumes unterstützt.

Drittens kommt eine neuartige Datenmanagement-Strategie [1] zum Einsatz, welche die Kosten für die Evaluierung von bereits gegebenen Anonymisierungen im Suchbaum verringert. Mit diesen Methoden können Sortier-Schritte drastisch reduziert werden. Einen Vergleich bisheriger Methoden mit dem Ansatz von Bayardo und Agrawal zeigt folgende Tabelle 9:

Algorithm	**für den Einsatz in der Praxis geeignet?**	**k-Anonymität garantiert?**
Sweeney-Datafly	**ja**	**nein**
Sweeney-Min Gen	**nein**	**optimal**
k-OPTIMIZE	**ja**	**optimal**

Tabelle 9: Vergleich verschiedener Methoden zur k-Anonymisierung [10]

Zur Generalisierung eines Attributes muss zuerst seine Domäne in Intervalle partitioniert werden. Zum Beispiel kann das Attribut Alter mit der Grunddomäne [1],[2],…,[30], entsprechend in die Domäne {[1, 10], [11, 20], [21,30]} generalisiert werden. Die Grunddomäne enthält alle möglichen Werte, welches das Attribut annehmen kann. Die Generalisierung in Form von Domänen kann dann als Unterteilung in Teilmengen der Grunddomäne beschrieben werden. Um die Schreibweise zu verkürzen, wird nur der kleinste Wert übernommen und oben erwähnte Domäne kann dann als {1, 11, 21} angeschrieben werden.

Um eine Tabelle in *k*-anonyme Form zu bringen, werden die Werte zunächst geordnet. Anschließend folgt die Unterteilung in Intervalle mittels oben bereits erwähnter Methode. Da diese Unterteilung das Minimum der späteren Generalisierungen festlegt, sollten die Intervalle nicht zu groß gewählt werden. Zuletzt erfolgen das Ordnen der Werte und das fortlaufende Nummerieren der Intervalle. Bayardo und Agrawal zeigen in ihrer Arbeit folgendes Beispiel (Abbildung 17):

Age			Gender		Marital Status			
<[10-29]	[30-39]	[40-49]>	<[M]	[F]>	<[Married]	[Widowed]	[Divorced]	[Never Married]>
1*	**2**	**3**	**4***	**5**	**6***	**7**	**8**	**9**

Abbildung 17: Ordnung der Domänen einer Tabelle mit 3 Attributen und neun möglichen Werten [10]

Die einzelnen Werte der Attribute können nun anhand ihrer Position identifiziert werden. Somit muss der kleinste Wert jeder Attributs-Domäne (markiert mit *) in jeder gültigen Generalisierung erscheinen, z.B. {[1, 4, 6]}.
Fügt man nun weitere Intervall-Startwerte zur Anonymisierung hinzu, verkleinern sich die Intervalle, und man erreicht eine Tabelle mit zusätzlichen Informationen und höherer Aussagekraft. Zum Beispiel ergeben sich durch die Anonymisierung {[1, 4, 5, 6 ,8]} folgende Intervalle in Bezug auf Abbildung 17:

- [10-49] (Attribut AGE)
- [M] & [W] (Attribut GENDER)
- [Married oder Widowed] und [Divorced oder Never Married] (Attribut MARITAL STATUS)

Die Anonymisierung, die demnach dann den höchsten Informationsgehalt besitzt, ist jene aller Werte {[1, 2, 3, 4, 5, 6, 7, 8, 9]} da hier alle Attribute übernommen wurden.
Da natürlich eine Anonymisierung, die möglichst viele Informationen der Originaltabelle enthält, Gefahr läuft nicht *k*-anonym zu sein, wird eine systematische Suche im Lösungsraum durchgeführt, um ein *k*-anonymes Ergebnis zu bekommen. Eine solche Suche wird durch das sogenannte OPUS-Framework [1] ermöglicht. Dieses Framework verwendet eine *„set-enumeration-search"*-Strategie [41] mit dynamischer Baumneuanordnung (Abbildung 18) und besitzt spezielle Abbruchbedingungen (genannt *„pruning"*), die auf Kosten in Bezug auf Rechenzeit ausgelegt sind, worauf im nächsten Absatz noch näher eingegangen wird.
Die *„set-enumeration-search"*-Strategie [41] arbeitet nach dem Prinzip der systematischen Erweiterung aller Subsets eines gegebenen Alphabets mittels Aufspannen eines Baumes, beginnend bei der Wurzel. Die systematische Erweiterung des Baumes für das Alphabet {1, 2, 3, 4} sieht mit der *„set-enumeration-search"*-Strategie wie folgt aus:

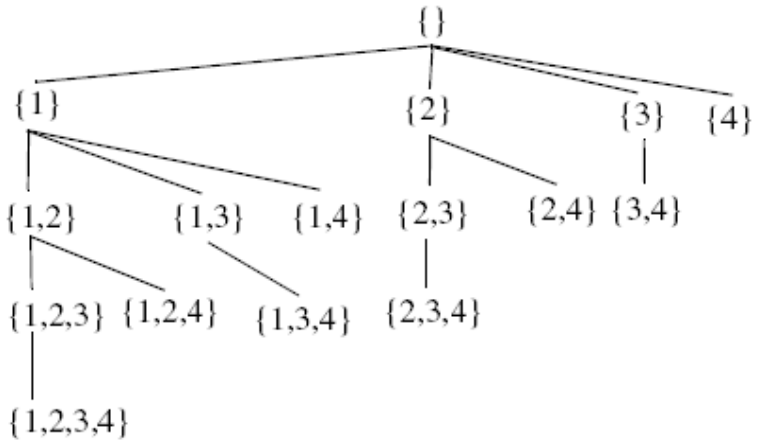

Abbildung 18: systematische Erweiterung des Baumes mit *set-enumeration-search*-Strategie [1]

Jeder Knoten in diesem Baum repräsentiert eine Anonymisierung, die auf ihre Kosten hin geprüft werden soll. Die Werte eines Knotens, die an die Blätter weitergegeben werden können, nennt man *„tail-set"*. Diese Werte stehen in geschwungenen Klammern und repräsentieren den jeweiligen Knoten, und können nur an direkte Kinder (Blätter) der Knoten weitergegeben werden - und zwar so, dass jedes Kind den ersten noch nicht vergebenen Wert erhält.

Theoretisch wäre nun nur noch ein Vergleich aller erzeugten Knoten untereinander notwendig, um die optimale Lösung zu finden. Da der Suchraum aller Knoten im Normalfall jedoch viel zu groß ist, um in akzeptabler Zeit Ergebnisse zu bekommen, kann dieser Ansatz so nicht direkt umgesetzt werden.

Bayardo und Agrawal haben ihren Algorithmus so implementiert, dass er beginnt nach Abbruchbedingungen zu suchen *(„pruning"*), um dadurch den Suchraum zu verkleinern. Die Suche findet innerhalb der Knoten, Blätter und *tail-sets* statt. Eine solche Abbruchbedingung gilt als gefunden, wenn ein Knoten höhere Kosten verursacht als die bisher gefundene optimale Lösung. Genauere Angaben zu den Abbruchbedingungen bzw. zur Kostenberechnung können in [1] nachgelesen werden.

Eine zweite Möglichkeit den Suchraum klein zu halten, ist das Entfernen unnötiger Werte, welches in Abbildung 20 gezeigt wird. Es wird vorausgesetzt, dass ein Datenset bereits durch Anonymisierung in 5 Äquivalenzklassen unterteilt wurde. Wird nun ein weiterer Wert hinzugefügt, unterteilt sich die Anonymisierung erneut (Abbildung 19). Sind die neuen Klassen kleiner als *k*, kann dieser neu hinzugekommene Wert als unnötig betrachtet und gelöscht werden.

Abbildung 19: Werden die Äquivalenzklassen durch Hinzufügen eines neuen Wertes in Klassen unterteilt (dargestellt durch die strichlierten Linien) die alle kleiner als k sind, wird dieser Wert als unnötig betrachtet und entfernt. [1]

Schließlich kann auch durch das Neuanordnen des *tail-sets* ein positiver Performance-Zuwachs hinsichtlich der Rechenzeit erreicht werden. Weiters wird dieser Zuwachs durch das Löschen der richtigen Werte unterstützt. So teilt jedes neue Element *im tail-set* diesen Knoten in weitere Anonymisierungen. Die Anzahl der so entstehenden neuen Äquivalenzklassen wird für jedes Element berechnet, und im *tail-set* absteigend angeordnet. Dadurch werden Anonymisierungen, die viel verändern würden, möglichst bald getestet, wohingegen Anonymisierungen, die wenig bewirken bis zum Schluss aufgehoben werden.

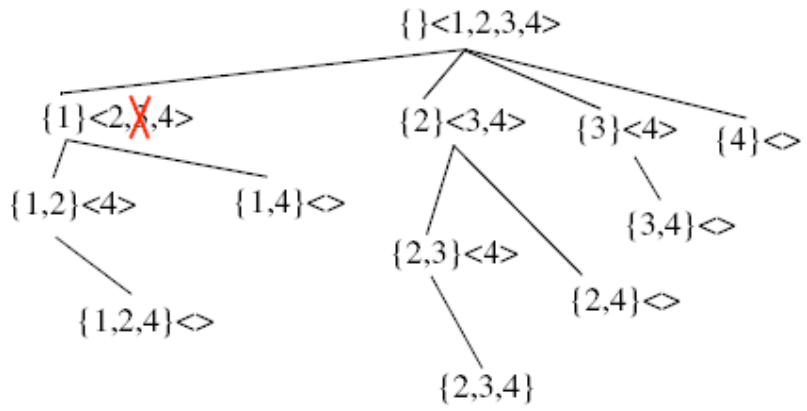

Abbildung 20: Entfernen des Wertes 3 aus dem *tail-set* von Knoten {1} [1]

Zur effizienten Berechnung der Äquivalenzklassen wurden diese in 3 Kategorien unterteilt [1]:

- Kategorie 1 – Äquivalenzklassen, die sich beim Spalten der Wurzelknotens ergeben
- Kategorie 2 – Äquivalenzklassen, die durch das Hinzufügen von Werten in den *tail-sets* entstehen
- Kategorie 3- alle Äquivalenzklassen, die in dem Suchbaum vorkommen können (inkl. derer die entstehen, wenn den *tail-sets* neue Werte hinzugefügt werden).

Ein simpler Ansatz wäre das Durchsuchen und neu Sortieren des kompletten Baumes nach jeder Veränderung. Aufgrund oben erwähnter Kategorisierung und weiterer Unterteilung der Äquivalenzklassen, ist es jedoch nicht notwendig jedes Mal den kompletten Datensatz neu zu durchsuchen. Stattdessen reicht das Durchsuchen der neuen relevanten Klassen aus.

Für das Benchmarken [1] des k-OPTIMIZE Algorithmus wurde ein Datenset mit 9 Attributen und insgesamt 30.162 Tupeln verwendet. Dieses Datenset unterstützte sehr feines Partitionieren des Alters (mögliche Partitionierungen: *adult_fine* und *adult_coarse*, siehe [38]), nämlich genau 1 Wert pro vorhandenem Alter, das zu einem Alphabet von über 160 Werten führte *(„adult_fine“*).
Für einen weiteren Test wurde das Attribut Alter in 15 Partitionen unterteilt, welches ein Alphabet mit immer noch 100 Werten als Resultat hatte [1]. Hierfür wurde die sogenannte „Coarser“ Partitionierung *(„adult_coarse“*) verwendet, welche die Reduktion der Flexibilität zur Erkennung einer Anonymisierung durch den Algorithmus zur Folge hat.
Als Metrik kam die *„Discernibility Metric“* (Kapitel 3.4.2) zum Einsatz. Als weitere Einstellung wurde die Unterdrückung von Werten erlaubt / nicht erlaubt / nur teilweise erlaubt (Attribut *„sup_limit“* [1]).
Der Algorithmus wurde in *C++* implementiert und auf einer dedizierten 2,8 Ghz *Intel Xeon* Maschine, mit *Linux* OS (Kernel 2.4.20) und *gcc* v3.2.1 ausgeführt. Für die Sortierschritte wurde die von *gcc* zur Verfügung gestellte *C* Funktion *qsort* [42] verwendet.

K-OPTIMZE liefert die billigste Anonymisierung hinsichtlich der Laufzeit in Relation zu im Vorhinein definierten minimalen Kosten. Diese Anonymisierung entsteht aus einem Subset des gegebenen Baumes. Existiert keine Anonymisierung welche billiger als jene ist, die im Vorhinein mit Hilfe der Kosten *c* definiert wurde, wird mit Hilfe der *PRUNE* Funktion der Baum so beschnitten, dass dies möglich wird.

Pseudocode der K-OPTIMIZE Funktion und der *PRUNE*-Funktion [1]:

```
K-OPTIMIZE(k, head set H, tail set T, best cost c)
    ;; This function returns the lowest cost of any
    ;; anonymization within the sub-tree rooted at
    ;; H that has a cost less than c (if one exists).
    ;; Otherwise, it returns c.
    T ← PRUNE-USELESS-VALUES(H, T)
    c ← MIN(c, COMPUTE-COST(H))
    T ← PRUNE(H, T, c)
    T ← REORDER-TAIL(H, T)
    while T is non-empty do
        v ← the first value in the ordered set T
        H_new ← H ∪ {v}
        T ← T − {v}        ;; preserve ordering
        c ← K-OPTIMIZE(k, H_new, T, c)
        T ← PRUNE(H, T, c)
    return c
```

```
PRUNE(k, head set H, tail set T, best cost c)
    ;; This function creates and returns a new
    ;; tail set by removing values from T that
    ;; cannot lead to anonymizations with cost
    ;; lower than c
    if COMPUTE-LOWER-BOUND(k, H, H ∪ T) ≥ c
        then return ∅
    T_new ← T
    for each v in T do
        H_new ← H ∪ {v}
        if PRUNE(H_new, T_new − {v}, c) = ∅
            then T_new ← T_new − {v}
    if T_new ≠ T then return PRUNE(H, T_new, c)
    else return T_new
```

Abbildung 21: Pseudocode von K-OPTIMIZE und der Funktion „PRUNE“ [1]

Einstellungen, die für den Test verwendet wurden [1]:

- für *k* wurden folgende Werte verwendet: 1000, 500, 250, 100, 50, 25, 10 und 5
- Anzahl der erlaubten Unterdrückungen: keine / max. 100 / keine Begrenzung (*inf*)
- grobe / feine Partitionierung des Attributes *Alter* (*coarse* / *fine*)
- verwendete Metriken: CM oder DM (siehe Kapitel 3.4)

Lässt man die Kosten außer Betracht, und beobachtet nur die Performance der K-OPTIMIZE Funktion, ergeben sich folgende Graphen Abbildung 22:

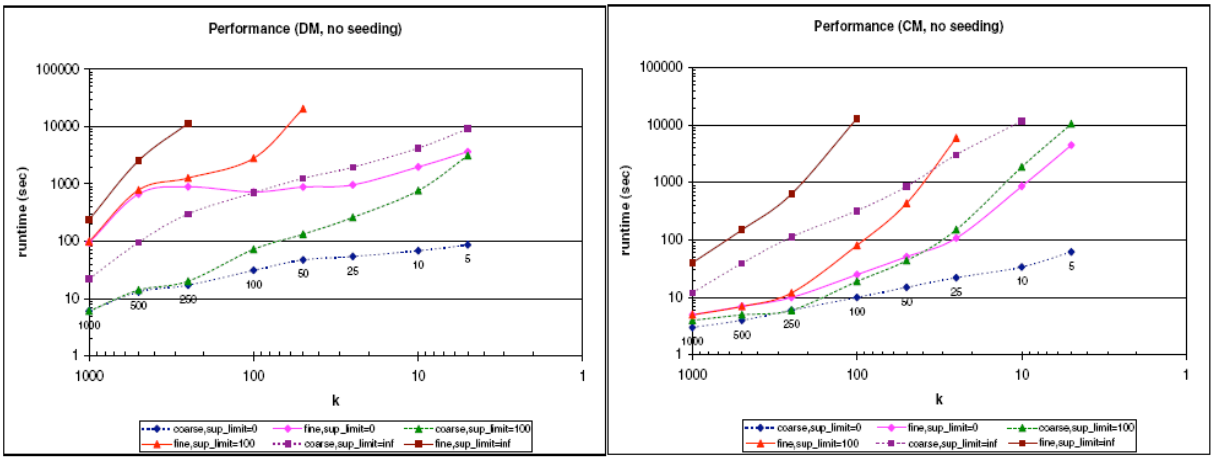

Abbildung 22: Performance der K-OPTIMIZE Funktion [1]

Wie sich hier beobachten lässt, performt der Algorithmus für das nahezu gesamte Spektrum von *k* sehr gut wenn keine Unterdrückung erlaubt ist (sup_limit = 0). Erlaubt man eine Unterdrückung, ist zu beobachten, dass der Algorithmus für kleiner werdende *k* mit längerer Laufzeit reagiert. Wie erwartet fordert auch der erhöhte Suchraum bei feiner Partitionierung signifikant höhere Rechenzeiten.

