



# Testbericht PID-Generator

Dipl.-Tech. Math. Markus Schmidberger

## 1. Einleitung

Im Rahmen des Kompetenznetzes Leukämie [1] und in Zusammenarbeit mit der AG-Studienzentralen hat das Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE [2], LMU München) den PID-Generator installiert und getestet.

Dieser Bericht stellt für die Mitglieder des Kompetenznetzes eine Zusammenfassung der Funktionalität des PID-Generator, der zugehörigen rechtlichen Grundlage und des Testergebnisses dar.

## 2. Produktbeschreibung

Der PID-Generator ist eine Software, die dazu dient, eine definierte Population mit pseudonymen (nichtsprechenden) Identifikatoren zu versehen. Dazu wird jeweils ein Satz von personenidentifizierenden Daten (IDAT) in einen **PersonenID**entifikator (PID) transformiert. Berücksichtigt wird dabei, dass die personenidentifizierenden Merkmale fehlerbehaftet und zeitlich veränderlich sein können.

Der PID-Generator wurde von der Uni Mainz in einem TMF-Projekt entwickelt und steht allen Mitgliedsnetzwerken der TMF [3] zur Verfügung.

### Datenbank

Der Kern des PID-Generators ist eine Datenbank, die die bisher eingegebenen Fälle – die personenidentifizierenden Daten im Klartext oder in umkehrbar oder unumkehrbar verschlüsselter Form - zusammen mit dem jeweiligen PID speichert. Wird ein neuer Fall eingegeben, wird durch einen Abgleichs- (Match-) Algorithmus [8],[10] festgestellt, ob dieser Fall schon erfasst ist und der bereits gespeicherte PID auszugeben ist oder ob ein neuer PID erzeugt werden muss. [6]

### Betriebsmodi

Der PID-Generator ist als Web-Service konzipiert, d.h. die Eingabe erfolgt über ein einfaches Web-Formular. Darüber hinaus sind ein interaktiver Konsolenbetrieb sowie ein Batch-Betrieb möglich. Für administrative Datenpflege, etwa bei nachträglichem Erkennen eines Homonyms, ist ein Direktzugriff auf die Datenbank zu nutzen. [6]

### PID-Erzeugung

Der PID wird als kryptographisch verschlüsselte laufende Nummer erzeugt und mit Prüfzeichen versehen. [6][7]

## 3. Rechtliche Grundlage - Datenschutzkonzept

Der PID-Generator ist ein Teil des „Generischen Konzepts für den Datenschutz in medizinischen Forschungsnetzen“ [9] der TMF. Forschungsdaten die außerhalb des Behandlungszusammenhanges bearbeitet werden müssen exportiert und dabei pseudonymisiert oder anonymisiert werden. Des Weiteren wird die Trennung der Daten gefordert, d.h. die personenbezogene Zuordnung von medizinischen Daten (MDAT) soll unabhängig von den medizinischen Daten und ohne Zugriff auf sie verwaltet werden. Die Forschungsdatenbank dagegen verwaltet medizinische Daten ohne Zugriff zur Identifikation des Patienten.

In der ersten Stufe erstellt der PID-Generator eine nichtsprechende Zeichenkette als Patientenidentifikator, d.h. die IDAT und PID werden in einer zentralen Personenliste gespeichert. Die Aufnahme eines Patienten in die Liste bedarf der informierten

Einwilligung des Patienten; sie muss Gegenstand der Patienteninformation und der Einwilligung sein.

In der zweiten Stufe wird von einem Treuhänder im Vorgang der Pseudonymisierung das Pseudonym (PSN, z.B. als kryptographische Transformation des PID) erzeugt und zusammen mit den MDAT an die Forschungsdatenbank weitergeleitet. Nur der Treuhänder darf die Zuordnung PID zur PSN kennen bzw. entschlüsseln.

