PMV forschungsgruppe

versorgungsepidemiologie
qualitätssicherung
sekundärdatenanalyse

Ergebnisbericht und Manual
Pseudonymisierungsdienst »Sekundärdaten«

Implementierung eines Pseudonymisierungsdienstes mit Treuhänderstelle
und Erstellung einer Pseudonymisierungssoftware
unter besonderer Berücksichtigung der Anforderungen
bei der Pseudonymisierung von Gesundheits- und Sozialdaten
für die Sekundärdatenanalyse
- Darstellung der Organisationsstruktur,
Programmbeschreibung und Installationsanleitung -

Teilprojekt im Projekt DS 3.1 »Pseudonymisierungsdienst«
der Arbeitsgruppe »Datenschutz und Datensicherheit«
der Telematikplattform für medizinische Forschungsnetze (TMF)

PMV
forschungsgruppe

Autor und Korrespondenzadresse
Peter Ihle
PMV forschungsgruppe, Universitätsklinikum zu Köln
Herderstraße 52-54, 50931 Köln
Tel.: (+49/0)-221-478-6548
Fax.: (+49/0)-221-478-6766
Peter.Ihle@medizin.uni-koeln.de

Danksagung
Für die konzeptionelle Beratung bedanke ich mich bei Frau Schmidtmann, Institut für medizinische Epidemiologie und Informatik der Universität Mainz, recht herzlich. Bedanken möchte ich mich auch bei Frau Lautz aus dem gleichen Institut für die gewissenhafte Durchführung der Pseudonymisierung.

Auch den Datenschutzbeauftragten der AOK Hessen und der KV Hessen, hier vor allem Herrn Dr. Staerk, der das Projekt durch seine Begeisterung für technische Details vorangetrieben hat, gilt unser Dank. Nicht zuletzt bedanke ich mich auch bei Frau Dr. Wellbrock und Herrn Wehrmann vom Hessischen Datenschutzbeauftragten, die das Projekt »Versichertenstichprobe AOK Hessen/KV Hessen« in rechtlicher und datenschutzrechtlicher Hinsicht beraten haben.

Version 1.01

Kurzfassung

Projektziel

Versichertenbezogene Gesundheits- und Sozialdaten - insbesondere Daten der Gesetzlichen Krankenversicherung (GKV) - werden zunehmend für andere als zu den primären Erhebungszwecken genutzt, vor allem für Fragen der Epidemiologie, Versorgungsforschung, Gesundheitsberichterstattung oder Gesundheitsökonomie (Sekundärdatenanalyse). Auch im seit 1. Jan. 2004 geltenden Gesundheitsmodernisierungsgesetz (GMG) ist eine Sammlung dieser Daten gesetzlich verankert worden. Das es sich bei den Gesundheitsdaten um besondere Daten im datenschutzrechtlichen Sinn handelt, ist neben methodischer und medizinischer Expertise auch ein datenschutzrechtlich geprüftes Datenschutzkonzept mit Pseudonymierungsdienst ein notwendiger Bestandteil für die Erhebung und Nutzung dieser Daten. Dieser Pseudonymisierungsdienst wurde in einem Projekt der Arbeitsgruppe "Datenschutz und Datensicherheit" der Telematikplattform für medizinische Forschungsnetze (TMF) konzipiert und mit Einrichtung eines Datentreuhänders implementiert. Hierbei ging es sowohl um die organisatorische Lösung als auch um die Umsetzung in eine Pseudonymisierungssoftware.

Rahmenbedingungen

Im Rahmen eines Kooperationsprojekts der AOK Hessen, der KV Hessen, des Hessischen Sozialministeriums und der PMV forschungsgruppe (Universitätsklinikum zu Köln) wurde eine Zufallsstichprobe aus AOK-Versicherten (Versichertenstichprobe AOK Hessen/KV Hessen) und in Zusammenarbeit mit dem Kompetenznetz »Maligne Lymphome« eine erkrankungsspezifische Stichprobe von Lymphompatienten erhoben. Nach Selektion der versichertenbezogenen Daten in der AOK Hessen und der KV Hessen werden diese in einer Vertrauensstelle, hier das Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI) der Universität Mainz, pseudonymisiert und an die Auswertungsstelle jeweils per CD und PGP-transportverschlüsselt übermittelt.

Ergebnisse

Die in der Vertrauensstelle notwendige Pseudonymisierungssoftware wurde vom Autor dieses Berichts in der Programmiersprache "C" für das Betriebssystem Windows 98SE/ME geschrieben. Der verwendete symmetrische Verschlüsselungsalgorithmus »Blowfish« steht lizenzfrei zur Verfügung. Das Programm verarbeitet die einer CD entsprechende Datenmenge, in der Regel bis zu mehrere Mio. Datensätze, innerhalb einer halben Stunde. Neben versichertenbezogenen Kennzeichen werden auch institutionsbezogene Angaben wie Kassenarztnummern oder Institutionskennzeichen stationärer Einrichtungen pseudonymisiert.

Schlussfolgerung und Ausblick

Mit dem Pseudonymisierungsdienst für GKV-Daten konnte ein sowohl in organisatorischer als auch technischer Hinsicht datenschutzrechtlich unbedenkliches und geprüftes Verfahren implementiert werden. Der Dienst kann auch in anderen Umgebungen, z. B. beim Export klinischer Daten in wissenschaftliche Datenbanken eingesetzt werden, d. h. immer in Bereichen, in denen Daten schnell, sicher und kostengünstig pseudonymisiert werden müssen. Der Dienst kann von den Mitgliedern der Telematikplattform zu den jeweils gültigen Konditionen genutzt werden.

Inhaltsverzeichnis

Verzeichnis der Abbildungen und Tabellen

Abbildungen

Tabellen

01 Einleitung

Daten der Gesetzlichen Krankenversicherung werden verstärkt für Forschungszwecke erhoben und genutzt, insbesondere für Fragen der Versorgungsforschung, der Gesundheitsberichterstattung und nicht zuletzt für gesundheitsökonomische Analysen. Vorteile dieser Sekundärdatenanalyse sind vor allem der Personenbezug auch über längere Beobachtungszeiträume, die Abbildung aller Sektoren (ambulant, stationär, Pflege) als auch das Fehlen von Erinnerungs- oder Interviewerbias sowie fehlende Selektion durch Verweigerung. Zudem stehen diese Daten relativ kostengünstig zu Verfügung. Nicht zuletzt lassen sich durch die sekundäre Analyse auch Führungsdaten für die Organe der Selbstverwaltung und für die Gesundheitspolitik erzeugen. Diese Vorteile sind auch vom Gesetzgeber erkannt worden, der in dem seit 1. Jan. 2004 geltenden Gesetz zur Modernisierung der Gesetzlichen Krankenversicherung (GMG) die Erhebung und Nutzung dieser Daten verankert hat (SGB V §303a-f).

Notwendiger Bestandteil für die Erhebung und Nutzung dieser Daten ist neben methodischer und medizinischer Expertise auch ein datenschutzrechtlich geprüftes Datenschutzkonzept.

Ein solches Konzept wurde im Rahmen eines Methodenforschungsprojektes im Auftrag des Statistischen Bundesamtes (Ihle et al. 1999) konzipiert und in der »Versichertenstichprobe AOK Hessen/KV Hessen« mit den zuständigen Datenschutzbeauftragten diskutiert und realisiert. Zentraler Bestandteil dieses Konzepts ist der Datentreuhänder für die Pseudonymisierung der Patientendaten. Unter ökonomischen Gesichtspunkten was es notwendig, die beim Treuhänder anfallenden Routinearbeiten zu automatisieren und den Personalaufwand auf ein Minimum zu reduzieren. Gleichzeitig mussten die Standards zur Pseudonymisierung von Massendaten abgebildet werden. Die Software sollte skalierbar sein, um sie an unterschiedliche Umgebungen spezifisch anpassen zu können. Die entwickelte Software erfüllt diese Anforderungen.