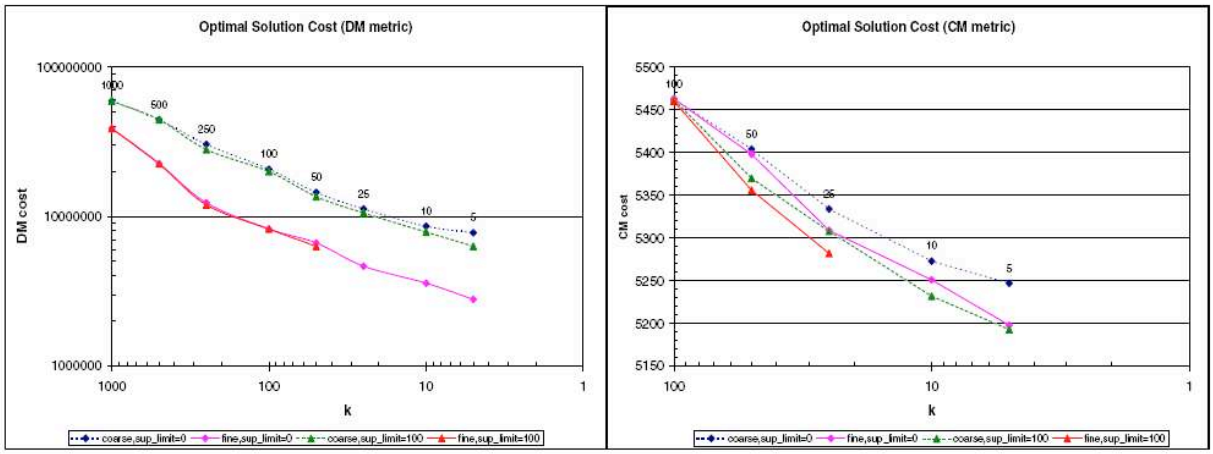

Abbildung 23: Optimale Kosten [1]

Betrachtet man den Algorithmus hinsichtlich seiner Kosten (Abbildung 23) kann man erkennen, dass hier die Variante mit feiner Granularität bez. des Attributes *Alter* für beide Metriken die besten Ergebnisse liefert. Interessant ist ebenfalls dass eine unendliche Anzahl von Unterdrückungen das Ergebnis nicht verändern konnte, weshalb die Graphen hier weggelassen wurden.

Auf einen detaillierten Vergleich mit anderen Algorithmen wird in [1] verzichtet. Es wird lediglich erwähnt, dass „greedy-Algorithmen“ [39] meist sehr schnell terminieren und das Ergebnis meist nicht optimal ist und den Anforderungen der *k*-Anonymität (vgl. Kapitel 3.2) nicht genügt. Hingegen produzieren „genetic algorithms“[44] und „simulated annealing“ [43] Anonymisierungen höherer Qualität, konvergieren jedoch zu langsam. Leider werden zu diesem Algorithmus keine namentlichen Beispiele in [1] erwähnt.

Der Ansatz, der laut Bayardo und Agrawal eine annehmbare Alternative bietet, ist eine 2-phasige „hill-climbing“ Methode [1], welche die Vorteile von greedy und stochastischen Methoden vereint. Diese Methode durchläuft zuerst eine Anonymisierungsphase, gefolgt von einer Generalisierungsphase, in der Werte mit hohen Kosten iterativ entfernt werden. Dies geschieht solange bis keine Optimierung hinsichtlich der Kosten erreicht werden kann. Danach erfolgt die 2. Phase der Optimierung, in der Werte hinzugefügt werden, bis wieder keine weitere Optimierung mehr möglich ist. Diese Me-

thode besitzt kein Abbruchkriterium, weshalb der Algorithmus händisch gestoppt werden muss. Folgende Abbildung zeigt einen Vergleich von K-OPTIMIZE *(„complete“)* und dem stochastischen Ansatz:

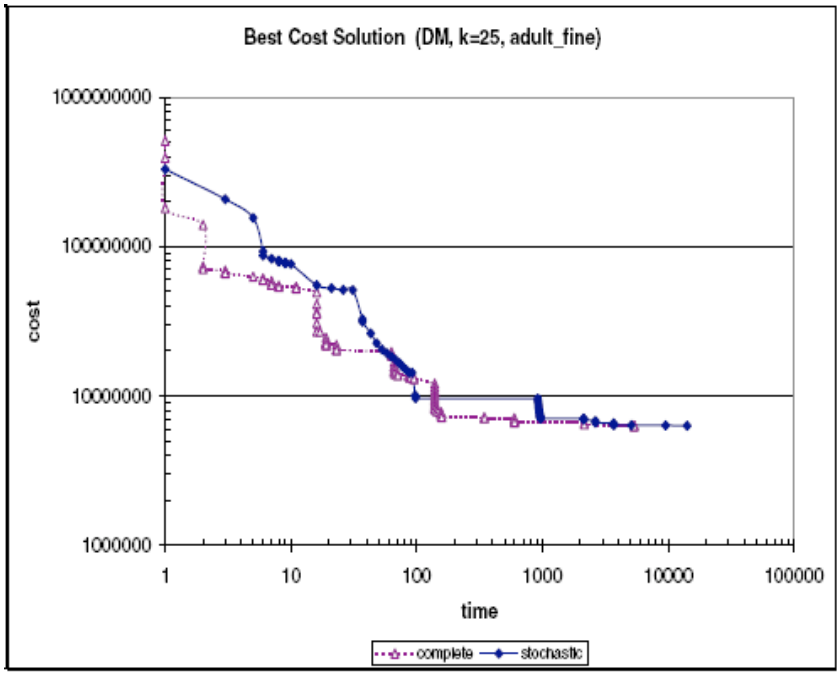

Abbildung 24: Vergleich von K-OPITMIZE mit stochastischer Methode [1]

Obwohl K-OPTIMIZE nicht das schnelle Finden einer Lösung als Ansatz wählt (sondern eine *k*-optimale Lösung), ist deutlich zu sehen, dass dieser Algorithmus die optimale Lösung fast 3 mal schneller als die stochastische Variante finden kann (5.359 gegenüber 14.104 Sekunden). An dieser Stelle lässt sich festhalten, dass K-OPTIMIZE aufgrund seiner Laufzeit und der Qualität der gefundenen Lösung, die Basis für einen Algorithmus bietet, welcher in der Praxis eingesetzt werden kann.

3.3.6. Multidimensionale k-Anonymität

In ihrer wissenschaftlichen Arbeit „Mutlidimensional *k*-Anonymity“ [45] stellen LeFevre, De Witt und Ramkrishnan einen weiteren neuen Ansatz zur *k*-Anonymisierung vor. Multidimensionale *k*-Anonymität ist ein Modell der multidimensionalen Partitionierung für *k*-Anonymisierung. Im Gegensatz zur eindimensionalen Partitionierung, wird bei der mehrdimensionalen Methode nicht nur ein Wertebereich (Dimension) für die Generalisierung eines Attributes herangezogen, sondern mehrere. Optimale *k*-Anonymisierung, welche dieses Modell verwendet, ist NP-hard. Dies wird auch von A. Meyerson und R. Williams in [28] bzw. [1] deutlich gezeigt. Der Begriff „Np-hard“ bezeichnet besonders aufwändig zu berechnende Probleme und stammt aus der Komplexitätstheorie. Im Worst

Case ist die Obergrenze der Äquivalenzklassen im multidimensionalen Fall O(k), wohingegen diese Grenze im eindimensionalen Fall linear mit der Anzahl der Datenbankeinträge wächst. Eine weitere Variation des multidimensionalen Falles hat eine Obergrenze von 2k [45].
Aufgrund dieser Ergebnisse stellen LeFevre, De Witt und Ramakrishnan einen neuen multidimensionalen *k*-Anonymisierungsalgorithmus vor [45], der schneller als optimale Algorithmen wie „MinGen“ (Kapitel 3.3.1) arbeitet, und trotzdem zufriedenstellende Resultate erzeugt.
In einer relationalen Datenbank werden Attributen Domänen mit bestimmten Wertebereichen zugeordnet. Globale Generalisierung zum Zweck der *k*-Anonymisierung versucht diese Domänen der Quasi-Identifiers zu generalisierten oder in veränderte Werte über zu führen.
Globale Generalisierung kann in zwei Subklassen unterteilt werden, eindimensionale und multidimensionale globale Generalisierung [45]. Bei einer eindimensionalen globalen Generalisierung gibt es pro Quasi-Identifier eine Funktion, welche die Domäne des jeweiligen Attributs generalisiert, wohingegen bei der multidimensionalen Generalisierung nur eine einzige Funktion für alle Quasi-Idfentifier verwendet wird.

Eindimensionale Domänen-Partitionierung:
Bei der eindimensionalen globalen Generalisierung ist eine vorausgehende eindimensionale Partitionierung der Domäne jedes Attributs notwendig. Es erfolgt eine eindimensionale, einander nicht überlappende, Unterteilung jeder Domäne in Intervalle, die den gesamten Wertebereich umfassen. Werte, die innerhalb eines Intervalls liegen, werden bei der Generalisierung auf diesen Wertebereich abgebildet (Tabelle 11 , Attribute *Age* und *Zipcode*) [45].

Multidimensionale Domänen-Partitionierung:
Hierbei erfolgt keine Unterteilung der einzelnen Domänen, sondern die des gesamten Wertebereichs aller Attribute (ein *x*-dimensionaler Raum für *x* Attribute) in nicht überlappende, multidimensionale Regionen zur Durchführung der multidimensionalen, globalen Generalisierung [45].

Um die Vorgehensweise zu veranschaulichen, wurde die folgende Tabelle 10, welche medizinische Daten enthält, einmal mittels eindimensionaler (Tabelle 11) und einmal mittels mehrdimensionaler Generalisierung (Tabelle 12) in *k*-anonyme Form gebracht.

Patient Data

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
27	Female	53712	AIDS
28	Male	53711	Hang Nail

Tabelle 10:Ausgangstabelle mit Patientendaten [45]

Age	Sex	Zipcode	Disease
[25-28]	Male	[53710-53711]	Flu
[25-28]	Female	53712	Hepatitis
[25-28]	Male	[53710-53711]	Brochitis
[25-28]	Male	[53710-53711]	Broken Arm
[25-28]	Female	53712	AIDS
[25-28]	Male	[53710-53711]	Hang Nail

Tabelle 11: eindimensionale Anonymisierung der Ausgangstabelle mitt. Generalisierung von *Zipcode* [45]

Age	Sex	Zipcode	Disease
[25-26]	Male	53711	Flu
[25-27]	Female	53712	Hepatitis
[25-26]	Male	53711	Brochitis
[27-28]	Male	[53710-53711]	Broken Arm
[25-27]	Female	53712	AIDS
[27-28]	Male	[53710-53711]	Hang Nail

Tabelle 12: mehrdimensionale Anonymisierung der Ausgangstabelle [45]

Der Unterschied beider Ansätze ist sofort erkennbar: Wird bei eindimensionaler Generalisierung (Tabelle 11) das gleiche Attribut auch immer auf denselben Wertebereich der *k*-anonymisierten Tabelle abgebildet, ist dies bei Verwendung der multidimensionalen Methode (Tabelle 12) nicht so. Hier wird für ein und dasselbe Attribut „53711“ zweimal keine Veränderung vorgenommen, und einmal wird das Attribut in die Generalisierung [53710-53711] überführt. Somit ist diese Form der Generalisierung flexibler und kann sich den Ausgangsbedingungen bzw. Anforderungen besser anpassen.

Um sich die Partitionierung besser vorstellen zu können, werden die Attribute $A=\{A_1, A_2, \ldots, A_x\}$ der Tabelle T in einen x-dimensionalen Raum projiziert, wobei jedes Tupel der Tabelle einen Punkt in diesem Raum darstellt.

Folgende Abbildung 25 stellt die Ausgangstabelle (a), ihre eindimensionale Partitionierung der Domäne des Attributs *Zipcode* (b) und ihre multidimensionale Partitionierung der Domänen der Attribute *Zipcode* und *Age* (c) dar.

