Evaluation des Matchalgorithmus bei der Generierung von Patienten-Identifikatoren in medizinischen Forschungsnetzen

Moormann J, Pommerening K

Institut für Medizinische Biometrie, Epidemiologie und Informatik, Johannes-Gutenberg-Universität Mainz, Deutschland
moormann@imbei.uni-mainz.de

Einleitung und Fragestellung

Das Kompetenznetz POH (Pädiatrische Onkologie und Hämatologie) setzt bereits seit dem Jahr 2003 den so genannten PID (Patienten-Identifikator) ein, um Probanden verschiedener Studien ein eindeutiges, dabei aber nichtsprechendes Identifikationsmerkmal zuzuordnen. Die hierfür verwendete Software, der PID-Generator [1], wurde in einem Projekt der TMF (Telematikplattform für Medizinische Forschungsnetze e.V.) weiterentwickelt und findet zunehmend auch in anderen Kompetenznetzen Verwendung. Durch seine Funktionalität unterstützt er die Einrichtung einer Patientenliste im Rahmen des Pseudonymisierungsdienstes, der von den generischen Lösungen der TMF zum Datenschutz der Forschungsnetze [2, 3] gefordert wird. Im PID-Generator ist eine deterministische Record-Linkage-Methode implementiert. Dieses Matchverfahren zum Abgleich neu eingegebener mit bereits vorhandenen Personendaten kann dabei über verschiedene Konfigurationsoptionen gesteuert werden. Um Aussagen über die Qualität des Verfahrens treffen zu können, ist es z. B. notwendig, die Homonym- und Synonymfehlerraten zu ermitteln oder zumindest abzuschätzen. Von Interesse ist außerdem, mit welcher relativen Häufigkeit die verschiedenen Matchergebnisse – und in Abhängigkeit davon bestimmte Rückmeldungen und Aktionen – auftreten. In diesem Zusammenhang wird auch das Vorkommen von Mehrfachanfragen für einen Patienten ausgewertet.

Material und Methoden

Eine Besonderheit des PID-Generators ist seine flexible Konfiguration mit zahlreichen Einstellungsmöglichkeiten. Die Attribute, die zu jeder Person erfasst werden, und jene, die für das Matchverfahren herangezogen werden, sind für jedes Forschungsnetz frei wählbar. Des Weiteren wird zwischen Pflichtfeldern und optionalen Feldern unterschieden, und mehrere Normalisierungs- und Transformationsfunktionen stehen zur Verfügung. So wurde z. B. für den Abgleich von Namen ein spezielles Verfahren implementiert, das die Namen in ihre Bestandteile zerlegt und dadurch auch kreuzweise Vergleiche der Komponenten ermöglicht. Außerdem ist ein Abgleich über phonetische Codes vorgesehen, um die vergleichsweise häufig auftretenden Eingabefehler bei Namen aufzufangen.
Auch die durchzuführenden Aktionen bei einem bestimmten Matchergebnis (z.

B. Update der Daten, neuen PID erzeugen, Nachricht an den Administrator schicken, Rückmeldung an den Benutzer) sind konfigurierbar.
Um eine Evaluierung des Matchverfahrens durchführen zu können, werden definierte Konfigurationen vorausgesetzt. Hierbei werden vor allem die in den Kompetenznetzen POH und AHF (Angeborene Herzfehler) bereits real verwendeten Konfigurationen berücksichtigt. Des Weiteren wird der Einfluss der Einbeziehung phonetischer Codes, der Geburtsdaten und der optionalen Felder geprüft.
Für die Auswertung der Matchmethode werden u. a. Patientendaten aus der Referenzdatenbank des Universitätsklinikums Mainz verwendet. Für die Abschätzung möglicher Homonymfehlerraten werden auch externe Quellen sowie Erfahrungen aus Landeskrebsregistern herangezogen.
Mit Hilfe der Protokollierung der Anfragen wird das Auftreten der verschiedenen Matchergebnisse untersucht, während das Vorkommen von Mehrfachanfragen anhand des Realbetriebes des Kompetenznetzes POH überprüft wird.

Ergebnisse

Bei den verschiedenen Tests wird ein Schwerpunkt auf die in den Kompetenznetzen POH und AHF eingesetzten Konfigurationen gelegt. Für jede Einstellung werden die Homonym- und Synonymfehlerraten ermittelt, so dass die Eignung des Matchverfahrens bzw. der Einfluss der Einbeziehung von phonetischen Codes und verschiedenen Datenfeldern evaluiert werden kann. Des Weiteren wird über Erfahrungen mit der bisherigen Nutzung, dabei auftretende Probleme und die Benutzerakzeptanz berichtet.

Diskussion

Aufgrund der großen Flexibilität in der Konfiguration des PID-Generators ist es nicht möglich, eine für alle Szenarien gültige Aussage zu treffen. Insbesondere die Synonym- und Homonymfehlerraten sind im hohen Maße abhängig von der Qualität (Korrektheit und Vollständigkeit) der eingegebenen Datensätze und davon, ob die Krankenversichertennummern im Matchverfahren verwendet werden können. Das Vorkommen von Mehrfachanfragen hängt auch von organisatorischen Gegebenheiten ab. Installationen, die sich an den Konfigurationen der Kompetenznetze POH und AHF orientieren, erhalten jedoch durch die ermittelten Ergebnisse Anhaltspunkte für die Qualität und Zuverlässigkeit ihrer eingesetzten Matchmethode.
Die Ergebnisse liefern die Grundlage für weitergehende Verfeinerungen des Matchalgorithmus, z. B. die Erweiterung um stochastische Matchmethoden.

Literatur

[1]	Pommerening K, Wagner M. Ein Pseudonymisierungsdienst für medizinische Forschungsnetze. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 2001;32: 251
[2]	Reng CM, Debold P, Adelhard K, Pommerening K. Generisches Datenschutzkonzept für Forschungsnetze in der Medizin. Im Druck

[3]	Reng CM, Debold P, Adelhard K, Pommerening K. Vernetzte medizinische Forschung – Akzeptiertes Datenschutzkonzept. Dtsch Arztebl 2003; 100: A 2134-2137 [Heft 33]