Im Folgenden wird der Pseudonymisierungsdienst »Sekundärdaten« dargestellt und diskutiert. Anschließend wird die für die Pseudonymisierung notwendige Software und die Installation in einer Standardumgebung beschrieben.

02 Anforderung

2.1 Pflichtenheft

Aufgrund der existierenden organisatorischen Rahmenbedingungen sowie der datenschutzrechtlichen Aspekte wurde zu Beginn des Projektes ein Pflichtenheft erstellt, in dem die Anforderungen an den Pseudonymisierungsdienst sowie an die zu erstellende Software fixiert wurden:

- Schnelles Offline-Verfahren zur Verarbeitung großer Datenmengen.
- Ressourcen schonendes Verfahren in der Treuhänderstelle (Niedriger Arbeitsaufwand, kurze Laufzeiten).
- Standard-Hardware (PC nach Industriestandard) mit Standardbetriebssystem (Intelrechner mit MS-WINDOWS).
- Die Daten sollten im ASCII-Format geliefert und ebenso nach Pseudonymisierung weiterverarbeitet werden.
- Die Informationen sollten zeilenweise (in relationaler Form) verarbeitet werden. Die hierdurch entstehende Doppelung von Informationen steht der Vorteil einer leichten Verarbeitung und der Existenz von Export und Importroutinen der gängigsten Datenbanksysteme gegenüber.
- Pseudonymisierung von n Datenfeldern, da neben den Versichertennummern auch Leistungserbringerkennzeichen wie Kassenarztnummern, Institutionskennzeichen von stationären Einrichtungen und nichtärztlichen Leistungserbringern pseudonymisiert werden. Trennung von Personen- und Institutionskennzeichen und (sensiblem) Datenbereich.
- Sichtung des Datenteils nur durch datenliefernde und auswertende Stelle. d. h. die Treuhänderstelle erhält nur Einblick in den für die Pseudonymisierung notwendigen Teil , der sensible medizinische Datenbereich bleibt kryptographisch verschlüsselt [Ihle et al.: (2001): Die Vertrauensstelle im Rahmen der Sekundärdatenforschung – Lösungsansätze zum Problem der Datenkonzentration. Gesundheitswesen Sonderheft 1:6-12.].
- Programm sollte parametrisierbar sein, um unterschiedliche Pseudonymisierungsprofile zu pseudonymisieren
- Mehrere Projekte sollen nebeneinander bedient werden können. Hierfür ist ein Authentifizierungsverfahren zu implementieren.
- In einem LOG-File sollen die Arbeitsschritte dokumentiert werden.
- Das Programm sollte transportverschlüsselte Daten entschlüsseln und die pseudonymisierten Daten wieder transportverschlüsseln können.
- Es sollten starke, geprüfte und möglichst lizenzfreie Algorithmen zum Einsatz kommen
- Die Pseudonymisierungssoftware soll auch den datenliefernden Stellen zur Verfügung stehen, damit dort mit Testpseudonymisierungen die korrekte Vorbereitung der Daten für die Pseudonymisierung geprüft werden kann.

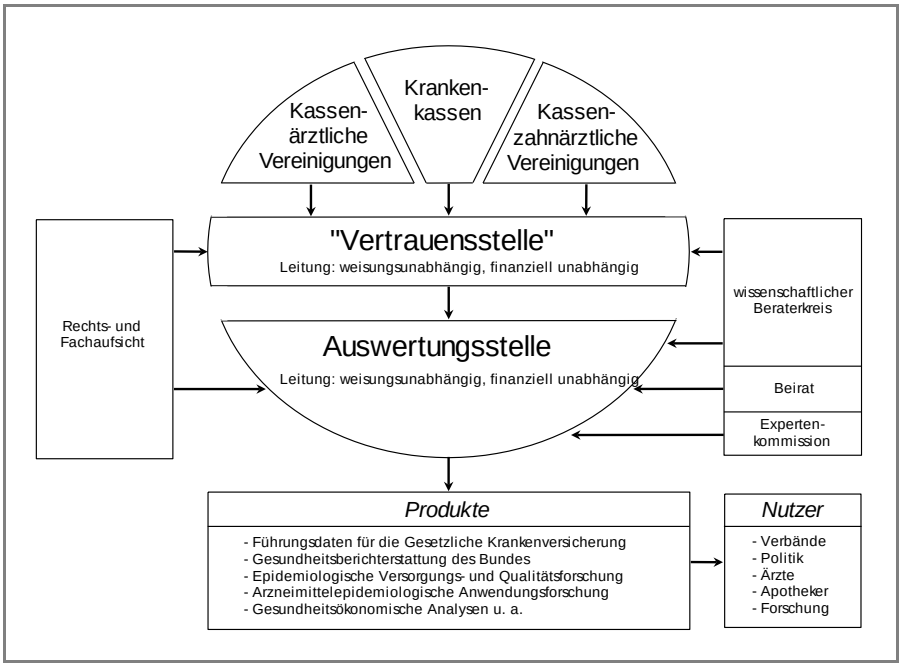
Gleichzeitig wurde in einem sehr frühen Zeitpunkt festgelegt, dass der Datenabgleich bzw. das Matchingverfahren (Erkennen von Homonymen und Synonymen) nicht Aufgabe der Treuhänderstelle sein sollte, sondern noch im Datenerhebungskontext - also noch in der datenerhebenden bzw. daten-

liefernden Stelle durchgeführt werden. Dieses Vorgehen deckt sich auch mit dem Vorgehen innerhalb der generischen Datenschutzkonzepte der TMF.

Die Datentreuhänderstelle muss organisatorisch, räumlich getrennt und weisungsungebunden gegenüber anderen in das Datenschutzkonzept eingebundenen Stellen sein.

Bei der Implementierung des Datentreuhänders wurde inhaltlich Bezug genommen auf das Methodenforschungsprojekt »Versichertenstichprobe aus der Gesetzlichen Krankenversicherung«, welches von der Forschungsgruppe Primärmedizinische Versorgung (PMV) im Auftrag des Statistischen Bundesamtes durchgeführt wurde [Ihle et. al: (1999): GKV-Versichertenstichprobe. Wirtschaft und Statistik 9:742-749].

Abb. 1 **Organisationsstruktur einer Versichertenstichprobe aus der Gesetzlichen Krankenversicherung**



nach Ihle, Peter; Köster, Ingrid; Schubert, Ingrid; Ferber, Liselotte von; Ferber, Christian von (1999): GKV-Versichertenstichprobe. Wirtschaft und Statistik 9:742-749.

2.2 Aufgaben des Datentreuhänders

Zentrale Aufgabe der Datentreuhänderstelle ist die Pseudonymisierung von Patientendaten.

Im Rahmen dieser Aufgabe müssen auch logistische Aufgaben übernommen werden. So werden die Daten der datenliefernden Stelle auf CD (PGP-transportverschlüsselt) an die Datentreuhänderstelle verschickt. Nach erfolgter Pseudonymisierung wird diese CD - sozusagen als Quittung - an die einsendende Stelle zurückgeschickt. Die pseudonymisierten Daten werden auf CD

(ebenfalls PGP-transportverschlüsselt) an die Auswertungsstelle verschickt. Damit verbleiben nach erfolgreicher Pseudonymisierung keine Daten in der Datentreuhänderstelle.