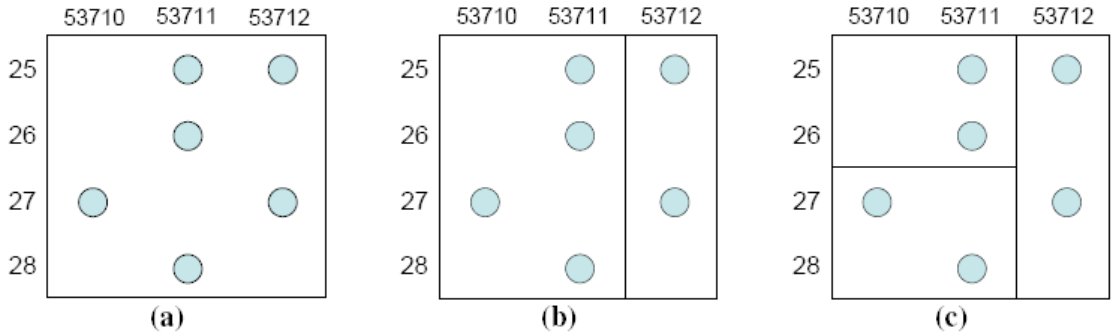

Abbildung 25: Darstellung der Ausgangstabelle (a) und ihrer eindimensionalen (b) und mehrdimensionale (c) Partitionierung. [45]

Dank Abbildung 25 ist der Vorteil multidimensionaler Partitionierung sofort ersichtlich: Durch die flexible Einteilung des Raumes sind mehrere, kleinere Regionen mit weniger Tupeln möglich, was dazu führt, dass die Werte nicht so stark generalisiert werden müssen, was wiederum zu einer höheren Datenqualität führt. Für eine korrekte Darstellung der Tupel im Raum, hätte man bei obiger Abbildung auch noch das Geschlecht mit einbringen müssen, darauf wurde aber in [45] aufgrund der Übersichtlichkeit verzichtet, da auch so schon der Vorteil der Methode erkennbar ist.

Eine weitere Größe, die eine wichtige Rolle bei diesem Ansatz spielt, ist die maximale Größe einer Region, die bei einer multidimensionalen Partitionierung entstehen kann. Wie im vorigen Absatz bereits erwähnt, ist eine Unterteilung in möglichst kleine Regionen gewünscht, da dadurch nur geringe Abänderungen der Daten notwendig sind.

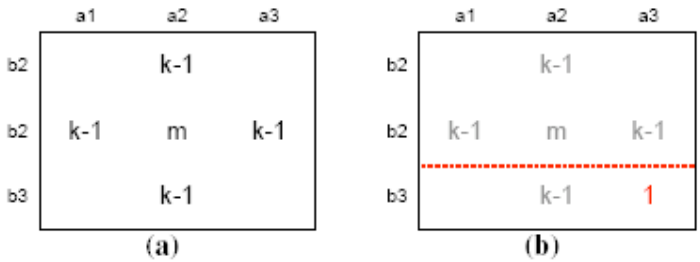

Abbildung 26: Erst nach Hinzufügen eines Punktes (b) ist ein Schnitt möglich [45]

Abbildung 26 zeigt eine Menge von Punkten im 2-dimensionalen Raum, welche erst nach Hinzufügen eines weiteren Punktes (b) teilbar wird. Grund dafür ist die Voraussetzung, dass ein Schnitt, der die Fläche partitioniert, achsenparallel sein muss und nur dann möglich ist, wenn die Anzahl der Punkte in den neu entstandenen Bereichen mindestens k beträgt [45].
Daraus folgt: Eine Partitionierung existiert nur für die Mengen von Punkten P im x-dimensionalen Raum mit $|P| > 2x(k-1) + m$, wobei m die maximale Anzahl von Kopien eines Punktes aus P ist.
Die maximale Anzahl der Punkte (= Tupel) eines Intervalls ist im Fall der eindimensionalen Partitionierung, nicht wie im multidimensionalen Fall, unabhängig von der Gesamtanzahl der Tupel der Tabelle T, sondern wächst linear mit der Anzahl der Datensätze und beträgt $O(|T|)$ [45].

Ein weiterer, etwas lockerer Ansatz der strikten multidimensionalen Partitionierung wird ebenfalls von LeFevre, De Witt und Ramakrishnan in [45] vorgestellt. Diese sogenannte relaxed multidimensionale Partitionierung ist eine multidimensionale lokale Generalisierung, und erlaubt eine Überlappung der Regionen in die der Wertebereich eingeteilt wird.
Gibt es keine, den Regeln und Anforderungen entsprechende, Möglichkeit einer strikten multidimensionalen Partitionierung, wird auf die relaxed multidimensionale Partitionierung zurückgegriffen.

Würde man obige Tabelle 10 zur Veranschaulichung heranziehen und eine 3-anonyme Tabelle aus ihr generieren wollen, bei welcher der *Zipcode* das einzige Quasi-Identifier Attribut ist, müsste man bei jedem Tupel auf die Generalisierung [53710-53712] zurückgreifen. Wendet man in diesem Fall jedoch die relaxed multidimensionale Partitionierung an, ist ein Überlappen der Bereiche erlaubt und man erhält folgende Tabelle 13:

Age	Sex	Zipcode	Disease
25	Male	[53710-53711]	Flu
25	Female	[53710-53711]	Hepatitis
26	Male	[53711-53712]	Brochitis
27	Male	[53710-53711]	Broken Arm
27	Female	[53711-53712]	AIDS
28	Male	[53711-53712]	Hang Nail

Tabelle 13: relaxed multidimensionale Partitionierung für ein einziges Quasi-Identifier Attribut Zipcode [45]

Multidimensionale Partitionierung kann in zwei Schritten erreicht werden: Im ersten Schritt werden multidimensionale Regionen definiert, die den Domänenbereich umfassen, im zweiten Schritt wird die Tabelle, aufgrund der erfolgten Partitionierung, in *k*-anonyme Form gebracht. In [45] wird ein Algorithmus vorgestellt, der eine Abwandlung eines Baumkonstruktionsalgorithmus darstellt, der mittels kleinerer Modifikationen auch für die relaxed multidimensionale Partitionierung angepasst werden kann. Der Rechenaufwand des Algorithmus lässt sich mit $O(nlogn)$ beziffern, wobei hier mit *n* die Anzahl der Tupel, die in Tabelle *T* vorhanden sind, bezeichnet wird. Dieser Algorithmus ist flexibel bei der Auswahl der Dimension für die Partitionierung in Regionen. Jede Region umfasst minimal *k* und höchstens $2x(k-1)+m$ Punkte, wobei hier *m* die maximale Anzahl der Kopien eines beliebigen Punktes ist.

Diese Lösung ist lt. [45] nur einen konstanten Faktor weit von der optimalen Lösung entfernt. Die optimale Generalisierung (hierbei wurde die Durchschnittsmetrik (Kapitel 3.4.3) verwendet) der Tabelle *RT* (Tabelle 10) ergibt den Wert $C_{AVG}(RT) >= 1$, der sich von der „worst-case"-Generalisierung durch relaxed multidimensionale Partitionierung, nur um den Faktor zwei unterscheidet: $C_{AVG}(RT) <= 2$.

Experimente und Ergebnisse [45]:

Das Hauptziel der Bemühungen LeFevres, De Witts und Ramakrishnans ist eine hohe Qualität der Ergebnisse, die mithilfe ihres Greedy-Algorithmus erzeugt wurden. In ihrer Arbeit vergleichen sie die Performanz ihres Greedy-Algorithmus mit der von Incognito (Kapitel 3.3.4) (Full-Domain Generalization [36]) und K-Optimize (Kapitel 3.3.5) (single-dimensional Partitionierung [1][33][35]). Als Ausgangspunkt dienten sowohl künstlich erzeugte, als auch reale, natürlich Daten. Wie zu erwarten war, schnitt der vorgestellte multidimensionale Algorithmus generell besser ab, als Algorithmen, die einen eindimensionalen heuristischen Ansatz [30][69][70] bzw. eine stochastische Suche [71][72] heranziehen.

Zur Veranschaulichung des Unterschieds der erzeugten Daten, kann man Abbildung 27 heranziehen, welche die verschiedenen Ergebnisse, welche durch eindimensionale (erste Zeile der Abb.) und multidimensionaler (zweite Zeile der Abb.) Partitionierung erreicht wurden, deutlich visualisieren:

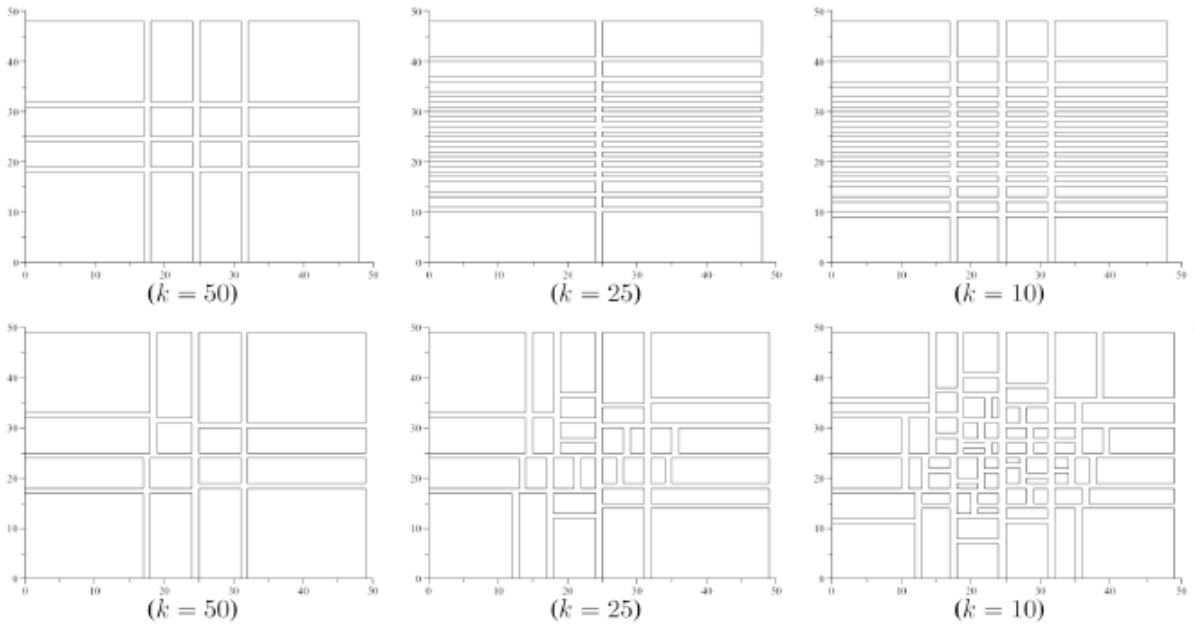

Abbildung 27: Vergleich von eindimensionaler und multidimensionaler Partitionierung [45]

Man sieht, dass der multidimensionale Ansatz eine höher variante Verteilung der Daten zur Verfügung stellt, und dadurch eine geringere Generalisierung der Daten erreicht, wohingegen der eindimensionale Ansatz die Daten zu großen Regionen zusammenfasst, was wiederum auch eine sehr starke Generalisierung bedeutet. Weiters ist auch eine Linearisierung der Attribute in der ersten Zeile zu beobachten.

Für weitere Ergebnisse inklusive einer Veranschaulichung anhand von Diagrammen sei auf [45] verwiesen.

3.4. Metriken für die Bestimmung von optimalen Generalisierungen

Gibt es mehrere *k*-minimale Generalisierungen einer Tabelle (Kapitel 3.3) kann durch die Berechnung bestimmter Metriken die beste Lösung in Hinsicht der Datenqualität gefunden werden. Hierbei wird der Informationsverlust, der durch die Generalisierung der Tabelle entsteht, berechnet. Auf diesem Ergebnis baut schlussendlich die Auswahl der Metrik auf.

3.4.1. Präzisionsmetrik

Der Grad der Veränderung wird bei der Präzisionsmetrik („Precision Metric" – „Prec") [30] durch das Verhältnis zwischen der Generalisierungsstufe einer Zelle *h* und allen möglichen Generalisierungsstufen angegeben. Die Präzision der Tabelle wird mittels Eins minus der Summe der Veränderungen, normalisiert durch die Gesamtzellenanzahl, angegeben.

Definition [5]: Gegeben sei eine Tabelle $PT(A_1,\ldots,A_{Na})$, und DGH_A sei die „domain generalization hierarchy" der Attribute *A*; Tabelle $RT(A_1,\ldots,A_{Na})$ sei eine Generalisierung der Tabelle *PT*. Die Präzision der Tabelle *RT* geschrieben *Prec(RT)* berechnet sich folgendermaßen (wobei *N* für die Anzahl der Tupel der Tabelle *PT* steht):

$$Prec(RT) = 1 - \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N} \frac{h}{|DGH_{Ai}|}}{|PT| \cdot |N_A|}$$

Formel 5: Berechnung der Domain Generalization Hirarchy [3]

Beispiel: Wenn *PT = RT*, so ist jeder Wert in der Grunddomäne ($h = 0$) und *Prec(RT)* = 1. Befindet sich hingegen jeder Wert in der höchsten Generalisierungsstufe, ist jedes $h = |DGH_{Ai}|$ und *Prec(RT)* = 0.

3.4.2. Unterscheidbarkeitsmetrik

Bei der Umwandlung einer Tabelle in k-anonyme Form werden viele Tupel bzw. Spalten auf die gleichen Werte generalisiert. Die Unterscheidbarkeitsmetrik („Discernibility Metric" – „DM") [1] versucht zu berechnen wieviele verschiedene Werte pro Tupel bzw. Spalten nach der Generalisierung noch erhalten geblieben sind.
Die Berechnung der Discernability Metric erfolgt mittels zweier Summen. Ist ein Tupel von anderen Tupeln nicht mehr unterscheidbar, wird es mit einem Strafpunkt im Ausmaß der identischen Datensätze versehen. Fällt ein Tupel in eine Äquivalentklasse der Größe *j* von identischen Tupeln, so bekommt es einen Strafpunkt der Größe *j* zugeordnet. Dies erfolgt durch Berechnung der ersten Summe. Wird ein Datensatz unterdrückt (entfernt), so bekommt er einen Strafpunkt äquivalent zur Ge-

samtanzahl aller in der Tabelle vorhandenen Tupel, zugewiesen. Dies erfolgt durch die Berechnung der zweiten Summe der Formel.

Definition [5]: Gegeben sei die Tabelle *PT*, und die Tabelle *RT* sei eine Generalisierung der Tabelle *PT*. |*E*| bezeichnet die Größe der Äquivalenzklasse, in der sich ein Tupel befindet, und |*PT*| steht für die Anzahl der Tupel der Originaltabelle *PT*.

$$C_{DM}(RT) = \sum_{\forall |E| \geq k} \left| E^2 \right| + \sum_{\forall |E| < k} |E|\,|PT|$$

Formel 6: Berechnung der Discernability Metric [3]

3.4.3. Durchschnittliche, normalisierte Äquivalenzklassengrößenmetrik

Die „Normalized Average Equivalence Class Size Metric“ [45] berechnet ebenfalls wie die “Discernibility Metric” die Qualität der Generalisierung mittels der Äquivalenzklassengröße, ist aber leichter und schneller zu berechnen.