Mit diesem Konzept gibt es im Forschungsnetz nur eine Stelle, die zur Auflösung von Pseudonymen in der Lage ist. Der Grundsatz der informellen Gewaltenteilung, der generell zwischen behandelndem und forschendem medizinischen Personal gefordert wird, kann somit effizient realisiert werden und Risiken, die in einem Fehlverhalten beteiligter Personen liegen können, mit hoher Sicherheit ausgeschlossen werden.

Sollte der PID mit den IDAT und den MDAT in den Behandlungseinrichtungen und mit den MDAT allein in der Forschungsdatenbank gespeichert werden – eine heute oft gebräuchliche Lösung - muss Betreibern und Nutzern der Datenbank der Zugriff auf die IDAT verwehrt werden. Eine Gefahr besteht jedoch darin, dass der behandelnde Arzt Informationen der Patienten kennt und eine Aufdeckung der PID möglich ist. Aus diesen Gründen sollte dieses Konzept bzw. Vorgehen vermieden werden; eine zwei-stufige Pseudonymisierung ist vorzuziehen.

Dieses Konzept wird z.B. vom Kompetenznetz Pädiatrische Onkologie und Hämatologie eingesetzt, jedoch mit unumkehrbarem PID [10].

### **3.1. Revision des Datenschutzkonzeptes**

Durch verschiedenste Anwendungen in den letzten Jahren wurde ein eindeutiger Weiterentwicklungsbedarf des Datenschutzkonzeptes der TMF deutlich. Einige Lücken und Problembereiche der bisherigen Konzepte, beispielsweise ein schlüssiges Konzept für Biomaterialien oder eine ausreichende softwareseitige Fundierung der Konzepte, werden schon jetzt von der TMF angegangen. Darüber hinausgehend gibt es aber eine Reihe wichtiger Punkte (z.B. Datenschutz bei Registern), die eine Weiterentwicklung der Datenschutzkonzepte notwendig erscheinen lassen.

Im Projekt V039-03 das seit Februar 2006 bei der TMF genehmigt ist wird ein modulares Neukonzept entworfen. Dies soll in 2008 abgeschlossen werden. [4]

## **4. Installation am IBE**

Der PID-Generator in der aktuellen Version 1.1. wurde im Februar 2007 am IBE auf einem UNIX-System (Debian) mit einer PostgreSQL Datenbank installiert.

### **4.1. Testumgebung**

Die Testumgebung steht jedem Mitglied des ELN (somit auch des Kompetenznetz Leukämie) zur Verfügung. Die Domain lautet: <http://pidgen.ibe.med.uni-muenchen.de/> Als Zugangsdaten müssen die Daten aus der ELN-Mitgliederdatenbank verwendet werden.

### **4.2. Erfahrung aus Installation**

Der PID-Generator ist in der Programmiersprache C geschrieben und sollte sich somit auf jedem UNIX-System problemlos kompilieren lassen. Durch unterschiedliche Versionen der C Compiler waren bei der Installation am IBE einige Anpassungen im Code nötig; diese haben aber keinen Einfluss auf die Funktionalität.

Für MS-Windows-Systeme ist eine installationsfertige Version vorhanden, diese wurde nicht getestet.

Als Datenbank wird PostgreSQL direkt unterstützt, hierfür ist eine Anpassung des PostgreSQL Client Authentication Configuration File (pg\_hba.conf) nötig. Andere Datenbanken können über eine ODBC-Schnittstelle angebunden werden, diese wurde nicht getestet.

### **4.3. Anforderungen und fehlende Funktionen**

Der PID-Generator wurde nach den Anforderungen im generischen Datenschutzkonzept [9] der TMF entwickelt. Die Tests und die Analyse am IBE haben gezeigt, dass die Anforderungen gut umgesetzt sind.

In der aktuellen Version fehlen jedoch folgende Funktionen bzw. Eigenschaften, die für einen produktiven Einsatz sehr wünschenswert wären.