Der Datentreuhänder muss lediglich den zur Pseudonymisierung notwendigen Schlüssel verwahren. Nur er hat Kenntnis von diesem Schlüssel und teilt dieses Geheimnis mit keiner anderen Institution. Da dieser Schlüssel über die Laufzeit eines Projekts , z. T. über mehrere Jahre, vorhanden sein muss, wird empfohlen, mehrere Sicherungskopien dieses Schlüssel zu generieren. Diese können zum einen lokal verwahrt werden, um bei Diskettenfehler schnell eine Sicherungskopie zur Hand zu haben, oder auch räumlich getrennt von der Datentreuhandstelle z. B. im Tresor einer Bank, so dass auch bei Zerstörung des Originalschlüssels durch Gewalteinwirkung (z. B. durch Gebäudebrand) eine Sicherungskopie vorhanden ist.

Der Datentreuhänder benötigt im Rahmen seiner Aufgaben keine Einsicht in den medizinischen Datenteil, diese werden lediglich chiffriert von der datenliefernden Stelle an die Auswertungsstelle durchgereicht. Er führt insbesondere auch keine Plausibilitätskontrollen der Daten durch.

2.3 Beschlagnahmesicherheit

Bei der Nutzung von medizinischen Daten für die Forschung muss die Beschlagnahmesicherheit der Daten diskutiert werden.

Dieser Punkt ist im Pseudonymisierungsdienst aus verschiedenen Gründen ohne Relevanz. Zum einen werden keine Daten dauerhaft beim Datentreuhänder gespeichert, da die gelieferten Daten wieder an die einsendende Stelle zurückgeschickt werden. Aber auch zum Zeitpunkt der Pseudonymisierung sind keine sensiblen Daten vorhanden. Der Datentreuhänder hat zu keinem Zeitpunkt Einsicht in den medizinischen Datenteil. Zudem sind die Originalkennzeichen von Patient und Leistungserbringer bereits durch nichtsprechende, nicht rückrechenbare Identifikationsnummern ersetzt. Eine Nutzung dieser Daten für Dritte ohne weitere Angaben ist damit nicht erkennbar.

03 Programmbeschreibung

3.1 Softwareentwicklung

Da die Anforderungen an die zu erstellende Software umfassend und detailliert in einem Pflichtenheft (s. o.) fixiert waren, konnte die Programmierung sehr zielorientiert durchgeführt werden. Die Entwicklungszeit konnte vor allem durch Einbindung von vorhandenen Softwaremodulen, die innerhalb der PMV forschungsgruppe bei der Erhebung von Sekundärdaten erstellt worden waren, mit einem Mannmonat sehr kurz gehalten werden.

Das Programm wurde in der Programmiersprache C (Borland C Version 3.1) unter Windows 98 SE geschrieben. Das ablauffähige Programm wurde unter Windows 98 SE und Windows ME getestet und lauffähig installiert.

Als Algorithmus wurde Blowfish, entwickelt von Bruce Schneier, eingesetzt. Es handelt sich hierbei um ein Blockverfahren, welches lizenzfrei zur Verfügung steht. Die Schlüssellänge von bis zu 56 bit ist lang genug, um Bruteforce-Angriffen standzuhalten. Bisher sind keine erfolgreichen Angriffsversuche gegen dieses Verfahren bekannt (Schneier, B.: Angewandte Kryptographie. Protokolle, Algorithmen und Sourcecode in C. Addison Wesley, Deutschland. 1996).

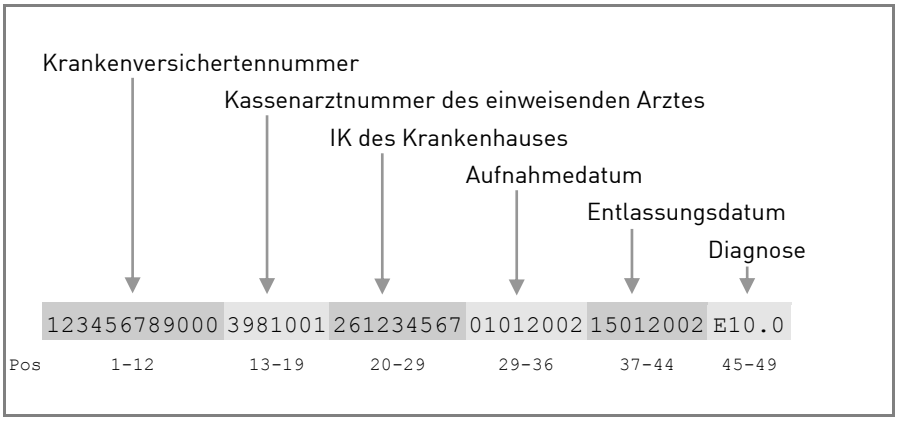
ASCII-Format

Für die Realisierung wurde vereinbart, kein spezifisches Dateiformat zu wählen. Die Daten liegen als reine Textdateien im ASCII-Format mit abschließendem Zeilenende (hexadezimal `0x0D` und `0x0A`) vor. Die einzelnen Daten befinden sich an festen Spaltenpositionen mit fester Feldlänge, üblicherweise ohne spezifische Spaltentrennzeichen. Die Satzlänge kann allerdings variabel sein, so dass im jeweils letzten Feld des Datensatzes unterschiedlich lange Einträge, z. B. Klartextdiagnosen möglich sind, die bei fester Satzlänge mit der entsprechenden Anzahl Leerzeichen aufgefüllt werden müssten.

3.2 Beispieldatensatz

Im Folgenden wird ein typischer Datensatz beschrieben, der im weiteren Verlauf beispielhaft verschlüsselt werden soll. Es handelt sich hierbei um einen Datensatz, der einen stationären Aufenthalt eines Versicherten dokumentiert. Angegeben neben der Versichertennummer, die Kassenarztnummer des einweisenden Arztes, das Institutionskennzeichen des behandelnden Krankenhauses, Aufnahme- und Entlassungsdatum, und die primäre Entlassungsdiagnose.

Beispieldatensatz für die Pseudonymisierung



Die Originalkennzeichen von Patient (PAT) und Leistungserbringer (LE = Arzt und Krankenhaus) werden in einem ersten Schritt durch nichtsprechende Identifikationsnummern - hier jeweils von der Länge 8 (Pnnnnnnn=Patient, Annnnnnn= Arzt, Knnnnnnn=Krankenhaus) ersetzt. Dadurch werden die Daten zum einen schon zu einem sehr frühen Zeitpunkt der Datenerhebung depersonalisiert, zum anderen können dabei notwendige Matchingverfahren durchgeführt werden. So werden beispielsweise unterschiedliche Kennzeichen wie Rentenversichertennummer und Krankenversichertennummer oder auch unterschiedliche Krankenversichertennummern eines Patienten zu einer einzigen eindeutigen Patientenidentifikationsnummer zusammengeführt. Dieser Testdatensatz befindet sich als Datei `TEST1.P01` auf der Installations-CD.