Definition: Gegeben sei die Tabelle *PT*(A_1,…,A_{Na}) mit der Anzahl an Tupel |*PT*|; Tabelle *RT*(A_1,…,A_{Na}) sei eine Generalisierung der Tabelle *PT* . Die „Normalized Average Equivalence Class Size Metric“ der Tabelle *RT* $C_{AVG}(RT)$ wird wie folgt berechnet:

$$C_{AVG}(RT) = \left(\frac{|PT|}{Anzahl\ der\ Äquivalenzklassen} \right) / (k)$$

Formel 7: Berechnung der Normalized Average Equivalence Class Size Metric [3]

Alle 3 Metriken berechnen die Qualität der Generalisierung im Zuge einer *k*-Anonymisierung von Datentabellen und machen somit eine Interpretation der Ergebnisse bez. der Verwendbarkeit der Daten möglich.

3.5. *l*-Diversität

Bei der *l*-Diversität handelt es sich um die Weiterentwicklung der *k*-Anonymität. Wie in Kapitel 3.2.1 gezeigt wurde, ist *k*-Anonymität anfällig für Angriffe, die Hintergrundwissen beinhalten. *l*-Diversität erweitert die *k*-Anonymität um weitere Beschränkungen und Forderungen, um Anonymität für sensible Daten zu gewährleisten.

Bei der im Kapitel 3.2.1 erwähnten Homogeneity- bzw. Background-Knowledge-Attack, spielt vorhandenes Hintergrundwissen eine große Rolle. Da der Angreifer die öffentliche *k*-anonyme Tabelle kennt, ist nicht auszuschließen, dass er sein Hintergrundwissen einsetzen wird, um sensible Attribute auszulesen oder Individuen zu identifizieren. Dies ist möglich wenn er weiß, dass eine spezielle Person in der Tabelle vorhanden ist. Weiters ist es wahrscheinlich, dass er über nicht sensible Attribute (sekundäre Identifikationsmerkmale) Bescheid weiß und einige sensible Attribute ausschließen kann. Eine weitere Möglichkeit wäre dann noch das Wissen über die Verteilung von sensiblen und nicht sensiblen Attributen, also demographisches Hintergrundwissen. Ein Beispiel dafür wäre, das geringe Erkrankungsrisiko an Krebs bei Japanern [6].
Um das mögliche Hintergrundwissen eines Angreifers bei der Anonymisierung einer Tabelle zu berücksichtigen, werden mittels des Bayes-Theorems [46] in [1] die Begriffe *„prior belief“* und *„observed belief“* erläutert. Der Begriff *„prior belief“* steht für das eigentliche Hintergrundwissen einer Person bevor sie noch Einsicht in die *k*-anonyme Tabelle hatte. Da sich das Wissen nach dem Einblick in die Tabelle ändert, ist das Ziel, dieses *„observed belief“* genannte Wissen, möglichst äquivalent zum *„prior belief“* zu halten. In anderen Worten ausgedrückt, soll die Differenz zwischen *„prior belief“* und *„observed belief“* möglichst gering sein. Der Angreifer soll also durch den Einblick in die Tabelle nur an wenige bzw. keine neuen Informationen gelangen können.
Weiters werden die Begriffe *„positive disclosure“* [47] und *„negative disclosure“* [47] erläutert. Kann ein sensibles Merkmal einer Person, nach Bekanntwerden der generalisierten Tabelle, mit hoher Wahrscheinlichkeit identifiziert werden, so ist die Rede von *„positive disclosure“*. Dies war bei Alices oben erwähnter Homogeneity-Attack (Kapitel 3.2.1.1) der Fall, bei der sie das sensible Merkmal Bobs ausfindig machen konnte.
Können Werte des sensiblen Attributes mit hoher Wahrscheinlichkeit ausgeschlossen werden (vergl. „Background-Knowledge Attack“, Kapitel 3.2.1.2), so ist die Rede von *„negative disclosure“*.
Aber nicht immer sind *„positive“* und *„negative disclosure“* so fatal wie bei den beiden vorgestellten Angriffsszenarien (Kapitel 3.2.1). Alice kann beispielsweise keinen großen Nutzen daraus zie-

hen, wenn sie weiß, dass Bob nicht an Ebola leidet, da ihr *„prior belief“* für dieses Ereignis ohnehin klein war und somit die Differenz zwischen *„prior belief“* und *„observed belief“* ebenfalls gering ausfällt.

Doch bei diesem Ansatz gibt es auch Nachteile. Die Verteilung der sensiblen und nicht-sensiblen Attribute in der Gesamtbevölkerung Ω sind meistens nicht bekannt. Diese bräuchte man aber, um den *„observed belief“* zu berechnen. Außerdem muss bei der Veröffentlichung der Daten mit einbezogen werden, dass Angreifer unterschiedliche Stufen des Hintergrundwissens besitzen. Hier kann man sich also nicht sicher sein, was der Angreifer nun wirklich weiß. Dies könnte auch negativ für den Angreifer ausfallen wenn Bob eine für sein Alter ungewöhnliche Krankheit hat, und der Angreifer den Zusammenhang zwischen Krankheit und Alter kennt und diese Krankheit somit ausschließt. Ein Angreifer, der nicht über dieses Wissen verfügt, könnte Bobs Krankheit wiederum als sensibles Attribut in Erwägung ziehen.

Die mögliche Antwort auf die oben gezeigten Probleme und Unzulänglichkeiten bildet die *l*-Diversität. Das Wort Diversität, welches für „Vielfalt“ oder auch „Mannigfaltigkeit“ steht, bezieht sich auf das sensible Attribut, das pro *k*-Gruppe – eine Gruppe, die aus mindestens *k* Tupeln besteht, die sich nicht durch die Kombination ihrer sekundären Identifikationsmerkmale unterscheiden lässt, mindestens *l* verschiedene Werte annehmen muss ($l \geq 2$) [1].

Definition *l*-Diversität:

Eine *k*-Gruppe ist *l*-divers, wenn das sensible Attribut der Gruppe zumindest *l*-verschiedene Ausprägungen aufweist. Eine Tabelle *T* ist *l*-divers, wenn jede *k*-Gruppe *l*-divers ist [1].

Durch die Einführung dieser Bedingung wird die „Background-Knowledge-Attack“ (Kapitel 3.2.1.2) deutlich erschwert. Der Angreifer muss nun $l - 1$ sensible Werte durch *„negative disclosure“* ausschließen können, um an die tatsächlichen Werte zu kommen. Wandelt man eine Tabelle entsprechend der Definition von *l*-Diversität ab, so kann aufgrund der verschiedenen Werte auch keine „Homogeneity-Attack“ (Kapitel 3.2.1.1) mehr durchgeführt werden. Ein Beispiel dafür zeigt die Tabelle 14, die eine 3-diverse Abwandlung der Tabelle 1 darstellt, und somit nicht mehr durch die beiden erwähnten Angriffe sensible Informationen preisgeben kann. Es ist deutlich erkennbar, dass Tabelle 14 hinsichtlich des sensiblen Attributes 3-divers ist (Diagnose ist in 3 verschiedenen Ausprägungen vorhanden). Weiter ist jede k-Gruppe für sich ebenfalls 3-divers (siehe jeweils die ersten, zweiten und dritten 4 Zeilen).

	Non-sensible Attribute			*sensible Attribute*
	PLZ	**Alter**	**Nationalität**	**Diagnose**
1	1030	28	USA	Alkoholismus
2	1010	29	Österreich	Alkoholismus
3	1011	21	Japan	Manisch Depressiv
4	1030	23	Österreich	Manisch Depressiv
5	1110	50	Indien	Drogenabhängig
6	1110	55	USA	Alkoholismus
7	1150	47	Österreich	Manisch Depressiv
8	1140	49	Österreich	Manisch Depressiv
9	1030	31	Österreich	Drogenabhängig
10	1030	37	Indien	Drogenabhängig
11	1010	36	Japan	Drogenabhängig
12	1012	35	Österreich	Drogenabhängig

	Non-sensible Attribute			*sensible Attribute*
	PLZ	**Alter**	**Nationalität**	**Diagnose**
1	103*	≤ 40	*	Alkoholismus
4	103*	≤ 40	*	Manisch Depressiv
9	103*	≤ 40	*	Drogenabhängig
10	103*	≤ 40	*	Drogenabhängig
5	11**	> 40	*	Drogenabhängig
6	11**	> 40	*	Alkoholismus
7	11**	> 40	*	Manisch Depressiv
8	11**	> 40	*	Manisch Depressiv
2	101*	≤ 40	*	Alkoholismus
3	101*	≤ 40	*	Manisch Depressiv
11	101*	≤ 40	*	Drogenabhängig
12	101*	≤ 40	*	Drogenabhängig

Tabelle 14: 3-diverse Abwandlung der Tabelle 1

Folgende Vorteile von *l*-Diversität lassen sich gegenüber der *k*-Anonymität herausstreichen:

- *l*-Diversität schützt vor „Homogeneity-„ und „Background-Knowledge-Attacken“
- *l*-Diversität benötigt kein Wissen über die Verteilung der Attribute in der Population
- verschiedene Stufen von Hintergrundwissen müssen nicht extra in Betracht gezogen werden, da als Voraussetzung für *l*-Diversität der Unterschied zwischen *prior* und *observed belief* kleingehalten bzw. gar nicht vorhanden sein darf.
- *l*-Diversität, ist wie *k*-Anonymität, monoton. Das bedeutet, dass eine weitere Generalisierung einer *l*-diversen Tabelle wieder *l*-divers ist. Diese Eigenschaft kann man sich bei der Umwandlung einer Tabelle zu einer *l*-diversen Tabelle zunutze machen. Um schneller ans Ziel und zu einer Lösung zu kommen, verwenden viele Algorithmen Suchstrategien (z.B. „set-

enumeration-search“, siehe Kapitel 3.3.5, Abbildung 18) im Lösungsraum, welche die Suche an einem Ast vorzeitig abbrechen können, wenn gewisse Abbruchkriterien erfüllt sind. Ohne die Eigenschaft der Monotonie könnten durch den vorzeitigen Stopp wichtige Lösungen verloren gehen [1].

Die bisher getroffenen Aussagen über die *l*-Diversität beruhen auf der Annahme, dass pro Datensatz immer nur ein sensibles Attribut vorhanden ist. Tabelle 15 zeigt, dass sich bei Vorkommen mehrerer sensibler Attribute und der Einhaltung der *l*-Diversität, neue Herausforderungen ergeben. Bei dieser Tabelle handelt es sich um eine *k*-Gruppe, die sowohl auf das Attribut *Medikation* als auch auf das Attribut *Diagnose* bezogen 3-divers ist (jeweils ohne das andere Attribut). Kann ein Angreifer, der weiß, dass sich eine bestimmte Person in der Tabelle befindet, auch ausschließen, dass diese Person manisch depressiv ist, ist der Angriff bereits erfolgreich, da das zweite sensible Merkmal identifiziert wurde.

	PLZ	Alter	Nationalität	Diagnose	Medikation
1	102*	< 50	*	Manisch Depressiv	Medikament A
2	102*	< 50	*	Ödipus	Medikament B
3	102*	< 50	*	Drogen	Medikament C
4	102*	< 50	*	Alkohol	Medikament C

Tabelle 15: Tabelle mit 2 sensiblen Attributen (Diagnose und Medikation)

Eine Möglichkeit, solche Angriffe zu verhindern bzw. es dem Angreifer zu erschweren, wäre die zufällige Anordnung der Werte eines sensiblen Attributes innerhalb einer *k*-Gruppe, um die Korrelation zwischen den beiden Attributen aufzubrechen. Da dies aber eine derart große Veränderung der Daten darstellt, wäre dieser Ansatz für wissenschaftliche Studien inakzeptabel. In [1] wird eine alternative Methode skizziert, die sich besser dafür eignet.

Wir gehen von einer Tabelle mit den sensiblen Attributen $S (V_1, V_2, \ldots, V_m)$ aus. Zunächst werden die Attribute $V_1, \ldots, V_m$ ignoriert, und die Tabelle wird in eine *l*-diverse Tabelle – bezogen auf Attribut *S* – umgewandelt. $\phi = 1 - 1 / l$ und t_s sei die Zahl der Tupel mit dem Wert *s* des Attributs *S* innerhalb einer *k*-Gruppe. Für jedes *s'* innerhalb einer k-Gruppe werden $[\phi t_{s'}]$ Tupel mit der Ausprägung *s'* ausgewählt, und die Werte der Attribute $V_1, \ldots, V_m$ für diese Datensätze unterdrückt, wodurch zusätzliche Sicherheit gewährleistet wird. Der Informationsgehalt sinkt dadurch im Kontext dazu ab. Trotzdem hat diese Methode auch ihre Vorteile:

- Die Information des sensiblen Attributes S bleibt vollständig erhalten

- Die Information der restlichen sensiblen Attribute $(V_1, \ldots, V_m)$ bleibt teilweise erhalten
- Die Information durch die Korrelation der einzelnen Attribute wird nicht verfälscht.

In [1] wird zu Zwecken der Performanz der *l*-Diversitäts Algorithmus mit einem „normalen“ *k*-Anonymisierungsalgorithmus verglichen (der verwendete *k*-Anonymisierungsalgorithmus wird nicht weiter namentlich erwähnt). Bei dem Vergleich wurde die Verarbeitungszeit der beiden Algorithmen für eine Datenbank (UCI Machine Learning Repository [59]) mit 45.222 Tupel gemessen. Die Tabelle wurde dabei in eine *l*-diverse Form gebracht. Unten stehende Abbildung 28 zeigt das Ergebnis des Vergleichs. Wie zu sehen ist, nimmt die Verarbeitungszeit bei steigender Anzahl sekundärer Identifikationsmerkmale bei beiden Algorithmen in etwa gleich schnell zu.

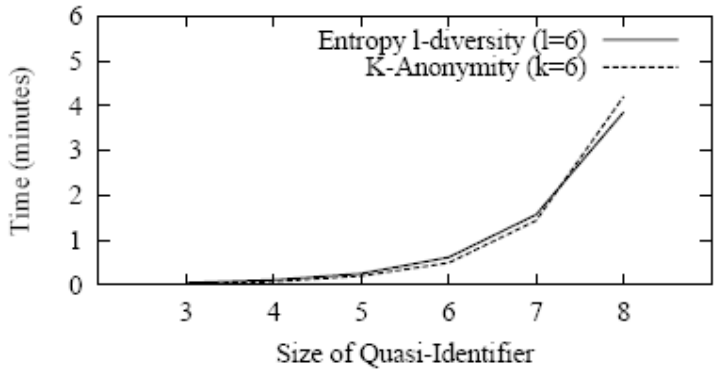

Abbildung 28: Vergleich l-Diversität und k-Anonymität [1]

Ein weiterer wichtiger Faktor, welcher neben der Geschwindigkeit des Algorithmus noch von Bedeutung ist, ist die Brauchbarkeit der erhaltenen Daten. Dabei wurden folgende Metriken eingesetzt:

- Präzisionsmetrik [30] (Kapitel 3.4.1)
- Unterscheidbarkeitsmetrik [1] (Kapitel 3.4.2)
- durchschnittliche, normalisierte Äquivalenzklassengrößenmetrik [45] (Kapitel 3.4.3)

Das Ergebnis der Versuche war, wie bei der Verarbeitungszeit, annähernd äquivalent. Beide getesteten Algorithmen schnitten in etwa gleich ab. Auch die Generalisierung der Daten war bei *l*-Diversität nicht unbedingt stärker, als es für das Erreichen einer *k*-anonymen Tabelle notwendig gewesen wäre. Für weitergehende Vergleiche und Tests in Bezug auf Performanz wird auf [1] verwiesen.