- Weitere Phonetiken

Es sind nur die Kölner und Hannoveraner Phonetik (für deutschen Sprachraum) integriert. Das Einbinden von Phonetiken für den nicht-deutschen Sprachraum ist nur mit erheblichem Programmieraufwand möglich.

- Administrations-Tools

Jegliche Art von Administrationstools (z.B. Log-Rotation, Benachrichtigung bei Fehlern, Depseudominierungs-Oberfläche) fehlen.

- Validierung gemäß AMG / GCP
- Vollständige bzw. fehlerfreie Dokumentation

Im PID Anwenderworkshop der TMF am (26.9.2006 in Berlin) wurden diese Probleme auch angesprochen und Lösungsvorschläge wurden diskutiert. [5]

Aus softwaretechnischer Sicht entspricht die Implementierung teilweise nicht mehr den heute angestrebten Standards.

#### **4.4. Testbericht**

Für den Test wurde ein Datensatz aus dem GPOH – freundlicher Weise von Herr M. Sariyar zur Verfügung gestellt – mit 16.195 Datensätzen verwendet. Teilweise sind Patienten mehrfach enthalten, teilweise mit geänderten Daten (Umzug, Namensänderung). Vor allem bei diesen „doppelten“ Datensätzen und bei der zweiten Eingabe des Datensatzes wurde das Verhalten des PID-Generators genau getestet.

Es konnten kein auffälliges bzw. fehlerhaftes Verhalten des PID-Generators festgestellt werden. Die Ergebnisse wie in [8] und [10] beschrieben haben sich bestätigt. Der Test hat vor allem Erkenntnisse über mögliche Längen von Namen und fehlerhaften Postleitzahlen ergeben. Des Weiteren ist aufgefallen, dass es Patienten aus dem deutschsprachigen Raum gibt, die nicht-deutsche Namen haben. D.h. für diese Namen funktionieren die integrierten Phonetiken nicht.

Eine Aussage zu Fehleranfälligkeit in Abhängigkeit der ID-Felder kann nicht gemacht werden, da keine Aussagen zur Anzahl der Fehleingaben vorliegen. An Hand von Tests konnte hierzu auch keine Aussage getroffen werden.

#### **4.5. ToDo's für Einführung Im Kompetenznetz Leukämie**

Folgende Punkte müssten für den Einsatz des PID-Generators im Kompetenznets Leukämie geklärt werden.

- Einheitliche Erhebung der IDAT

Die Erhebung der IDAT muss möglichst einheitlich sein. Als Basis wird der Datensatz der Versichertenkarte, ein Alternativname und Kennzeichen der meldenden Klinik. D.h.: Vorname, Nachname, Alternativname, Geburtsdatum, PLZ, Wohnort, Geschlecht, Kliniknummer

- Rückidentifizierung

Eine Rückidentifizierung ist technisch möglich. Diese muss durch ein Regelwerk abgesichert werden.

- Einwilligung des Patienten

Die Aufnahme eines Patienten in die Patientenliste bedarf der informierten Einwilligung des Patienten; sie muss Gegenstand der Patienteninformation und der Einwilligung sein. Dies muss von der entsprechenden Studienzentrale umgesetzt werden. Der PID-Generator ist somit nur für zukünftige Studien einsetzbar.

- Datentreuhänder

Das IBE könnte die treuhändischen Aufgaben für die Patientenliste übernehmen, solange der PID mittels einer Pseudonymisierungssoftware in eine PSN umgewandelt wird, d.h. durch den PID keine direkte Beziehung zu den MDAT herstellbar ist. Der PID-Generator und ein Pseudonymisierungsdienst kann somit nicht gleichzeitig vom IBE betrieben werden.