Ersetzen der Originalkennzeichen durch Identifikationsnummern

```
P0000341 A0001234 K0030422 01012002 15012002 E10.0
Pos  1-8      9-16     17-24     25-32     33-40   41-45
```

Die medizinischen Daten (MDAT), in diesem Fall werden Aufnahme- und Entlassungsdatum sowie die Diagnose in einem weiteren Arbeitsschritt kryptographisch chiffriert. Durch den jeweils gewählten Algorithmus kommt es hierdurch i. d. R. zu einer Veränderung der Satzlänge:

Chiffrieren der medizinischen Daten

```
P0003412 A0001234 K0030422 ?aj&%kBb$l-9+;JKkx*%t/Hj8%
Pos  1-8      9-16     17-24                25-50
```

Dieser Datensatz wird von der Datentreuhänderstelle pseudonymisiert. Durch die Art der Datenaufbereitung ist sichergestellt, dass der Datentreuhänder nur diejenigen Informationen erhält, die er für die Erfüllung seiner Aufgaben - die Pseudonymisierung - tatsächlich benötigt. Er erhält keine Originalkennzeichen der Patienten und Leistungserbringer, sondern nur die nicht sprechenden Identifikationsnummern, aus denen er das jeweilige Pseudonym generieren

kann. Er erhält weiterhin keinen Einblick in die medizinischen Daten; diese werden verschlüsselt »durchgereicht« und erst von der Auswertungsstelle wieder dechiffriert.

Um die Pseudonymisierung durchführen zu können, benötigt der Datentreuhänder noch Angaben darüber, an welchen Stellen des Datensatzes sich die jeweiligen Identifikationsnummern befinden, bzw. an welchen Positionen sich Daten befinden, die nicht pseudonymisiert werden müssen. Im gezeigten Beispieldatensatz muss der Datentreuhänderwissen, dass an Position 1 eine 8-stellige Patientenidentifikationsnummer steht, an Position 9 die 8-stellige Identifikationsnummer eines Kassenarztes und an Position 17 die 8-stellige Identifikationsnummer eines Krankenhauses. Ab Position 25 bis zum Ende des Datensatzes befinden sich die verschlüsselten medizinischen Daten, die unverändert bleiben.

Mit Hilfe dieser Angaben kann der Datensatz pseudonymisiert werden. Die jeweiligen 8-stelligen Identifikationsnummern werden in diesem Fall jeweils durch ein 10-stelliges Pseudonym ersetzt; damit ändert sich auch die Satzlänge. Die verschlüsselten medizinischen Daten werden unverändert angehängt. Nach erfolgter Pseudonymisierung sieht der Datensatz wie folgt aus:

Pseudonymisierung

```
ak/eP7%c+9 7&ghZ:jl$p a*od/$adV1 ?aj&%kBb§l-9+;JKkx*%t/Hj8%
Pos     1-10        11-20       21-30                31-56
```

In der Auswertungsstelle können die medizinischen Daten wieder entschlüsselt werden (auch hierbei kommt es wieder zu einer Änderung der Satzlänge). Die jeweiligen Angaben von Patient und Leistungserbringern liegen jetzt als Pseudonym vor, die medizinischen Daten wieder im Originaltext:

Dechiffrieren der medizinischen Daten

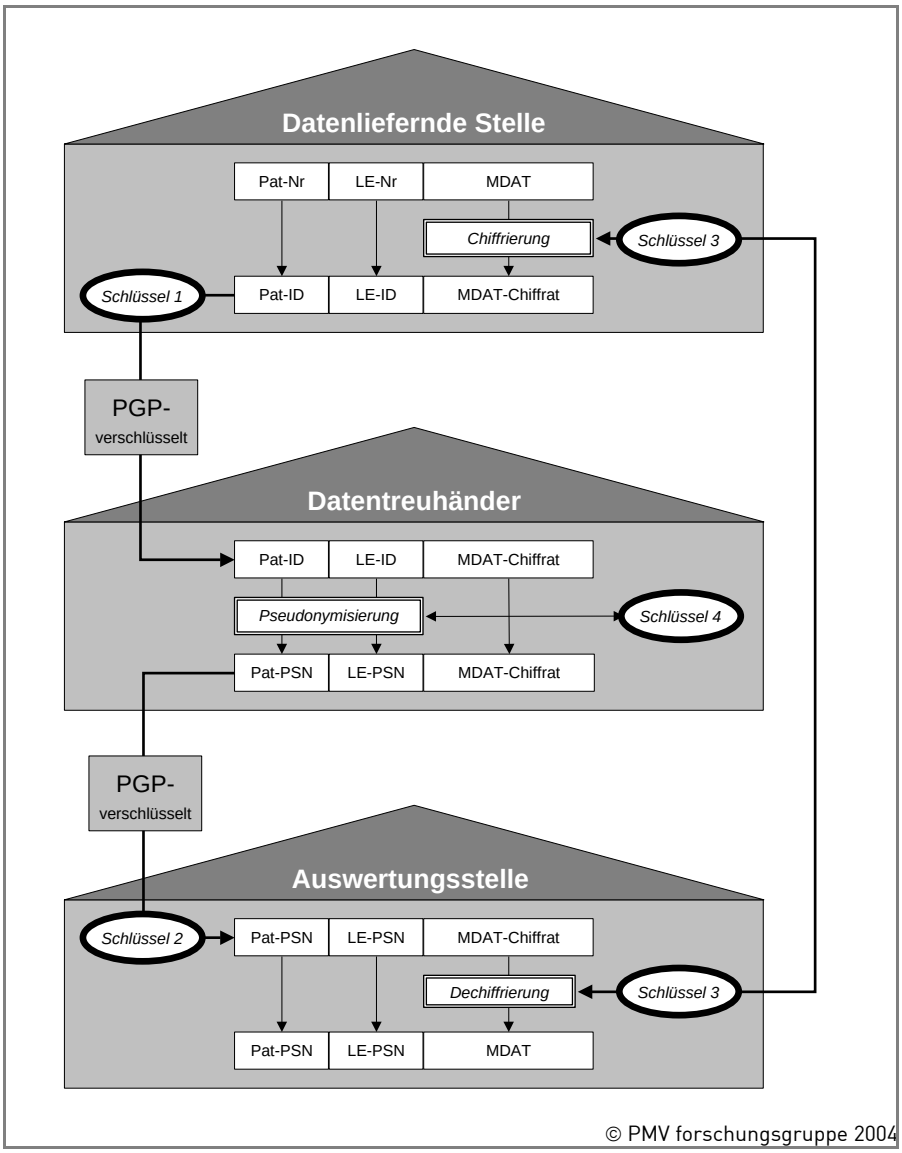
```
ak/eP7%c+9 7&ghZ:jl$p a*od/$adV1 01012002 15012002 E10.0
Pos     1-10        11-20       21-30      31-38     39-46   47-52
```

Auf dem Transportweg von der datenerhebenden Stelle zum Datentreuhänder und von dort zur Auswertungsstelle werden die Daten jeweils mit PGP verschlüsselt.

3.3 Datenfluss im Pseudonymisierungsdienst

In der nachfolgenden Abbildung werden die oben beschriebenen Einzelschritte zusammenfassend schematisch dargestellt.

Abb. 2 **Datenfluss im Pseudonymisierungsdienst**



Pat-Nr = Versichertennummer oder Patientennummer, **LE-Nr** = Leistungserbringernummer, z. B. die Kassenarztnummer oder das Institutionskennzeichen von Krankenhäusern., **PSN** = Pseudonym, **MDAT** = Medizinische Daten (Diagnosen, Leistungen, etc.).

3.4 Austausch von Passwörtern

Im Rahmen des Pseudonymisierungsdienstes werden Daten verschlüsselt und entschlüsselt. Die hierfür notwendigen Schlüssel müssen zu Beginn des Projekts ausgetauscht werden (vgl. hierzu auch Abb. 2). Dies erfolgte im laufenden Projekt durch persönlichen Kontakt der Mitarbeiter aus den jeweiligen Organisationen, kann aber auch durch entsprechend sichere Verbindungen, z. B. mit PGP-verschlüsselter Mail erfolgen.