Am Ende dieses Abschnitts sei festzuhalten, dass aufbauend auf den Erkenntnissen aus diesem Teil der Arbeit, *l*-Diversität als eine durchaus gelungene und brauchbare Weiterentwicklung von *k*-Anonymität gesehen werden kann. Performance-Tests und die Brauchbarkeit der Daten lassen diesen Schluss folgern und zeigen, dass *l*-Diversität höheren Schutz der Daten, bei etwa gleicher Verarbeitungsdauer im Vergleich zu *k*-Anonymität, bieten kann.

4. Beschreibung des Fallbeispiels SPICS Soul

In Kapitel 3 wurden unterschiedliche Methoden aus dem Gebiet der Anonymisierung (Kapitel 3.3) und Konzepte der Telematikplattform für Medizinische Forschungsnetze für die Pseudonymisierung für medizinische Forschungszwecke (Kapitel 3.1) vorgestellt. Der Punkt in dem sich Anonymisierung und Pseudonymisierung am stärksten unterscheiden, ist die Wiederherstellungsmöglichkeit der Daten. Ist bei der Anonymisierung eine sehr starke Verfremdung der Daten (z. B. durch Generalisierung) zu beobachten, ist bei der Pseudonymisierung, mit Hilfe von Patientenlisten und Pseudonymisierungsdiensten, eine gewollte Wiederherstellung des Personenbezugs möglich.
Ein wichtiger Aspekt von SPICS Soul ist, dass die teilnehmenden Forschungseinrichtungen autonom handeln sollen. Zentral ausgelegte und orientierte Konzepte sollen vermieden werden. Das heißt, dass eine Gewaltentrennung unter den Forschungseinrichtungen stattfinden muss. Jede Forschungseinrichtung hat ihre eigene Patientendatenbank und ihre eigene Prüfkomponente, welche die eingehenden Anfragen bearbeitet.
Auf Grund der Definitionen von Anonymität (Kapitel 2.4) und Pseudonymität (Kapitel 2.5), ist bei der Verwendung von Daten für wissenschaftliche Forschungszwecke, eine Anonymisierung ausreichend, da hierfür eine Rückidentifikation nicht notwendig ist. Geht man von einer klinischen Forschungsstudie (z.B. über die Wirkungsweise und Symptome bzw. Nebenwirkungen) eines Medikamentes aus, ist ein Personenbezug nicht zwingend notwendig. Hingegen kann eine Rückidentifikation in klinisch fokussierten Behandlungsszenarien oder im Lehrbereich gewollt sein (Mitteilung der Diagnose, Besprechung des Krankheitsverlaufs oder der Behandlungsmöglichkeiten eines bestimmten Patienten).

4.1. Allgemeine Beschreibung

Das Projekt SPICS Soul an der Forschungsgruppe Industrial Software (INSO) am Institut für Automation der TU Wien dient zur Speicherung von medizinischen Behandlungsdaten von Patienten zu Forschungszwecken. Diese Daten können zwischen Ärzten und medizinischen Einrichtungen für Forschung und Lehre ausgetauscht werden.
Kernpunkt dieser Arbeit aus Datenschutz- und Informationssicherheitssicht sind sensible Daten, die in SPICS gespeichert werden. Dies können u.a. das Geburtsdatum, Geschlecht, Krankheitsbild und andere ausgewählte Behandlungsdaten (Medikation, Dauer der Behandlung) sein.

Sollen Datenpakete, die oben erwähnten Daten enthalten, zwischen Ärzten oder Institutionen ausgetauscht werden, muss der Arzt die gewünschten Daten per Anfrage anfordern. Genau zu diesem Zeitpunkt soll eine Komponente überprüfen, ob diese Daten weitergegeben werden können ohne die Anonymität der Patienten zu gefährden, oder ob dies nicht der Fall ist und eine Abfrage abgelehnt werden muss.
Diese Entscheidung bedarf zuvor genauer Betrachtung der angeforderten Daten (um welche Daten handelt es sich? Wie groß ist die Grundmenge?) und muss aufgrund von verschiedenen Kriterien gefunden werden, die viele Attribute der Patienten mit einbeziehen (Alter, Nationalität, Geschlecht u.v.m.).

Wie bereits in Kapitel 3.2.1 gezeigt wurde, reicht in vielen Fällen die klassische *k*-Anonymisierung der Daten per Definition (Kapitel 3.2) nicht mehr aus. Sind die anonymisierten Daten nicht divers genug, ist es einem Angreifer möglich sensible Daten in den Paketen zu identifizieren und diese einer Person eindeutig zuzuordnen.
Eine große Rolle bei der Entscheidungsfindung spielt die Kombination und die zugrundeliegende Menge (die Grundmenge) der auszutauschenden Daten. Diese wiederum stehen in enger Verbindung mit weiteren Kriterien wie statistischen Verteilungen und Häufigkeiten. Es ist bereits hier zu sehen, dass die Entscheidung für den Austausch der Daten, ja oder nein, eine stark dynamische ist, die viele Aspekte mit einbeziehen muss. Im Endeffekt entscheidet der Arzt ob die Daten weitergegeben werden oder nicht. Die Prüfkomponente bildet für diese Entscheidung lediglich eine Grundlage.

4.2. Anforderungen an Informationsflüsse in klinisch fokussierten Forschungsnetzen

In klinisch fokussierten Forschungsnetzen steht die Ableitung der wissenschaftlichen Daten aus dem Behandlungsprozess im Mittelpunkt. Durch Zusammenführung der Daten aus dem Behandlungsprozess für Forschungszwecke kann als Nebeneffekt die klinische Befundkommunikation verbessert werden. Laut [11] darf, kann und soll die wissenschaftliche Nutzung der hier in einer klinischen Datenbank zusammengeführten Informationen nicht online erfolgen, sondern nur im asynchronen Zugriff auf, eigens an die wissenschaftliche Fragestellung adaptierte, exportierte Teilmengen der Behandlungsdaten.

4.2.1. Klinische Kompetenz

Die Behandlung von Patienten mit chronischen oder besonders schwer behandelbaren Krankheiten erfolgt immer öfter in Kooperation mehrerer Ärzte. Durch das Hinzuziehen spezialisierter Mediziner sowie besonders erfahrener Behandlungsteams kann so eine höhere Diagnostik- / Therapiemöglichkeit erlangt werden, als dies allein im Bereich der Regelversorgung möglich ist. Die besonderen Aufgaben spezialisierter Zentren sind [1]:

- Entwicklung von Verfahren und Therapien solcher Erkrankungen
- Evaluierung solcher Verfahren
- Einführung dieser Verfahren auf breiter Basis im klinischen Alltag
- Beratung der Patienten die auf der Suche nach einer zweiten Meinung über die Möglichkeit ihrer Behandlung sind
- Beratung der an der Versorgung beteiligten Ärzte in Bezug auf die Behandlung ihrer Patienten

Da verschiedene Kompetenzen im medizinischen Bereich heutzutage auch verteilt sein können, wird die Behandlung komplexer Krankheitsprobleme auf zahlreiche Expertenschultern verteilt, um den für den Patient höchsten Effizienzgrad, mit dem Ziel des optimalen Behandlungserfolges, zu erreichen. Die Datensammlung für Forschungszwecke soll somit dem behandelnden Arzt einen elektronischen Zugriff auf Vorbefunde oder Befunde mitbehandelnder Ärzte ermöglichen. Durch diesen resultierenden Vorteil soll nicht zuletzt eine Motivation der Ärzte und Patienten für die Teilnahme am klinisch fokussierten Forschungsnetz erfolgen [1].

4.2.2. Wissenschaftliche Kompetenz

Für die Erarbeitung neuer Behandlungsverfahren von Patienten mit chronischen, seltenen oder besonders schwerer Krankheiten ist die Betrachtung großer Populationen notwendig. Die Anzahl von Patienten, die für eine schnelle Umsetzung der Behandlungsverfahren benötigt wird, ist hier in der Regel zu klein, um international Schritt halten zu können. Dies führt dazu, dass die wissenschaftliche Forschung für die Erprobung neuer Behandlungsverfahren gar nicht in Betracht gezogen wird, und

die kommerzielle Markteinführung nicht auf Ergebnissen aus dem eigenen Land basiert. Durch die Bündelung der wissenschaftlichen Kompetenz verschiedener spezialisierter Forschungsgruppen, kann aber trotzdem Spitzenforschung betrieben werden. Trotz der Zusammenarbeit im Bereich der Forschung und der Unterstützung neuer Kommunikationsmedien, treten gerade im Bereich der klinischen Forschung und Lehre große Probleme auf [1]:

- Die Verfügbarkeit von Informationen aus dem Behandlungsprozess ist für die klinische Forschung unentbehrlich. Dies können Angaben über den Krankheitsverlauf, Risikofaktoren, Behandlungsverfahren, oder wie im Fall von SPICS Soul, auch Daten hinsichtlich der Medikation sein. Durch Verteilung biologischer Proben ist der Zusammenhang von Behandlungsinformation und erhobener Laborbefunde nicht immer gegeben. Biologisches Material wird an geeignete spezialisierte Zentren gesendet und für die spätere Untersuchung aufbewahrt

- Die Zusammenführung klinischer Daten aus dem Behandlungsprozess mit den Ergebnissen der wissenschaftlichen Untersuchung ist aus wissenschaftlicher Sicht wünschenswert. Die Voraussetzung dafür ist eine redundante Erfassung klinischer Daten die oft aus Zeitmangel nicht oder in minderer Qualität erfolgt.

4.2.3. Anforderungen der Nutzer

Patienten, Ärzte und Wissenschaftler, die am Forschungsprozess beteiligt sind, versprechen sich durch ihre Beteiligung am Forschungsnetz unterschiedliche Vorteile. Die Netzstrukturen in einem klinisch fokussierten Forschungsnetz sollen nicht nur der Sammlung von forschungsrelevantem Wissen, sondern auch der Verbesserung des Forschungsflusses in der klinischen Medizin dienen.

Patienten:
Patienten sind in der Regel, nach vorhergehender Einwilligungserklärung, bereit, ihre im Laufe einer Erkrankung gewonnener Daten zur Erforschung ihrer Erkrankung, zur Verfügung zu stellen. Um diesen Willen zur Kooperation nicht uneingeschränkt geltend zu machen, ergeben sich folgende Ansprüche an den Umgang patientenbezogener Daten und biologischer Proben[1]:

- Die Erfassung und Auswertung der Daten und Proben soll dem eigenen Behandlungsprozess zu Gute kommen.

- Forschungsergebnisse, welche für den Patienten relevant sein könnten, sofern sie zum Beispiel seine Gesundheit betreffen, sollen diesem auch mitgeteilt werden.
- Die Identifikation eines Patienten darf nur den im Behandlungsprozess Beteiligten möglich sein aber nicht z.B. wissenschaftlich tätigen Mitarbeitern im Forschungsnetz.
- Der Missbrauch patientenbezogener Daten soll durch eine klare Abgrenzung für deren Verwendung vermieden werden (z.B. Behandlungsdaten, die für die Behandlung, nicht aber für die Forschung erforderlich sind)
- Um einen unkontrollierten Gebrauch bzw. Missbrauch der Daten auszuschließen, darf eine Weitergabe an externe, nicht ins Forschungsnetz eingebundene Dritte, nicht erfolgen.
- Ziele in der Forschung im Bereich des Netzes und erzielte Ergebnisse sowie Informationen über Personen, welche patientenbezogene Daten verwaltet haben, müssen transparent sein.
- Ein Zurückziehen der Kooperation mit dem Forschungsnetz muss jedem Patienten jederzeit möglich sein.

Reng et al. erwähnen in [1], dass die Zusammenarbeit von Patient und Forschungsnetz durch geeignete Einverständniserklärungen und Aufklärungen zur Datenerhebung geregelt werden.

Behandelnde Ärzte:

Um Diagnostik und Therapie für Patienten verbessern zu können, erwarten Ärzte, die im Behandlungsprozess tätig sind, eine Optimierung der Prozessstrukturen in einem klinisch fokussierten Netz. Darüber hinaus erwarten sie ebenfalls eine Verminderung redundanter Arbeitsvorgänge. Daraus ergeben sich die nachfolgenden Ansprüche [1]:

- Informationen, betreffend Patienten, müssen verwechslungsfrei und fehlerlos sein. Patientenakten sollen bei einer Wiedervorstellung zum Zweck der Fortschreibung einer Krankengeschichte verfügbar sein.
- Um den Informationsfluss und –Status aller Beteiligten, sowohl aus diagnostischen als auch therapeutischen Bereichen, zu gewährleisten, ist eine lückenlose patientenbezogene und zeitnahe Zusammenführung sämtlicher Informationen des Forschungsdatensatzes als zwingend anzusehen.
- Keine Doppelerfassung von Daten – eine Ableitung der wissenschaftlich relevanten Daten aus den klinischen Daten, ist aus Gründen der Arbeitserleichterung und der Qualitätssicherung anzustreben

- Der behandelnde Arzt muss in der Lage sein mit Patienten Kontakt aufnehmen zu können, um einen neuen Behandlungsprozess, der sich auf Grund von wissenschaftlichen Untersuchungen ergibt, zu besprechen.
-

Wissenschaftler:

Gerade bei chronischen und besonders schweren Erkrankungen sind die Rückgriffe auf fallbezogene, frühere Informationen oftmals von großem Interesse, wenn Prognose und Therapieeffekte betrachtet werden sollen. Somit erwarten die am Forschungsnetz beteiligten Wissenschaftler einen besseren klinischen Bezug ihrer Forschung. Folgende Punkte beschreiben die Ansprüche der Wissenschaftler [1]:

- Zentrumsübergreifende Zusammenführung von diagnostischen und therapeutischen Daten soll helfen, eine möglichst große Zahl von Patienten der wissenschaftlichen Evaluation zur Verfügung zu stellen.
- Eine übergreifende, epidemiologische Aus- und Bewertung der fallbezogenen Informationen muss durch Definition unterschiedlicher Sortieralgorithmen und Datenfilter möglich sein.
- Der Zusammenhang der klinisch erhobenen Daten mit den Ergebnissen der Forschung (biologische Proben) muss hergestellt werden können, um so die Wertigkeit der Untersuchung für den Behandlungsfall besser beurteilen zu können.