- Technische Sicherheit

Die Patientenliste bzw. der PID-Generator darf nur durch registrierte Teilnehmer genutzt werden. Dafür nötig ist eine starke Authentifizierung im SSL-Protokoll, Firewallfunktionen für das Rechnersystem und die Bereitstellung eines dedizierten Rechners für die Datenbank mit kontrolliertem Kanal zum Kommunikationsrechner. D.h. die Anschaffung eines entsprechenden neuen Servers durch das Kompetenznetz wäre erforderlich. Das IBE würde Installation und Administration übernehmen.

## 5. Fazit

Der PID-Generator erfüllt die im generischen Datenschutzkonzept [9] der TMF beschriebenen Anforderungen sehr gut. Technisch funktioniert der PID-Generator ebenfalls ohne Probleme, hat jedoch die in 4.3 angesprochenen Defizite. Aus datenschutzrechtlicher Sicht macht der Einsatz des PID-Generator für die Zusammenführung von Daten aus verschiedenen Studien nur im Gesamtkonzept der generischen Lösung der TMF zum Datenschutz sinn.

Aus den genannten technischen und datenschutzrechtlichen Gründen rät das IBE den Mitgliedern des Kompetenznetzes Leukämie vorerst vom alleinigen Einsatz des PID-Generators ab.

Sollte im Netzwerk trotzdem Interesse an der Einführung bzw. dem Einsatz des PID-Generators bestehen, müssten nochmals sowohl die technischen als auch datenschutzrechtlichen Anforderungen gemeinsam genau diskutiert werden. Vor allem die erwähnten ToDo's müssten konkretisiert werden.

Das Ziel der einheitlichen Umsetzung der Pseudonymisierung von Forschungsdaten im Netzwerk darf nicht aus den Augen verloren werden. Im nächsten Schritt sollten die Anforderungen und der Bedarf des Kompetenznetzes Leukämie an einem Pseudonymisierungsdienstes diskutiert werden. Dies sollte in enger Anlehnung an das neue Datenschutzkonzept der TMF erfolgen.

## Literaturverzeichnis

- [1] <http://www.kompetenznetz-leukaemie.de>
- [2] <http://ibe.web.med.uni-muenchen.de>
- [3] <http://www.tmf-ev.de>
- [4] TMF: Projektantrag und Votum zum Hauptprojekt (V039-03)  
[http://www.tmf-ev.de/site/DE/ext/binary/Projekte/Voten/2006-02-22/V039-03\\_DS-Konzept\\_II\\_PA\\_PAVotum.pdf](http://www.tmf-ev.de/site/DE/ext/binary/Projekte/Voten/2006-02-22/V039-03_DS-Konzept_II_PA_PAVotum.pdf)
- [5] TMF: Protokoll des Anwender-Workshops vom 26.9.2006  
[http://www.tmf-ev.de/site/DE/int/AG/DS/Projekte/PID-Generator/WS\\_09-2006/2006-09-26\\_Protokoll\\_WS-PID.pdf](http://www.tmf-ev.de/site/DE/int/AG/DS/Projekte/PID-Generator/WS_09-2006/2006-09-26_Protokoll_WS-PID.pdf)
- [6] PID-Produktbeschreibung; <http://www.staff.uni-mainz.de/pommeren/PID/PID-Produktbeschreibung.pdf>
- [7] A. Faldum/ K. Pommerening, An optimal code for patient identifiers. *Computer Methods and Programs in Biomedicine* 79 (2005), 81-88
- [8] Moormann J, Pommerening K: Patienten-Identifikatoren in medizinischen Forschungsnetzen: Evaluation des Matchalgorithmus. GMD S 2005.
- [9] Generische Lösungen der TMF zum Datenschutz für die Forschungsnetze der Medizin. Version 1.10 vom 1. Juli 2003; P. Debold, C.-M. Reng
- [10] Glock J, Herold R, Pommerening K: *Personal identifiers in medical research networks: Evaluation of the personal identifier generator in the Competence Network Paediatric Oncology and Haematology*. GMD Med Inform Biom Epidemiol. 2/2 (2006), Doc06.