Tab. 1 **Eingesetzte Schlüssel im Pseudonymisierungsdienst »Sekundärdaten«**

Schlüssel	Institutionen	Funktion
1	Datenliefernde Stelle + Datentreuhänder	PGP-Transport verschlüsselung
2	Datentreuhänder + Auswertungsstelle	PGP-Transport verschlüsselung
3	Datenliefernde Stelle + Auswertungsstelle	Chiffrierung der Medizinischen Daten mit Blowfish
4	Nur Datentreuhänder	Pseudonymisierung mit Blowfish

3.5 Formatdatei

Die Aufgabe der zu erstellenden Software ist nun, die oben beschriebenen Pseudonymisierungsschritte durchzuführen. Hierbei sollten durch die Datentreuhänderstelle möglichst wenig Eingaben zu tätigen sein, so dass die Pseudonymisierung mehr oder weniger automatisch ablaufen kann. Die hierfür notwendigen Ablaufparameter werden zusammen mit den zu pseudonymisierenden Daten von der datenliefernden Stelle in einer Formatdatei mitgeteilt, so. z. B. die Angaben, an welchen Stellen sich die zu pseudonymisierenden Datenteile befinden.

Die Formatdatei muss der 8.3-Konvention von MS-DOS entsprechen. Als Erweiterung ist »`PSD`« zwingend vorgeschrieben. Der Dateiname ergibt sich aus der Endung der zugehörigen Datendatei. Wenn sich die zu pseudonymisierenden Datensätze in der Datei »`TEST.P01`« befinden, muss die zugehörige Formatdatei »`P01.PSD`«heißen.

Im Folgenden ist die Formatdatei für den o. g. Beispieldatensatz gezeigt. Jeder Block wird durch das Schlüsselwort »`DATA`« eingeleitet und durch das Schlüsselwort »`END`« angeschlossen. Für den Beispieldatensatz ergeben sich damit 4 Blöcke: Pat-ID, Arzt-ID, Krankenhaus-ID und MDAT. Mit »`POS`« ist der Beginn und mit »`LEN`« die Länge der Angabe im Datensatz bezeichnet, auf das Schlüsselwort »`ENCODECODE`« folgt die jeweilige Kategorie. Mit dem Schlüsselwort »`NONE`« wird dem Programm mitgeteilt, dass keine Pseudonymisierung erfolgen soll. Üblicherweise wird die Länge des MDAT-Blockes mit der maximalen Datensatzlänge von 1024 angegeben. Das Programm liest den Datensatz aber nur bis zum jeweiligen Zeilenende.

Datei»P01.PSD«

```
...

DATA
  POS=1
  LEN=8
  ENCODECODE=PAT
END

DATA
  POS=9
  LEN=8
  ENCODECODE=ARZT
END

DATA
  POS=17
  LEN=8
  ENCODECODE=KH
END

DATA
  POS=24
  LEN=1024
  ENCODECODE=NONE
END
```

3.6 Schlüssel

Das Pseudonymisierungsprogramm liest aus der Formatdatei Position und Länge der zu pseudonymisierenden Identifikationsnummern. Ebenso erhält es die Angabe der Art des jeweiligen Kennzeichens, z. B. ob es sich um einen Pat-ID oder eine Arzt-ID handelt. Diese Angabe findet sich - wie oben beschrieben - als Eintrag unter dem Schlüsselwort `ENCODECODE`

In einer Projektdatei muss sich zu jedem ENCODECODE-Eintrag ein entsprechender Beschreibung enthalten sein. Dem Programm wird dadurch mitgeteilt, welcher Pseudonymisierungs-Algorithmus anzuwenden ist und welcher Schlüssel benutzt werden soll. Damit ist es möglich, dass für jede Art der Identifikationsnummer unterschiedliche Algorithmen bzw. Algorithmusversionen und/oder Schlüssel genutzt werden können.

Die jeweiligen Angaben finden sich in der Projektdatei. Zu jedem `ENCODECODE`-Eintrag findet sich ein Block, der von dem Schlüsselwort `ENCODE` eingeleitet und wie gewohnt mit `END` abgeschlossen wird. Dazwischen findet sich ein Eintrag `CODE`, dem der jeweilige Eintrag der `ENCODECODE`-Eintrages identisch sein muss. Nach dem Schlüsselwort `KEYCODE` steht der Name des anzuwendenden Schlüssels. Das Schlüsselwort `PROC` wird gefolgt von dem jeweiligen Algorithmus.

3.7 Projektdatei

Nachfolgend findet sich die Projektdatei zu dem Beispieldatensatz. Zu jedem Eintrag des Schlüsselworts ENCODECODE, in dem gezeigten Fall»PAT«, »ARZT« und »KH« findet sich ein entsprechender Block. Als Schlüsselname (KEYCODE) wird durchgehend KEY1 verwendet, als Algorithmus (PROC) das einzige bisher implementierte Verfahren (BLOWFISH_V1).

Projektdatei »PROJEKT1.PSP«

```
...

ENCODE
  CODE=PAT
  KEYCODE=KEY1
  PROC=BLOWFISH_V1
END

ENCODE
  CODE=ARZT
  KEYCODE=KEY1
  PROC=BLOWFISH_V1
END

ENCODE
  CODE=KH
  KEYCODE=KEY1
  PROC=BLOWFISH_V1
END
```

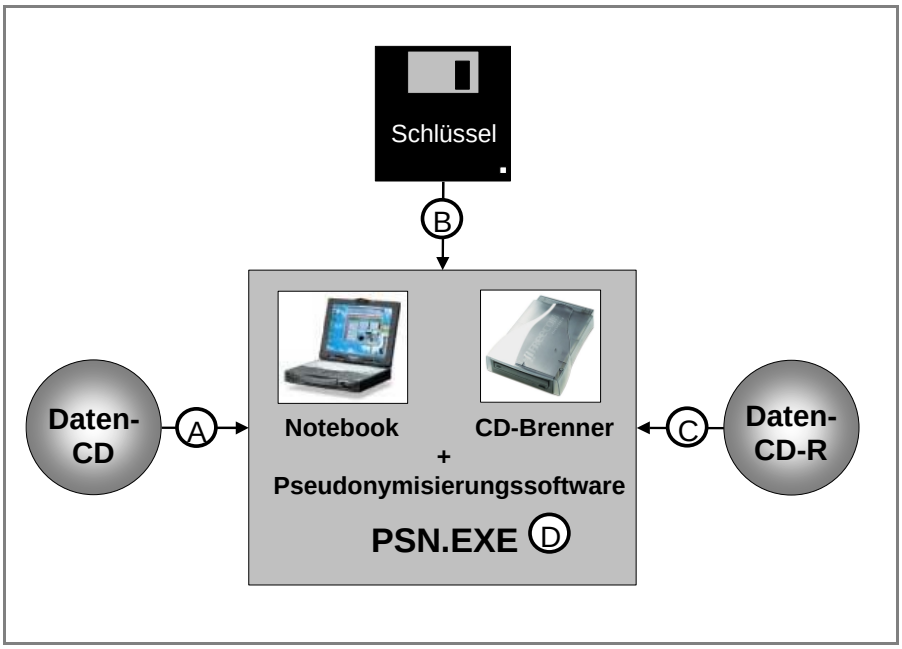
Der mit KEY1 bezeichnete Schlüsselname verweist auf eine Datei mit gleichem Namen, in dem sich der zu verwendende Schlüssel befindet. Als Dateierweiterung ist KEY zwingend vorgeschrieben. Die Schlüsseldatei im gezeigten Beispiel hat damit den Dateinamen KEY1.KEY. In ihr befindet sich ein 56-bit langer Schlüssel für den Blowfish-Algorithmus. Der Schlüssel seinerseits ist mit Blowfish verschlüsselt gespeichert.