4.3. Anforderungen an den Datenfluss in wissenschaftlich fokussierten Netzen

Forschungsdaten, die außerhalb des Behandlungszusammenhanges bearbeitet werden, sollen exportiert und dabei pseudonymisiert oder anonymisiert werden. Im Gegensatz dazu gestaltet das Konzept für forschungsfokussierte Netze nicht den Komplex der Erhebung und Nutzung von Daten in Behandlungszusammenhang, sondern ausschließlich die Prozessierung von Daten, die für die wissenschaftliche Forschung und in der Lehre bereit gestellt werden. Ein solches generisches Konzept muss die Bedingung erfüllen dass Forschungsdatenbanken große Kollektive abbilden können, die über einen längeren Zeitraum beobachtet werden, ohne die Vertraulichkeit der Informationen anzutasten. Daraus leiten sich laut Reng et al. [1] drei Grundforderungen zur Sicherung der Qualität und Vertraulichkeit der Daten ab:

- Die Einführung eines Verfahrens zur sicheren Identifikation eines Patienten, um die Verwendung von Daten, die den Patienten im Klartext identifizieren würden, zu vermeiden. Das Verfahren soll eindeutige nicht sprechende Zeichenketten als Patientenidentifikatoren (PID) generieren und die Erzeugung von Synonymen und Homonymen vermeiden, wenn der Patient zeitgleich von mehreren, zum Forschungsnetz gehörenden Einrichtungen, behandelt wird.
- Die Qualitätssicherung der Forschungsdaten, um Vollständigkeit und Plausibilität soweit als möglich zu sichern.
- Die Einführung eines treuhänderischen Dienstes (Pseudonymisierungsdienst), der die Schlüssel zur Aufdeckung von Pseudonymen zentral speichert. Für die Pseudonymisierung der Forschungsdaten soll ein Verfahren hoher Sicherheit eingesetzt werden, dass der Tatsache gerecht wird, dass große Mengen von Daten zentral gespeichert und für Forschung und Lehre verfügbar gemacht werden sollen.

4.4. Anforderungen an die Prüfkomponente für SPICS

Um eine Komponente, welche die gewünschten Anforderungen (Integrität, Anonyme Daten ohne gewünschte Rückidentifizierung für Forschungszwecke) erfüllt, zu entwerfen, müssen einige Aspekte betrachtet werden. Zwei wichtige Grundlagen für die Komponente bilden die Stichwörter *Grundmenge* und *Kombination*. Aufgrund der zum Austausch herangezogenen Daten (Grundmenge) und deren Kombination muss die Komponente eine Entscheidungsgrundlage dafür liefern, ob diese Daten verschickt werden können ohne, dass eine ungewollte Rückidentifikation möglich ist. Da eine Rückidentifikation für Forschungszwecke nicht zwingend notwendig, aber durch Attacken möglich ist, müssen Mechanismen der *k*-Anonymisierung, Pseudonymisierung und Diversität eingesetzt werden. Ebenfalls die Wahrscheinlichkeit bezüglich gewisser Häufigkeiten von Charakteristika einer Krankheit (Alter, Geschlecht, Nationalität,…) bzw. das Auftreten der Krankheit selbst, müsste in einer Datenbank gespeichert und in regelmäßigen Abständen gepflegt werden. Dies bedeutet einen sehr hohen administrativen und analytischen Aufwand, der betrieben werden muss, um aktuelle Daten als Entscheidungsgrundlage zur Verfügung stellen zu können.

Ausgehend von dem Beispiel, dass eine Forschungsinstitution oder ein Mediziner Daten zu einer spezifischen Krankheit oder zu bestimmten Symptomen von einer anderen Einrichtung anfordert, soll die Komponente typische Charakteristika dieser Krankheit (Auftreten bei/ab einem speziellen

Alter, geschlechtsabhängigkeit, vererbbar, wo tritt die Krankheit, geographisch gesehen, häufig auf? usw.) prüfen und mit den angeforderten Daten vergleichen.
Fallen die angeforderten Daten in ein bestimmtes Häufigkeitsmuster dieser speziellen Krankheit, muss aufgrund der Auftrittswahrscheinlichkeit dieser Musters entschieden werden ob Daten verschickt werden können, und ob die Grundmenge ausreichend groß ist, um eindeutige Identifikationen ausschließen zu können.

Abhängig von den Charakteristika der Krankheit vergleicht die Komponente dann dynamisch die einzelnen Attribute (Geb.Datum, Geschlecht, Nationalität,…) und „verfremdet" diese dann mit Methoden der Generalisierung, Unterdrückung oder Verfremdung, wie sie in Kapitel 3.2 vorgestellt wurden. So kann z.B. das Geburtsjahr auf die ersten beiden Stellen abgeschnitten werden (19XX bzw. 20XX wenn man davon ausgeht, dass es eine Krankheit ist, die bei einer bestimmten Altersgruppe gehäuft auftritt).

Die Einführung von Ranges, in denen die angeforderten Daten liegen sollen, ist in solch einem Fall ebenfalls zielführend. Ist die Krankheit altersabhängig, bekommt der anfordernde Mediziner auf diese Art und Weise repräsentative Daten geliefert. Trotzdem muss dann der Rest des Datensatzes betrachtet werden, um auch hier eine Identifikation ausschließen zu können. Weiß man, dass die zu identifizierende Person ein spezielles, auffälliges Merkmal aufweist (Alkoholismus, Drogenkonsum, Medikamentenabhängigkeit), müssen auch hier wieder die angeforderten Daten, datensatzweise untereinander verglichen werden, und eventuelle herausragende Peaks geglättet bzw. Auffälligkeiten gelöscht werden, sodass diese nicht mit übertragen werden.

Beispiel:
Unter dem Vorwand die Wirksamkeit eines Medikamentes X untersuchen zu wollen, möchte ein Angreifer (in diesem Fall ein Mitarbeiter der Forschungseinrichtung) herausfinden ob sein Nachbar, der seit geraumer Zeit in Behandlung ist, wirklich drogenabhängig ist.

Aufgrund des Anwendungsgebietes des Medikamentes X wird entschieden wie groß die Grundmenge sein muss und welche Kriterien sie erfüllen soll.
Zum Beispiel muss ausgeschlossen werden, dass Medikament X nur bei sehr wenigen Patienten in der Forschungseinrichtung verwendet wird – hier muss eine entsprechend hohe Grundmenge vorhanden sein. Die Grundmenge ist das erste Kriterium auf welches, in Abhängigkeit der angeforderten

Daten, geachtet wird. Dann wird der Entscheidungsbaum weiter verzweigt und man betrachtet weitere Daten, die angefordert worden sind. Sind Auffälligkeiten in diesen Daten zu beobachten (herausragendes Alter, Geschlechtsgewichtung, Gewichtungen im Allgemeinen), müssen diese beschnitten bzw. nicht übertragen werden sodass ein Gleichgewicht entsteht welches keine Rückschlüsse aufgrund von Gewichtungen erlaubt.

5. Evaluierung ausgewählter Anonymisierungs- und Pseudonymisierungstechniken

Im diesem 5. Kapitel der Arbeit erfolgt eine Reflexion der vorgestellten Konzepte auf die aufbauend versucht wird einen Lösungsvorschlag für die Komponente zur Überprüfung des Anonymisierungs- und Pseudonymisierungsgrades der Daten, die ausgetauscht werden sollen, zu finden. Die Anforderungen, die an ein pseudonymisiertes Forschungsnetz gestellt werden, sind in [1] erläutert. Diese unterscheiden sich wiederum im Einsatzzweck. Hierbei kann man zwischen Speicherung der Daten zu klinischen Zwecken und Verwendung der Daten für wissenschaftliche Forschungszwecke unterscheiden.

In Kapitel 3 wurde erwähnt, dass für den Aufbau eines solchen Netzes feststehen muss in welchem Kontext die Daten verarbeitet und gespeichert werden sollen. Ist der Aufbau eines Forschungsnetzes mit Langzeitspeicherung der Daten gewünscht, ist eine Pseudonymisierung nicht zwingend notwendig, da hierfür die Rückidentifikation der Patienten (z.B. zwecks Verständigung, vergleiche Kapitel 4.3) nicht erfolgen muss. In solch einem Fall werden Daten anonymisiert gespeichert. Dieser Umstand, kann aufgrund der in Kapitel 3.2.1 erwähnten Angriffsszenarien dazu führen, dass Daten rückidentifiziert werden, und somit die Attacke auf einen solchen Datensatz verwertbare Ergebnisse liefert.

Kapitel 3 gibt eine Übersicht der Methoden für die *k*-Anonymisierung (Kapitel 3.3) und die Modelle der Pseudonymisierung (Kapitel 3.1). Im kommenden Abschnitt erfolgt eine Evaluierung der Methoden, welche für das Projekt SPICS Soul geeignet sind.

Bei einer Pseudonymisierung werden Attribute durch ein Pseudonym ersetzt, welches ohne Referenzliste nicht mehr entschlüsselt werden kann. Eine solche Liste birgt ein großes Risiko, da ein Angreifer mit ihrer Hilfe an sensible Patientendaten gelangen kann. Löscht man den direkten Patientenbezug bei einer Pseudonymisierung indem man die Referenzliste vernichtet, können weitere für Forschungszwecke relevante Daten (Alter, Geschlecht, ethnische Herkunft, im Allgemeinen primäre

Identifikationsmerkmale) verloren gehen oder nicht mehr zugeordnet werden. Da solche Attribute kontextabhängig sind, ist es besser alle Daten zur Verfügung zu haben und zum Zeitpunkt der Anfrage zu entscheiden ob diese weitergereicht werden dürfen oder nicht. Eine Anonymisierung kann erwünschte Daten erhalten, während sie Andere durch Methoden der Generalisierung verfremdet. Durch das Wegfallen der Notwendigkeit einer Rückidentifikation wird zumindest das Risiko, welches durch Speichern einer Referenzliste entsteht, verringert.

MinGen:

In Kapitel 3.3.1 wurde der MinGen - Algorithmus vorgestellt, der mit Hilfe von minimalen Veränderungen, eine gegebene Tabelle in *k*-anonyme Form bringt. Der Algorithmus war nicht nur sehr komplex, sondern auch ausgesprochen ineffizient, was sich in einer hohen Laufzeit (siehe Formel 3 bzw. Formel 4) niederschlug.

Data-Fly:

Data-Fly (Kapitel 3.3.2) überzeugte durch viele Parameter, die vom Benutzer eingestellt werden können. Auch die Kennzeichnung der Attribute, welche übertragen werden sollen, sowie eine Gewichtung ob eine Identifikation „sehr wahrscheinlich“ oder „sehr unwahrscheinlich“ sei, kann vom Datenhalter festgelegt werden. Hier sei erwähnt, dass detailliertes Fachwissen des Datenhalters notwendig ist, um eine solche Entscheidung treffen zu können. Auch ein Missbrauch dieser Möglichkeit ist erdenklich wenn der Datenhalter absichtlich Daten mit einer niedrigen Wahrscheinlichkeit markiert. Für SPICS Soul ist dieser Fall aber irrelevant da die letztliche Entscheidung der Arzt trifft. Data Fly trifft undurchdachte Entscheidungen und kann so zu einer zu starken Generalisierung führen. Aufgrund der besseren Laufzeit als MinGen ist er jedoch, zumindest in dieser Hinsicht, für den Einsatz in der Praxis besser geeignet.

***μ*-Argus:**

Ebenfalls Einsatz in der Praxis findet *μ-Argus* von Statistics Netherlands [9], welcher eine Kombination von Global Recoding und Local Suppression ist. Diese Kombination erlaubt einen minimalen Verlust hinsichtlich der Datenqualität, testet aber nicht alle möglichen Kombinationen was zu Ergebnissen führt, die den Anforderungen der *k*-Anonymität nicht genügen. Spielt die Laufzeit keine Rolle, kann man den Algorithmus um die Möglichkeit des Testens aller Kombinationen erweitern, damit man garantiert *k*-anonyme Ergebnisse bekommt.

INCOGNITO:

Im Fall von ICOGNITO (Kapitel 3.3.4) kam ein Breitensuchalgorithmus zum Einsatz, der die Patiententabelle in Domain Generalization Hierarchies unterteilte. Bei dieser Methode wird Schritt für Schritt aus Gruppen von sekundären Identifikationsmerkmalen, eine *k*-anonyme Variante der Tabelle erstellt. Eine (Superroots-ICOGNITO) der zwei Modifikationen des Algorithmus (Superroots- und Cube-INCOGNITO) konnte einen nachweislichen (Abbildung 16) Performanz-Zuwachs gegenüber der Basisvariante des ICOGNITO Algorithmus erzielen. Cube-INCOGNITO hingegen, wurde mit zunehmender Größe der Datenbank immer langsamer.

k-OPTIMIZE:

Einen völlig anderen Ansatz verfolgt die in Kapitel 3.3.5 vorgestellte Methode namens k-OPTIMIZE. Während bei den vorangegangenen Varianten Originaldaten als Input herangezogen wurden, startet k-OPTIMIZE mit einem voll generalisierten Datensatz und wandelt diesen dann Schritt für Schritt, durch Hinzufügen von zusätzlichen Informationen, in minimal *k*-anonyme Form um. Der Algorithmus setzt beim optimalen Datenmanagement an und garantiert somit immer *k*-anonyme Ergebnisse mit minimal gehaltener Generalisierung – einem großen Vorteil hinsichtlich der Datenqualität. Obwohl hier nicht das schnelle Finden einer Lösung, sondern eine optimale Lösung mit hoher Datenqualität das Ziel ist, führt k-OPTIMIZE bis zu drei mal schneller (auch für breite Spektren von *k* (5-1000)) zu einer Lösung als stochastische Methoden, weshalb hier eine Empfehlung für den Einsatz bei SPICS Soul ausgesprochen wird. Generell sei festzuhalten, dass es gilt einen Kompromiss zwischen Datenqualität und Rechenzeit zu finden. Hohe Datenqualität fordert in den meisten Fällen auch hohe Rechenzeit. Als Belohnung erhält man dafür wiederum wertvolle Daten, die einer geringeren Generalisierung unterzogen wurden.

Multidimensionale Anonymität:

Den Ansatz der multidimensionalen Anonymität verfolgen LeFevre, De Witt und Ramakrishnan ebenfalls mit ihrem Vorschlag welcher eine hohe Datenqualität mit geringen Generalisierungen liefert. Auch dieser Ansatz kann aufgrund der erzielten Ergebnisse für den Einsatz bei SPICS Soul empfohlen werden. Beide Methoden (k-OPTIMIZE und multidimensionale *k*-Anonymität) erfordern jedoch einen hohen Aufwand zur Vorbereitung der Daten aber liefern dafür Daten, die den Anforderungen der *k*-Anonymität genügen. Aufgrund der erwähnten Angriffsszenarien auf diese, ist es je-

doch ratsam die Möglichkeiten der *l*-Diversität (Kapitel 3.5) in Erwägung zu ziehen und auf die erhaltenen Daten anzuwenden.
Wie in (Abbildung 28) veranschaulicht wird, ist kaum mehr ein Unterschied hinsichtlich der Rechenzeit zwischen *k*-Anonymität und *l*-Diversität zu sehen. Gibt es keinen Grund für eine sekundenschnelle Berechnung eines Ergebnisses der Komponente (dürfen Daten veräußert werden – ja oder nein?), sprechen wir uns für eine Methode aus, die eine *k*-Anonymisierung mit hoher Datenqualität zum Ergebnis hat welche daraufhin mit Methoden der *l*-Diversität immun gegen Angriffe, gemacht wird. Ein solcher Datensatz genügt den Anforderungen der *l*-Diversität und kann weitergegeben werden.

