Patienten-IDentifikatoren in medizinischen Forschungsnetzen: Evaluation des Matchalgorithmus

Jutta Glock, Klaus Pommerening; Institut für Medizinische Biometrie, Epidemiologie und Informatik der Johannes-Gutenberg-Universität Mainz

Begriffe

PID: Studienübergreifender, eindeutiger Patienten-Identifikator

PID-Generator: Verarbeitet PID-Anfragen, vergleicht bestimmte Merkmale des eingegebenen Datensatzes mit der Patientenliste, liefert bei einem Match den PID zurück oder generiert einen neuen PID und speichert den Fall in der Patientenliste

Homonymfehlerrate: Wie oft wird verschiedenen Patienten derselbe PID zugeteilt, d.h. fälschlicherweise gematched? Die Homonymfehlerrate hängt maßgeblich von den gewählten Datenfeldern ab („echte“ Homonyme vermeiden).

Synonymfehlerrate: Wie oft werden einem Patienten zwei oder mehrere PIDs zugeordnet, d. h. fälschlicherweise nicht gematched? Die Synonymfehlerrate hängt vor allem von der Datenqualität und organisatorischen Gegebenheiten (Häufigkeit von Mehrfacheingaben) ab und ist daher je nach Anwendung und Datenquelle unterschiedlich.

Ablauf einer PID-Anfrage

Der Anwender gibt die Patientendaten über eine Weboberfläche ein und erhält als Ergebnis einen PID oder eine Fehlermeldung.

PID: G3HU7FEP

Es wurde ein passender Fall gefunden.

Sie können den PID verwenden.

Anwender (Studienzentrale)

PID-Server

PID-Generator

Matchalgorithmus

		G3HU7FEP
Müller		Müller
Hans-Georg		Hans
26.02.1953	= ?	28.02.1953
männlich		männlich
		21614
Buxtehude		Buxtehude
Deutschland		Deutschland
unsicher		unsicher

Match → • **PID** zurückliefern • Datensatz aktualisieren

Kein Match → • Neuen **PID** erzeugen und • **PID** zurückliefern • Neuen Datensatz anlegen

PID

Patientenliste (PID-Datenbank)

Der PID-Generator vergleicht den eingegebenen Datensatz mit denen in der Patientenliste. Je nachdem, ob ein Match gefunden wird, liefert er einen bereits vorhanden PID zurück oder generiert einen neuen.

KPOH-Konfiguration

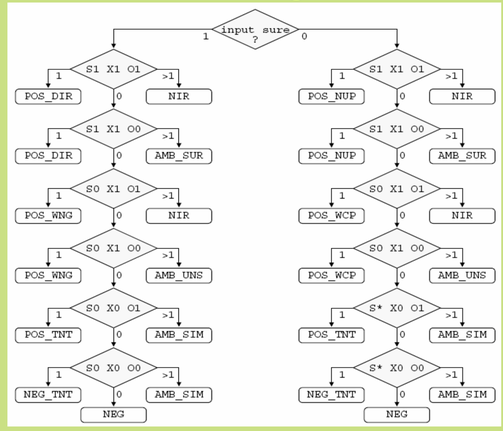
Das Matchverfahren ist insbesondere abhängig von der Wahl der Datenfelder, die zum Abgleich verwendet werden, und der Gestaltung des Entscheidungsbaums. Beides ist im PID-Generator frei konfigurierbar. Hier dargestellt sind die getesteten Spezifikationen des Kompetenznetzes Pädiatrische Onkologie und Hämatologie (KPOH).