3.8 Vorbereiten und Starten des Programms

Für einen Pseudonymisierungsvorgang wird der PC entsprechend den Einstellungen der Konfigurationsdatei bestückt (siehe hierzu auch Abb. 3).

In das CD-ROM-Laufwerk D: wird die CD mit den Originaldaten eingelegt (Schritt A), in den CD-Brenner kommt eine direkt beschreibbare CD (für die Vorbereitung siehe entsprechende Handbücher) (Schritt B), und in das Diskettenlaufwerk wird die Diskette mit dem Schlüssel eingelegt (Schritt C). Anschließend wird durch Eingabe von PSN an der Eingabeaufforderung das Programm gestartet (Schritt D).

Abb. 3 **Vorbereitung des PCs für de Pseudonymisierung**

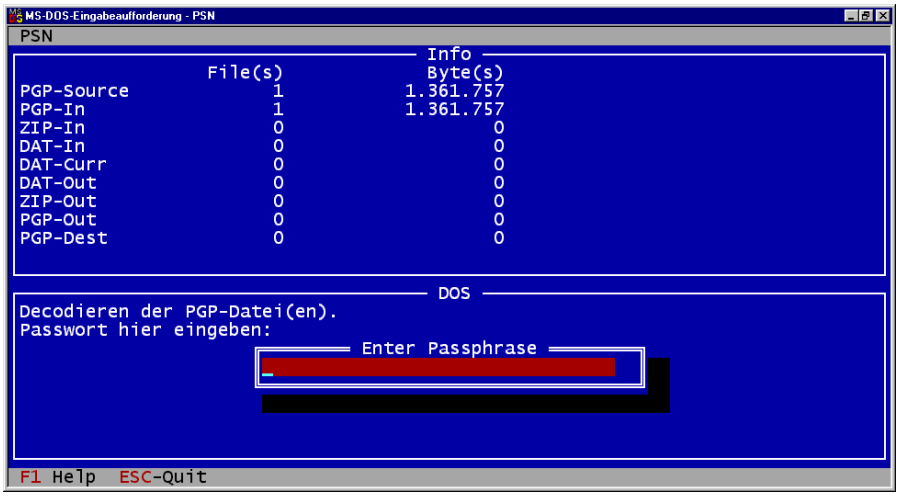


Nachdem das Programm gestartet wurde, fragt es nach kurzer Zeit drei Passwörter in der folgenden Reihenfolge ab (siehe auch Abb. 4):

1. Passwort für das Decodieren der von der datenliefernden Stelle gelieferten PGP-Dateien
2. Passwort für das Codieren der pseudonymisierten Daten mittels PGP
3. Passwort zum Decodieren der verwendeten Schlüsseldatei

Jedes Passwort muss jeweils zweimal hintereinander identisch eingegeben werden. Schlägt die Eingabe fehl, wird das Passwort erneut zweimal abgefragt.

Abb. 4 **Eingabemaske für die Passwörter**

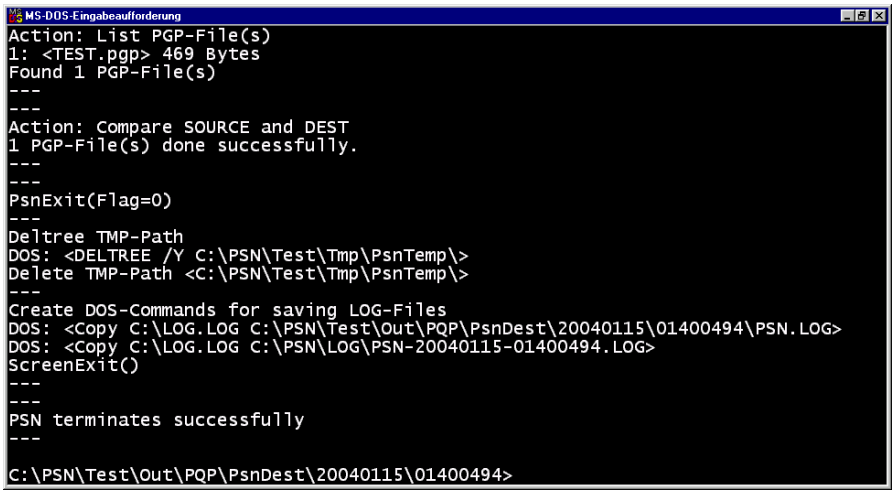


Nun läuft die Pseudonymisierung selbstständig ohne weitere Benutzereingaben statt.

Die Verarbeitungsgeschwindigkeit hängt von der technischen Ausstattung des PCs und dem Umfang der zu pseudonymisierenden Daten ab. Bei den von uns durchgeführten Pseudonymisierungen war eine CD mit ca. 300 MB gezippten Daten innerhalb von 15 bis 30Minuten verarbeitet.

Die pseudonymisierten Daten erhalten die gleichen Dateinamen wie die Originaldateien und werden in identischer Struktur zusammen mit den Formatdateien gezippt und in einer PGP-verschlüsselten Datei abgelegt. Die Original-CD und die CD mit den pseudonymisierten Daten haben dieselbe Datenstruktur. Nach erfolgreicher Pseudonymisierung meldet sich das Programm mit dem untenstehenden Bildschirminhalt zurück.

Abb. 5 **Abschlussbildschirm nach erfolgreicher Pseudonymisierung**



Anschließend wird die Original-CD entnommen und an die datenliefernde Stelle zurückgeschickt, die CD mit den pseudonymisierten Daten wird an die Auswertungsstelle geschickt. Die Diskette mit den Schlüsseln wir den jeweiligen Sicherheitsstandards entsprechend, z. B. im Datentresor, verwahrt. Alle während der Pseudonymisierung angelegten Arbeitsdateien werden nach erfolgter Pseudonymisierung automatisch gelöscht. Lediglich der LOG-File und eine Kopie der erzeugten PGP-Datei befinden sich noch im Arbeitsverzeichnis (`TMP`) im Unterverzeichnis gebildet aus dem aktuellen Zeitstempel (Datum und Uhrzeit mit Angabe von Sekunden und Hunderstelsekunde) z. B. `\TMP\20040115\09013253\`, also am 15. Jan. 2004 um 9:01 h 32,53 sec. Diese Dateien können nach Rückmeldung der Auswertungsstelle über die korrekte Verarbeitung der Daten manuell gelöscht werden.

3.9 Ablaufschema

Entsprechend den beschriebenen Einzelschritten läuft das Programm nach folgendem Schema ab:

Tab. 2 Ablaufschema einer Pseudonymisierung beim Datentreuhänder

Benutzereingaben	A1 Einlegen der CD mit den zu pseudonymisierenden Daten A2 Einlegen der beschreibbaren CD für die pseudonymisierten Daten A3 Einlegen der Diskette mit dem Schlüssel A4 Mit PSN an der Eingabeaufforderung das Programm starten A5 Eingabe der drei benötigten Passwörter für das Entschlüsseln der PGP-Dateien, dem Entschlüsseln der Schlüsseldateien, und für das PGP-Verschlüsseln der pseudonymisierten Dateien.
Ablauf des Pseudonymisierungs-programms ohne weitere Benutzereingaben	B1 Lesen der PSN.CFG-Datei mit den Pfadangaben für die Original-CD, für die pseudonymisierten Daten und die Schlüsseldatei(en). B2 Liste aller PGP-Dateien (1 bis n) anlegen und sukzessive verarbeiten
	C1. Entschlüsseln der n-ten PGP-transport verschlüsselten Datei C2. Entpacken der komprimierten ZIP-Datei C3 Liste aller Datendateien (1 bis m) anlegen und sukzessive verarbeiten
	D1 Zu der m-ten Datendatei die zugehörige Formatdatei (*.PSD) lesen D2. Die darin enthaltene Projektkurzbezeichnung und Projektmitglied lesen, Liste der zu pseudonymisierenden Arten von Identifikationsnummern (ID) mit Position, Länge und Schlüsselart lesen D3 In der passenden Projekdatei (*.PSP) lesen und prüfen, ob das Projektmitglied verzeichnet ist (Authentifizierung). D4 In der passenden Projekdatei (*.PSP) prüfen, ob alle ID-Arten gültig sind, ob der angegebene Algorithmus implementiert ist und ob der zugehörige Schlüssel als Datei mit der Endung `KEY` vorhanden. D5 Schlüsseldatei lesen und entschlüsseln D6 Pseudonymisierung der m-ten Datendatei durchführen D7 Wenn nötig, Schritte D1 bis D6 wiederholen
	C4 Komprimieren der pseudonymisierten Datei(en) zusammen mit der/den zugehörigen Formatdatei(en) in eine ZIP-Datei C5 Verschlüsselung der gezippten Datei mit PGP C6 Wenn nötig, Schritte C1 bis C5 wiederholen
	B3 Abspeichern der LOG-Datei im Verzeichnis \PSN\LOG B4 Meldung, dass die Pseudonymisierung erfolgreich war
Benutzereingriff	A6 Meldung durch Benutzer überprüfen A7 Entnehmen der Original-CD, der CD mit den pseudonymisierten Daten und der Diskette mit dem Schlüssel/den Schlüsseln

4.1 Installation

Die Installation wird im Folgenden für das Betriebssystem »MS-Windows 98 SE« beschrieben. Die Installation erfolgt ausschließlich mit MS-DOS Befehlen. Hierbei werden Kenntnisse der DOS-Umgebung und seiner Befehle vorausgesetzt. Eine Installationsroutine ist für das nächste Update geplant.

Hierzu muss zunächst in die Eingabeaufforderung gewechselt werden. Anschließend wird ein Verzeichnis angelegt und die Dateien von der Installations-CD hineinkopiert.

DOS-Befehl	Bemerkung
`C:`	
`CD\`	
`MD Psn`	
`CD Psn`	
`XCOPY D:\INSTALL\PSN\*.* /S/E/V`	Hierbei wird vorausgesetzt, dass sich die Installations-CD in Laufwerk D: befindet und das Installationsverzeichnis \INSTALL\PSN ist. Hiermit werden alle benötigten Programme in die jeweiligen Verzeichnisse kopiert.

Neben dem Programm PSN.EXE werden hierbei auch das benötigte Komprimierungsprogramm (`PkZip25.EXE` = Pkzip Version 2.5) in das Unterverzeichnis `PSN\EXE` kopiert. Die Programmnamen dürfen nicht umbenannt werden. Die Konfigurationsdatei `PSN.CFG` wird in das Verzeichnis `\PSN` kopiert, eine Projektdatei `PROJEKT1.PSP` in das Verzeichnis `\PSN\PSP`. Das jeweils gültige Manual (dieses Dokument) wird in das Verzeichnis `\PSN\DOC` kopiert.

Des weiteren werden noch zwei Umgebungsvariablen benötigt. Damit ist die Installation beendet.

Umgebungsvariablen

DOS-Befehl	Bemerkung
`PATH=%PATH%;C:\PSN\EXE`	Hiermit wird das Programmverzeichnis der Umgebungsvariable PATH hinzugefügt, so dass das Programm aus jedem Verzeichnis heraus durch Eingabe von PSN ausgeführt werden kann.
`SET PSNPATH=C:\PSN`	In der Umgebungsvariable `PSNPATH` (ohne schließenden Backslash) wird das Verzeichnis eingetragen, in dem die Konfigurationsdatei `PSN.CFG` erwartet wird.

PGP 6.5.8

Das lizenzfreie Verschlüsselungsprogramm PGP muss zusätzlich installiert werden. In der von uns durchgeführten Standardumgebung wurde die Version 6.5.8 installiert, genutzt wird die Kommandozeilenversion (zur Installation siehe entsprechende Manuale). In der Umgebungsvariable PATH muss ein entsprechender Eintrag auf das PGP-Verzeichnis enthalten sein.

4.2 Konfiguration

Nach erfolgter Installation muss der Rechner konfiguriert werden. Die Konfiguration bezieht sich auf zwei Bereiche: zum einen die Dateiorganisation für das Pseudonymisierungssoftware, zum anderen die Konfiguration des Projektes.

4.2.1 Konfigurationsdatei

Die Angaben zur Dateiorganisation werden in der Konfigurationsdatei `PSN.CFG` gespeichert. Diese *muss* sich im Verzeichnis befinden auf welche die Umgebungsvariable `PSNPATH` verweist:

- *Schlüsselverzeichnis* (`KEY`), hier sucht das Programm nach der Datei mit dem Schlüssel/den Schlüsseln für die Pseudonymisierung.
- *Eingabeverzeichnis* (`SOURCE`): hier erwartet das Programm die PGP-verschlüsselte Originaldaten.
- *Im temporären Arbeitsverzeichnis* (`TMP`) werden die benötigten Dateien zwischengespeichert und anschließend wieder gelöscht.
- In das *Ausgabeverzeichnis* (`DEST`) wird die PGP-verschlüsselte Datei mit den pseudonymisierten Daten gespeichert.

Im untenstehenden Kasten ist eine Konfigurationsdatei für eine Standardumgebung zu sehen. Der Schlüssel (`KEY`)befindet sich auf einer Diskette und liegt im Laufwerk»A:«. Die transportverschlüsselt(en) Eingabedatei(en) (`SOURCE`) lieg bzw. liegt/liegen im Rootverzeichnis einer CD im Laufwerk »D:«, das Ausgabeverzeichnis für die pseudonymisierten Daten (`DEST`) ist eine direktbeschreibbare CD im Brenner mit der Laufwerksbezeichnung »E:«. Schließlich wird noch ein temporäres Arbeitsverzeichnis (`TMP`) benötigt; dies ist

standardmäßig das Verzeichnis C:\TMP\. Bei der Eingabe der Verzeichnisnamen ist darauf zu achten, dass immer ein abschließender Backslash (\) angegeben ist.

Konfigurationsdatei »PSN.CFG«

```
PATH
  KEY      = A:\
  SOURCE   = D:\
  TMP      = C:\TMP\
  DEST     = E:\
END
```

Aufbau der Konfigurationsdateien

Die im Pseudonymisierungsdienst verwendeten Dateien haben alle den gleichen Aufbau. Ein Informationsblock wird jeweils durch ein Schlüsselwort (hier `PATH`) eingeleitet und mit `END` abgeschlossen. Dazwischen befinden sich beliebig viele Variablen, die jeweils mit dem Variablennamen und einem Gleichheitszeichen (z. B. `KEY`) eingeleitet werden, gefolgt von dem zugeordneten Wert. Der Wert kann auch Leerzeichen und wird mit dem Zeilenende abgeschlossen, Leerzeichen vor dem Zeilenende werden abgeschnitten. Die Einrückung ist beliebig und dient der besseren Lesbarkeit der Datei.

4.2.2 Projektkonfiguration

Ein neues Projekt benötigt eine Kurzbezeichnung von maximal 8 Buchstaben (nach der `DOS`-Konvention). Alle weiteren Angaben werden in einer Datei gespeichert, die den Namen der Projektkurzbezeichnung und die Endung `PSP` trägt, z. B. `PROJEKT1.PSP`.