5.1. Entwurf und Evaluierung der Komponente

Aufgrund der in Kapitel 4.4 beschriebenen Anforderungen wurde ein Entwurf für die Komponente erstellt. Eine Evaluierung der geeigneten Anonymisierungs-Algorithmen wurde in Kapitel 5 abgegeben. Als Grundlage des Entwurfs dienen 2 Datenbanken, die globale bzw. lokale Daten gespeichert haben. Die Komponente dient zur Entscheidungsfindung ob die abgefragten Daten weitergegeben werden dürfen oder nicht. Im Zuge dessen erfolgt eine bidirektionale Kommunikation zwischen der Komponente und den beiden Datenbanken. Abbildung 29 beschreibt einen Forschungsverbund, dessen Komponenten im Folgenden einzeln erläutert werden:

5.1.1. Globale Datenbank (Datenbank 1)

Datenbank 1 ist eine zentrale, relationale Datenbank, die globale medizinische Daten speichert und auf die alle im Verbund enthaltenen Forschungseinrichtungen (abfragende Stellen und Stellen, die Medikations- oder Patientendaten speichern bzw. verschicken) zugreifen können. Sie fungiert als statistischer Speicher und enthält im Idealfall fortwährend aktuelle Informationen zu Krankheiten und den in direkter Relation dazu stehenden Daten. Diese Daten können das Auftreten der Krankheit in Abhängigkeit von Alter, Geschlecht, Herkunftsland, Lebensweise oder die dafür vorgesehene Medikation, deren Erfolg und Nebenwirkungen, sein. Die Datenbank speichert somit alle zu einer Krankheit bekannten Fakten, um eine möglichst große und breit gefächerte Basis für die Gesamtpopulation zu schaffen. Dies ist im vorgestellten Modell eine Voraussetzung, um aussagekräftige Daten zu erhalten, um entscheiden zu können ob abgefragte Daten Seltenheitswert haben und somit ein Individuum aufgrund von Hintergrundwissen eindeutig identifiziert werden kann. Hier ist festzuhalten,

dass für die Pflege einer solchen umfassenden Datenbank großer Aufwand betrieben werden muss, da Daten aktuell zu halten sind und dafür in regelmäßigen Abständen ein Update notwendig ist. Wie die Pflege dieser Datenbank im Detail auszusehen hat, soll nicht Inhalt dieser Arbeit sein. Es sei dennoch erwähnt, dass die World Health Organization auf ihrer Website http://www.who.int [48] ein statistisches Informationssystem (WHOSIS – WHO Statistical Information System [49]) anbietet, welches Abfragen im Bereich der Gesundheitsstatistik für die momentan 193 WHO Mitgliedsstaaten erlaubt. Die Abfrage ist mit Hilfe eines Web-basierten Tools realisiert worden, welches eine Datenbank durchsucht. Diese enthält mehr als 70 Indikatoren (z.B. Sterberaten, Durchschnittsalter,...), welche in Abhängigkeit von demographischen Attributen gefiltert und abgefragt werden kann. Weiters erscheinen diese Daten jährlich in dem im Mai herausgegebenen World Health Statistics Report [50]. Hier sind nicht alle Daten, die für eine Datenbank, die ausreichend repräsentative Daten für die Prüfkomponente liefert, enthalten, jedoch bildet der World Health Statistic Report eine wichtige Grundlage aus der ebenfalls Daten für eine statistische Datenbank extrahiert werden können. Es ist sicherlich notwendig weitere medizinische Daten von Forschungseinrichtungen (weltweit), konkret in Bezug auf Medikation, in einer solchen Datenbank abzubilden.
Um in der globalen Datenbank gefundene Auffälligkeiten in Abhängigkeit der Abfrage auch auf lokale Einrichtungen und deren Patienten anwenden zu können, ist eine 2. Datenbank notwendig, mit Hilfe derer die Komponente abgefragte Datensätze mit globalen Daten und in weiterer Folge mit lokalen Daten der Einrichtung vergleicht.

5.1.2. Lokale Datenbank (SPICS-Soul) der jeweiligen Forschungseinrichtung (Datenbank 2)

Datenbank 2 ist ebenfalls eine relationale Datenbank, im konkreten Fall die Datenbank des Systems SPICS Soul, die Medikations- und Patientendaten der jeweiligen Forschungseinrichtung verwaltet. Diese Datenbank kann nur jeweils von der zugeordneten Forschungseinrichtung selbst verwaltet werden. Das heißt jede Einrichtung hat ihre eigene Datenbank, auf die andere Einrichtungen nicht zugreifen können. Somit ist festzuhalten, dass jede Einrichtung im Verbund Zugriff auf 2 Datenbanken hat: auf ihre eigene Datenbank, die Daten zu den behandelten Patienten speichert, und auf die zentrale, globale Datenbank, die globale Patientendaten von Einrichtungen auf der ganzen Welt verwaltet.
Diese 2 Datenbanken sind die Voraussetzung für die Arbeitsweise der Komponente welche im nächsten Abschnitt vorgestellt wird.

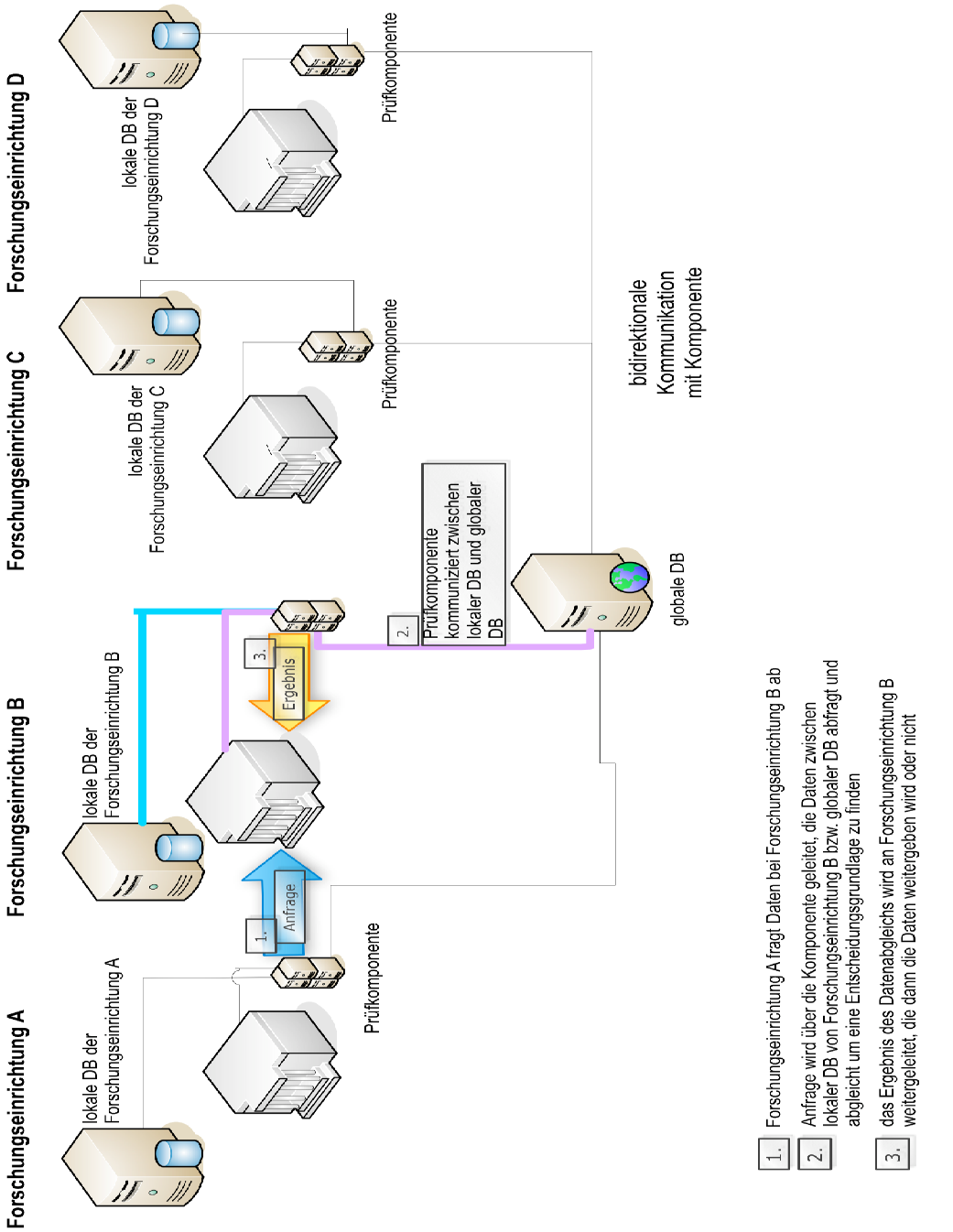

Abbildung 29: Schemenhafte Darstellung der Komponente im Forschungsverbund

5.2. Prüfkomponente

Voraussetzung für die Arbeit der Komponente sind die beiden in Kapitel 5.1.1 bzw. 5.1.2 erwähnten Datenbanken. Sie beinhalten die Daten, die für einen Vergleich der abgefragten Attribute mit lokalen und globalen Daten notwendig sind, und ermöglichen die weiterführende Entscheidungsfindung für die Erlaubnis der Weitergabe Dieser. Die Komponente geht wie folgt vor (Abbildung 30):
Forschungseinrichtung A fragt bei Forschungseinrichtung B Daten zu einem Medikament ab. Diese Daten werden als Paket in der Anfrage gebündelt und enthalten die Attribute Alter, Geschlecht, Gewicht, Medikationsdaten, Herkunftsland und Beruf. Ein solches Paket wäre eine plausible Anfrage wenn beispielsweise der Verdacht auf Symptome, die durch bestimmte Verhältnisse in der Berufswelt (Umwelteinflüsse, Stress, körperliche Arbeit) hervorgerufen werden, besteht. Die Komponente erhält dieses Paket als Input und schlüsselt es wieder in die einzelnen Attribute auf. In Abhängigkeit der einzelnen Attribute werden u.a. folgende Abfragen von Datenbank 1 erstellt:

- Welche Krankheiten sind in der abgefragten Berufsgruppe häufig?
- In Abhängigkeit der Medikation: welche Auffälligkeiten, (seltene) Attribute wie zum Beispiel äußerliche, optische Merkmale, gibt es hier?

Weitere plausible Abfragen sind natürlich möglich und sollen implementiert werden.

Attribute wie das Alter, Geschlecht oder Gewicht werden aus Performancegründen nicht explizit in der globalen Datenbank abgefragt sondern in Abhängigkeit der beiden oben erwähnten Attribute:

- Gibt es in der Berufsgruppe vorwiegend übergewichtige Personen, Männer oder Frauen, ethnische Gruppen?

Das Ergebnis könnte wie folgt aussehen: Die Medikation wird vorwiegend bei Patienten aus einer bestimmten Berufsgruppe angewandt, da diese in ihrem Beruf besonderen Verhältnissen ausgesetzt sind und öfters dieselben Symptome auftreten, wie zum Beispiel Rückenschmerzen oder Sehschwäche.

Die aufgrund dieser Abfragen gewonnen Ergebnisse aus Datenbank 1 werden im zweiten Schritt mit Datenbank 2 aller Forschungseinrichtungen verglichen. Sind dort Patienten in Behandlung, die Krite-

rien aus dem Ergebnis erfüllen, ist zu entscheiden welche Attribute weitergegeben werden. Ist die Herkunft bei einem einzelnen Patienten ein herausragendes Attribut, so ist auf die Weitergabe des Herkunftslandes zu verzichten. Genauso ist bei weiteren Attributen, die als einzigartig oder selten identifiziert wurden, vorzugehen. Dazu können Methoden der *k*-Anonymisierung verwendet werden (zb. Das Beschneiden von Informationen). Zum Schluss erfolgt der Schritt, welcher *l*-diverse Daten garantieren soll. Eine *k*-Gruppe ist *l*-divers, wenn das sensible Attribut der Gruppe zumindest *l*-verschiedene Ausprägungen aufweist (siehe Definition Kapitel 3.5) – somit erfolgt die Anwendung der *l*-Diversität wenn Daten, welche angefordert wurden aber ein potentielles Identifikationsrisiko darstellen, weitergereicht werden sollen. Reichen die vorhandenen Datensätze nicht dazu aus, kann man das Hinzufügen von Tupeln andenken. Hier muss aber wiederum beachtet werden, dass dies in einer Forschungsumgebung nicht sinnvoll ist, da Ergebnisse verfälscht werden können.

Aus datenschutzrechtlicher Sicht und basierend auf dem österreichischen Datenschutzgesetz in seiner gültigen Fassung ist für die Weitergabe der anonymisierten Daten keine Einwilligung notwendig, da in solch einem Fall alle primären Identifikationsmerkmale gelöscht werden. Da die vorgestellten Methoden jedoch einen relativ breiten Handlungsspielraum für das auswählen der Attribute in einem Datensatz erlauben, muss an dieser Stelle darauf geachtet werden, dass keine identifizierenden Attribute weitergegeben werden. Dies kann beispielsweise mit Hilfe eines 4 Augen-Prinzips oder einer entsprechenden Implementierung der Software, die keine Weitergabe solcher Daten erlaubt, realisiert werden.

Abbildung 30: Arbeitsweise der Prüfkomponente

6. Zusammenfassung und Ausblick

Die grundsätzliche Fragestellung unter welchen Bedingungen Daten weiter gegeben werden können lässt sich trotz sorgfältiger Recherche nicht eindeutig beantworten. Die Arbeit hat gezeigt, dass die Weitergabe von personenbezogenen Daten in der Medizin sensibel ist und es viele Möglichkeiten eines Angriffs auf eine Weitergabe geben kann.
Heute existieren mehrere Methoden und Konzepte für den Bereich der Anonymisierung und Pseudonymisierung, die möglichst viele Sicherheitsaspekte mit einbeziehen und auch in der Praxis im Einsatz sind.
Ausschlaggebend erscheint die Diversität der Daten, welche weitergegeben werden sollen. Doch Diversität ist ebenfalls nicht einfach zu erreichen. Für Diversität ist eine gewisse Grundmenge notwendig, welche garantiert, dass eine ausreichend große Anzahl an Daten vorhanden ist, aus denen man sich bedienen kann, und die es einem Angreifer erschweren mittels Homogenitäts- oder Hintergrundwissen-Attacken an Informationen zu gelangen.