Datenfelder

Feldname	Bedeutung	Wertebereich	Pflichtfeld	Relevanz [1]
lname	Nachname	string	Ja	+
aname	Alternativer Nachname (z. B. Geburtsname)	string	Nein	+
fname	Vorname(n)	string	Ja	+
bd	Geburtstag	0-31	Ja	+
bm	Geburtsmonat	0-12	Ja	+
by	Geburtsjahr	1000-9999	Ja	+
plz	Postleitzahl	string	Nein	-
loc	Wohnort	string	Nein	*
state	Land	string	Nein	*
sex	Geschlecht	[f | m | n]	Nein	*

[1] + wird immer beim Matchen verwendet
* wird optional beim Matchen verwendet
\- keine Matchrelevanz

Entscheidungsbaum



Legende

Rauten stellen Datenbankabfragen, abgerundete Rechtecke Resultate dar. Nach jeder Abfrage wird entweder 0, 1 oder mehr als 1 Match gefunden. Abhängig davon wird entweder eine neue Abfrage durchgeführt oder ein Resultat erreicht. Die Resultatnamen stehen jeweils für eine bestimmte Reaktion des PID-Generators. Nur bei Erreichen des Resultats NEG wird ein neuer PID generiert, in allen anderen Fällen wird ein bereits vorhandener PID oder eine Fehlermeldung zurückgeliefert. Die Kürzel in den Rauten stellen Filter dar, die auf die Datenbank angewendet werden.

S **S**ureness [0 = unsicher | 1 = sicher | * = beides]

X **Ex**actitude [0 = phonetische | 1 = exakte Übereinstimmung]

O **O**ptionality [0 = ohne | 1 = mit optionalen Daten]

Tests / Ergebnisse

Funktionstest mit fiktiven Daten

- Alle Pfade des Matchbaums werden durchlaufen
- Alle Tests liefern die erwarteten Ergebnisse

Realtests mit 14.915 Datensätzen mit und ohne Verschlüsselung der Datenfelder

- 14.913 neue PIDs
- und 2 Matche erzeugt
- Verschlüsselung hat erwartungsgemäß keinen Einfluss
- 0 Homonymfehler (da beide Matche korrekt)
- 6 Synonymfehler (beruhen meist auf Fehler im Geburtsdatum)

PIDs für das Deutsche Kinderkrebsregister (DKKR)

- 44.248 Datensätze vs. 2.579 bereits vorhandene Einträge in der Patientenliste
  - davon 527 Datensätze bereits mit PID

Ergebnisse

- 1.569 Matche (davon 1 Duplikat in DKKR-Daten)
- 53 sehr unsichere Matche (Resultat NEG_TNT, liefert Warnmeldung)
  - davon 10 in Übereinstimmung mit bereits vorhandenen PIDs
- 0 Homonyme innerhalb der 44.248 Datensätze
- 22 Synonyme

Fiktive Testdaten (Beispiele)

Sicher?	Name	Altname	Vorname	Geburtstag	Geschlecht	Resultat	PID
Ja	Albrecht		Anton	19.01.2001	m	NEG	G1QP56LV
Ja	Alt	Albrecht	Anton-Arndt	19.01.2001	m	POS_DIR	G1QP56LV
Nein	Meier		Moritz	12.11.2002	m	NEG	4VV50DPW
Nein	Maier		Mohritz	12.11.2002	m	POS_TNT	4VV50DPW
Ja	Veit		Verena	12.08.2003	w	NEG	MAPVTF8K
Ja	Veit		Viviane	12.08.2003	w	NEG	1HUZWWLY
Nein	Veit		Verena-Viviane	12.08.2003	w	NIR	???

Stand der GPOH-Patientenliste (20.07.2005)

PIDs gesamt		45.693
Mehrfachanfragen		
	4-6-fach	28
	3-fach	214
	2-fach	3.299
Resultate (seit 03.06.05)		
POS_DIR	(Match mit sicheren Daten bei sicherer Eingabe)	9
POS_NUP	(Match mit sicheren Daten bei unsicherer Eingabe)	1.247
POS_WNG	(Match mit unsicheren Daten bei sicherer Eingabe)	28
POS_WCP	(Match mit unsicheren Daten bei unsicherer Eingabe)	2.204
NEG_TNT	(Phonetische Ähnlichkeit, optionale Felder verschieden → kein PID)	123
NEG	(kein Match → neuer PID)	43.117