Diese Datei enthält zum einen die Langfassung der Projektbezeichnung (jeweils mit dem Schlüsselwort `NAME` eingeleitet), ein Verzeichnis aller Kooperationspartner (`MEMBER`), die ihrerseits wieder eine Kurzform (`CODE`) und n Klartexteinträge (`NAME`) erhalten können. Ebenso enthalten sind alle zulässigen Kodierungsvorgänge (`ENCODE`) mit Kurznamen (`NAME`) und dem zu verwendenden Algorithmus (`PROC`).

Projektdatei »PROJEKT1.PSP«

```
PROJECT
  NAME=Projektbezeichnung
  NAME=Laufzeit 1. Jan. 2004 bis 30. Juni 05
  NAME=Bewilligungsnummer A1B2C3/D4
END

MEMBER
  CODE=ABC
  NAME=Forschungsinstitut ABC
  NAME=Universität Neuberg
  NAME=Universitätsstraße 1
  NAME=12345 Neuberg
END

MEMBER
  CODE=XYZ
  NAME=Forschungsinstitut XYZ
  NAME=Vordere Gasse 13
  NAME=65432 Grossstadt
END

ENCODE
  NAME=PAT_ID
  PROC=MD5_BLOWFISH_V1
END

ENCODE
  NAME=ARZT_ID
  PROC=MD5_BLOWFISH_V1
END

ENCODE
  NAME=KH_ID
  PROC=MD5_BLOWFISH_V1
END
```

Nur die beiden Dateien PSN.CFG und die für das Projekt gültige Projektdatei, in dem aufgezeigten Beispiel PROJEKT1.PSP sind auf dem Pseudonymisierungs-PC anzulegen.

4.3 Daten und Formatdatei

Von der datenliefernden Stelle werden auf CD die zu pseudonymisierenden Daten geliefert. Diese sind wie oben beschrieben zeilenweise (relational) aufgebaut, wobei die Informationen spaltenweise angeordnet sind.

Im vorderen Teil befinden sich normalerweise die zu pseudonymisierenden Identifikationsnummern, im hinteren Teil die chiffrierten medizinischen Daten (MDAT) .

Zu jeder Datendatei, in unserem Fall soll diese `TEST1.P01` heißen, existiert eine zugehörige Formatdatei, die als Dateinamen die Endung der Datendatei

besitzen muss und als Endung PSD, im gezeigten Fall also P01.PSD. Es können mehrere Datendateien mit gleicher Dateiendung existieren, die alle auf die gleiche Formatdatei verweisen. Es können aber auch verschiedene Datendateien geliefert werden, wobei zu jeder Datei eine zugehörige Formatdatei existieren muss, ansonsten wird das Programm mit einer Fehlermeldung beendet. Datendateien und Formatdateien werden gemeinsam komprimiert (mit PkZip 2.5) . In einer ZIP-Datei befinden sich 1 bis n Datendateien und 1 bis m Formatdateien. Jede ZIP-Datei wird mit PGP transportverschlüsselt, wobei hier die "Conventional Encryption" gewählt wird.

Diese PGP-Datei stellt den Ausgangspunkt für einen Pseudonymisierungsvorgang dar. Auf der CD können mehrere PGP-Dateien gespeichert sein, wobei jede für sich den oben beschriebenen Konventionen entsprechen muss: In einer PGP-Datei befindet sich genau eine ZIP-Datei, in dieser 1 bis n Datendateien und zu jeder Datendatei eine zugehörige Formatdatei mit der Endung PSD.

Jede Formatdatei hat den unten gezeigten Aufbau. In Block PROJEKT ist der Kurzname des Projekts (CODE) verzeichnet, im Block MEMBER eine gültige Mitgliedskurzbezeichnung (CODE). In den DATA-Blöcken sind die jeweiligen Spalten mit Beginn (POS) und Länge (LEN) sowie der jeweiligen ID-Art (ENCODECODE) enthalten.

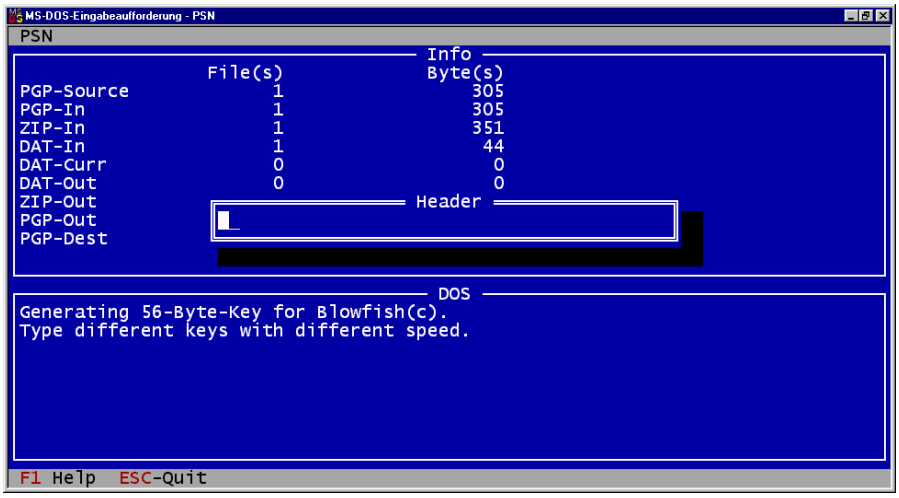
Datei»P01.PSD«

```
PROJECT
  CODE=PROJEKT1
END

MEMBER
  CODE=ABC
END

DATA
  POS=2
  LEN=7
  ENCODECODE=PAT
END

DATA
  POS=10
  LEN=7
  ENCODECODE=ARZT
END

DATA
  POS=18
  LEN=7
  ENCODECODE=KH
END

DATA
  POS=25
  LEN=1024
  ENCODECODE=NONE
END
```

Die Anzahl und Größe der zu liefernden PGP-Dateien sowie der in diesen enthaltenen Dateien sind nur durch das jeweilige Betriebssystem sowie der verwendeten Kapazität des Speichermediums (CD, DVD, ZIP-Laufwerk, MO-Laufwerke, externe Festplatten) beschränkt. Der Zugriff erfolgt über den Verzeichniseintrag der Konfigurationsdatei (`PSN.CFG`), in dem das Programm nach den entsprechenden PGP-Dateien sucht.

4.4 Schlüsselerzeugung

Die Generierung eines 56-bit-Schlüssels für `Blowfish` erfolgt für jedes Projekt bei der ersten Pseudonymisierung, es wird immer ein Schlüssel mit der für `Blowfish` zulässigen Maximallänge erzeugt. Wenn kein Schlüssel im Verzeichnis KEY gefunden wurde, wechselt das Programm automatisch in Schlüsselerzeugungsroutine. Der Benutzer wird nun aufgefordert, 56 Tastatureingaben zu tätigen. Wird die selbe Taste zweimal hintereinander gedrückt, so wird diese Eingabe abgelehnt. In die Generierung des Schlüssels gehen die gedrückten Tastencodes, sowie die zwischen zwei Tastatureingaben verstrichene Zeit ein.


MS-DOS-Eingabeaufforderung - PSN
PSN
Info
File(s) Byte(s)
PGP-Source 1 305
PGP-In 1 305
ZIP-In 1 351
DAT-In 1 44
DAT-Curr 0 0
DAT-Out 0 0
ZIP-Out
PGP-Out
PGP-Dest
Header
DOS
Generating 56-Byte-Key for Blowfish(c).
Type different keys with different speed.
F1 Help ESC-Quit