Trotzdem bleibt hier noch Raum für weitere Verbesserungen. Viele der, in der Arbeit vorgestellten, *k*-Anonymisierungsmethoden sind nicht ausgereift und könnten in vielerlei Hinsicht performanter arbeiten bzw. eine bessere Datenqualität erzeugen als es momentan der Fall ist. In einigen Fällen werden nicht alle Kombinationen getestet wodurch keine Garantie für *k*-anonyme Daten gegeben werden kann. Auch der Einsatz der *l*-Diversität hält sich im medizinischen Bereich in Grenzen und existiert momentan nur in Form des Vorschlags von Machanavajjhala et.al. in [1]. Praktische Erfahrungswerte sind keine vorhanden, was jedoch für einen breiten kommerziellen Einsatz in der medizinischen Forschung wünschenswert wäre.

Schließlich ist der in der Arbeit vorgestellte Vorschlag für die Prüfkomponente jener, der eine Kombination verschiedener existierender Methoden der *k*-Anonymisierung und der *l*-Diversität, gepaart im Einklang mit gesetzlichen Grundlagen miteinander vereint. Über die Umsetzung und die Ergebnisse dieses Modells kann zum heutigen Zeitpunkt mangels Praxis noch kein Urteil gefällt werden.

Literaturverzeichnis

[1] K.Ranneberg, A.Pfitzmann, G.Müller, Universität Freiburg, TU Dresden, Sicherheit, insbesondere mehrseitige IT-Sicherheit, 1996

[2] IT-Sicherheitskriterien, Kriterien für die Bewertung der Sicherheit von Systemen der Informationstechnik, 1. Fassung vom 11.1.1989, Köln**,** Zentralstelle für Sicherheit in der Informationstechnik**,** Bundesanzeiger 1989

[3] V. L. Voydock, S.T. Kent, Security Mechanisms in High-Level Network Protocols, ACM Computing Surveys 15 (1983), No. 2, June 1983, 135-170

[4] A. Jautz, Analyse und Umsetzung von Methoden zur Anonymisierung und Pseudonymisierung personenbezogener, medizinischer Daten, Magisterarbeit, Medizinische Universität Wien, Institut für Medizinische Informations- und Auswertsysteme, Wien 2006, 6pp

[5] Datenschutzgesetz 2000 (DSG 2000), BGBl. I Nr. 165/1999
http://www.dsk.gv.at
Abgerufen: 04.01.2008, 10:21 CET

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, l-Diversity: Privacy beyond *k*-Anonymity, In ICDE-2006

[7] A. Pfitzmann, M. Hansen, Anonymi*t*y, Unlinkability, Undet*e*ctability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology, Version 0.31 vom 15.02.2008
Archives: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

[8] K.Pommerening, M. Reng, P. Debold, S. Semler, Pseudonymisierung in der medizinischen Forschung – das generische TMF-Datenschutzkonzept
http://www.egms.de/pdf/journals/mibe/2005-1/mibe000017.pdf
Abgerufen: 16.08.2008, 11:22 CET

[9] Homepage der STRING-Kommission des BMAGs, MAGDA-LENA Richtlinie Version 2.0
http://www.meduniwien.ac.at/msi/mias/STRING/Hauptteil.pdf
Abgerufen: 07.03.2009, 11:33 CET

[10] Homepage der STRING Kommission des BMAGs, Kurzinformation
http://www.meduniwien.ac.at/msi/mias/STRING/Kurzin.html
Abgerufen: 07.03.2009, 11:40 CET

[11] C.M. Reng, P. Debold, Ch. Specker, K. Pommerening, Generische Lösungen zum Datenschutz für die Forschungsnetze in der Medizin, In: Schriftenreihe der Telematikplattform für Medizinische Forschungsnetze, Medizinisch Wissenschaftliche Verlagsgesellschaft, Berlin, 1: Teil B, 2005, S.44 - 65

[12] Homepage des US Department of Health and Insurance Services
http://privacyruleandresearch.nih.gov/
Abgerufen: 08.03.2009, 17:10

[13] Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule, Department of Health and Human Services, USA
http://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf
Abgerufen: 08.03.2009, 17:10

[14] M. Castells, Der Aufstieg der Netzwerkgesellschaft, UTB Verlag, 2004, S.56

[15] Homepage der Sozialversicherungs-Chipkarten Betriebs- und Errichtungsgesellschaft m.b.H. – SVC ,
http://www.chipkarte.at/portal/index.html?ctrl:cmd=render&ctrl:window=ecardportal.channel_content.cmsWindow&p_menuid=67375&p_tabid=2
Abgerufen: 28.03.2009, 17:29

[16] R. Dahinden, Risiken im industriellen Umfeld – Aspekte einer ganzheitlichen, umweltorientierten Risikobeurteilung, Dissertation, Hochschule St. Gallen für Wirtschafts-, Rechts- und Sozialwissenschaften, 1991, S.116

[17] R. Neumann: Risiko Organisation - organisiertes Risiko: Beiträge zur integrativ-systemorientierten Verarbeitung selbsterzeugter Risikopotentiale in und von Organisationen, Frankfurt a. M. et al., Peter Lang Verlag, 1995, S.22

[18] E. Muschick, P.H. Müller, Entscheidungspraxis: Ziele, Verfahren, Konsequenzen, 1. Aufl., Springer Verlag, Berlin 1987, S.108

[19] K.Pommerening, Pseudonyme – ein Kompromiss zwischen Anonymisierung und Personenbezug, Institut für Medizinische Statistik und Dokumentation der Johannes-Gutenberg-Universität, 55101 Mainz

[20] Organisation for Economic Cooperation and Development, Glossary of statistical Terms, 2007, S.643

[21] Designing Privacy Enhancing Technologies, International Workshop on Design Issues in Anonymity and Unobservability, Berkeley, CA, USA, July 25–26, 2000 Proceedings, S. 1-9

[22] D. Chaum, Security without identification: transaction systems to make Big Brother obsolete, Communications of the ACM, 28(10), 1985

[23] D. Chaum, T.P. Pedersen, Wallet databases with observers (extended abstract), In Advances in Cryptology, CRYPTO '92, Springer Verlag, 1992, S.89-105

[24] I. B. Damgård, Payment systems and credential mechanisms with provable security against abuse by individuals (extended abstract), In Advances in Cryptology, CRYPTO '88, Springer Verlag, 1988, S. 328-335

[25] L. Chen, Access with pseudonyms, In Ed Dawson and Jovan Golic, editors, Cryptography: Policy and Algorithms, Lecture Notes in Computer Science No. 1029, Springer Verlag, 1995, S. 232-243

[26] K. Ulsenheimer, Leitlinien, Richtlinien, Standards – Risiko oder Chanche für Arzt und Patient, Der Anaesthesist, Ausgabe 47, Nummer 2, Springer Verlag, 1998

[27] H. Franzki, Haftungsvoraussetzungen bei fehlerhafter Heilbehandlung. In: Jost, Langkau (Hrsg) Leitlinien in der Chirurgie, Darmstadt 1997, S 31–35

[28] A.Meyerson R. Williams, General k-anonymization is hard, Carnegie Mellon School of Computer Science. 2003 (CMU-CS-03-113), Forschungsbericht
http://reports-archive.adm.cs.cmu.edu/anon/2003/CMU-CS-03-113.pdf
Abgerufen: 29.03.2009, 14:55 CET

[29] A. Meyerson, R. Williams: On the Complexity of Optimal k-anonymity, In: Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2004, S. 223-228

[30] L. Sweeney, Achieving k-Anonymity Privacy Protection using Generalization and Suppression, In: International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, S. 571-588

[31] L. Sweeney, Guaranteeing anonymity when sharing medical data, the Datafly System, J Am Med Inform Assoc 1997, S.51-55

[32] A.J.Hundepool, L.C.R.J. Willenborg, Statistics Netherlands, τ- and μ-ARGUS: Software for Statistical Disclosure Control, 1997, S. 142-149

[33] K. LeFevre, D.J. De Witt, R. Ramakrishnan, Incognito: Efficient Full-Domain K-Anonymity, In: Proceedings of the ACM SIG-MOD International Conference on Management of Data, 2005, S.49-60.
http://www.cse.iitb.ac.in/dbms/Data/Courses/CS632/Papers/incognito.pdf
Abgerufen: 21.08.2008, 09:55 CET

[34] S. Brüssow, Uninformierte Suchstrategien (Blinde Suche). Universität Potsdam, Institut für Linguistik / Computerlinguistik, Proseminar Künstliche Intelligenz 2004

[35] P. Samarati, Protecting respondants' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6), November/December 2001.

[36] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998

[37] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10(5), 2002, S. 557-570

[38] R.J. Bayardo, R. Agrawal, Data Privacy Trough Optimal k-Anonymization, In: Proceedings of the 21st IEEE International Conference on Data Engineering, 2005, S.217-228

[39] T.H. Cormen, C. Leiserson, R L. Rivest, C. Stein: Introduction to Algorithms, MIT Press, 2001, S. 370 ff

[40] T.H. Cormen, C. Leiserson, R L. Rivest, C. Stein: Introduction to Algorithms, MIT Press, 2001, S. 253 ff

[41] R.Symon, Searching trough systematic set enumeration, In: Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning, 1992, S.539-550

[42] The Open Group Base Specifications Issue 6, qsort, IEEE Std 1003.1, 2004 Edition,
http://www.opengroup.org/onlinepubs/009695399/functions/qsort.html
Abgerufen: 12.04.2009, 10:45 CET

[43] J. Hromkovič, Theoretische Informatik, Formale Sprachen, Berechenbarkeit, Komplexitätstheorie, Algorithmik, Kommunikation und Kryptographie, 3.Auflage, Teubner Verlag, 2007, S.285ff

[44] M. Mitchell, An Introduction to Genetic Algorithms, Edition: reprint, illustrated, 1998, S.27ff

[45] K. LeFevre, D.J. De Witt, R. Ramakrishnan, Multidimensional k-Anonymity, University of Wisconsin, 2005, Forschungsbericht 1521, http://www.cs.wisc.edu/techreports/2005/TR1521.pdf
Abgerufen: 19.08.2008, 12:25 CET

[46] K.R. Koch, Einführung in die Bayes Statistik, Kapitel 2.1.8: Bayes Theorem, Springer Verlag, 2002, S.13-17

[47] S. Qing, H. Imai, G. Wang, Information and communications security (2007), In: Proceedings of the 9th international conference, ICICS 2007, Zhengzhou, China, December 12-15, 2007, S.145-145

[48] Homepage der WHO (World Health Organization)
http://www.who.int/
Abgerufen: 22.02.2009, 11:11 CET

[49] Homepage des WHOSIS (WHO Statistical Information System)
http://www.who.int/whosis/en/
Abgerufen: 22.02.2009, 11:11 CET

[50] Homepage des World Health Statistic Reports der WHO
http://www.who.int/whosis/whostat/en/
Abgerufen: 22.02.2009, 11:11 CET

[51] J. Pflüger, Skriptum zur Vorlesung Datenschutz und Datensicherheit, Wintersemester 2001/2002, TU Wien

[52] D. Stelzer, Kritik des Sicherheitsbegriffs im IT-Sicherheitsrahmenkonzept, Datenschutz und Datensicherung DuD 14, 1990, Nr.10 Oktober 1990, S. 501-506

[53] Canadian System Security Centre, The Canadian Trusted Computer Product Evaluation Criteria, Version 2.0, Final Draft, 1990, Communication Security Establishment, Government of Canada

[54] Canadian System Security Centre, The Canadian Trusted Computer Product Evaluation Criteria, Version 3.0e, 1993, Communication Security Establishment, Government of Canada

[55] L. Sweeney, Uniqueness of simple demographics in the u.s. population, Technical Report, 2000, Carnegie Mellon University

[56] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A.Zhu, k-anonymity: Algorithms and hardness, Technical report, 2004, Stanford University

[57] N.R. Adam, J.C. Wortmann, Security Control Methods for statistical databases: A comparative study, ACM Comput. Surv., 1989, S.515-556

[58] G.T. Duncan, S.E. Feinberg, Obtaining information while preserving privacy: A markov perturbation for tabular data, In: Joint Statistical Meetings, 1997, Anaheim, CA

[59] U.C. Irvine, Machine Learning Repository
http://www.ics.uci.edu/mlearn/mlearnrepository.html
Abgerufen: 26.06.2009, 21:01 CET

[60] B. Schneier, Secrets & Lies – Digital Security in a Networked World, John Wiley and Sons, 2000, Preface

[61] M. Bishop, Computer Security: Art and Science, Addison-Wesley, 2002, Kapitel 1.1, S. 3

[62] A.G. De Waal, L.C.R.J. Willenborg, Global Recodings and Local Suppressions in Microdata Sets, 1995, Report, Voorburg: Statistics Netherlands

[63] S. Tiourine, Set Covering Models for Statistical Disclosure Control in Microdata, Paper, 3rd International Seminar on Statistical Confidentiality, 1996, Bled

[64] A.G. De Waal, L.C.R.J. Willenborg, Elements of Statistical Disclosure Control, Springer Verlag Lecture Notes in Statistics, 2000

[65] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, August 1994, In: Proceedings of the 20th international Conference of very large Databases

[66] Resolution A/RES/45/95, General Assembly od United Nations: “Guidelines for the Regulation of Computerized Data Files”, 1990

[67] Directive 2002/58/EC of the European Parliament on: “The proceedings of persona Data”, Juli 2002

[68] Directive 95/46/CE of the European Parliament and the Council of the European Union: “On the protection of Individuals”, 1995

[69] B. Fung, K. Wang, P.Yu, Top-down specialization for information and privacy preservation, In: Proceedings of the 21st International Conference on Data Engineering, April 2005

[70] K. Wang, P. Yu, S. Chakraborty, Bottom-up generalization: A data mining solution to to privacy protection, In: Proceedings of the 4th IEEE International Conference on Data Mining, 2004

[71] V. Iyengar, Transforming data to satisfy privacy constraints, In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004

[72] W. Winkler, Using simulated annealing for k-anonymity, Research Report 2002-07, US Census Bureau Statistical Research Division, 2002

[73] Homepage der österreichischen Bürgerkarte
http://www.buergerkarte.at
Abgerufen: 18.07.2009, 17:02 CET

[74] XML Definition der Personenbindung, Version 1.2.2 vom 14.02.2005, Stabstelle IKT Strategie des Bundes
http://www.buergerkarte.at/konzept/personenbindung/spezifikation/aktuell/Personenbindung-20050214.pdf
Abgerufen: 18.07.2009, 17:05 CET

[75] Homepage der Maestro Bankomatkarte
http://www.bankomatkarte.at
Abgerufen: 18.07.2009, 17:18 CET

[76] Homepage des zentralen Melderegisters (ZMR) – allgemeine Informationen
http://zmr.bmi..gv.at/pages/allgemein.htm
Abgerufen: 18.07.2009, 17:21 CET

[77] Homepage des österreichischen Bundesministeriums für Finanzen
http://www.bmf.gv.at
Abgerufen: 18.07.2009, 17:21 CET