Beispieldatensatz für das Informationssystem Versorgungsdaten

Peter Ihle[1], Thomas Grobe[2], Christoph Beger[3], Matthias Löbe[3]

[1]PMV forschungsgruppe, Universitätsklinikum Köln
[2]AQUA - Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH, Göttingen
[3]Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE), Universität Leipzig

Motivation

Das Informationssystem Versorgungsdaten am DIMDI stellt auf Basis der Datentransparenzverordnung seit Anfang 2014 Routinedaten der gesetzlichen Krankenkassen für gesetzlich festgelegte Nutzergruppen zur Verfügung. Die Datenbasis als Vollerhebung aller GKV-Versicherten enthält damit Gesundheitsinformationen für knapp 90% der deutschen Bevölkerung und stellt demzufolge eine wichtige Quelle für die Versorgungsforschung dar. Auswertungen auf Basis dieser Datenquelle können Über-, Unter- und Fehlversorgung aufdecken und helfen, die medizinische Versorgung der Versicherten zu verbessern.

Die Nutzungsmodalitäten sind im SGB V und in der Datentransparenzverordnung geregelt. Das Antragsverfahren bietet aktuell zwei Möglichkeiten, die nach aktuellem Stand bisher etwa gleich häufig beantragt wurden: Der Antragsteller kann sich auf die Formulierung der Forschungsfrage sowie die Beschreibung der Methodik beschränken und die Skripterstellung den Analysten des Informationssystems Versorgungsdaten im DIMDI überantworten. Alternativ besteht die Möglichkeit, bei Antragstellung ein lauffähiges Skript (aktuell in Oracle SQL oder in SAS) einzureichen. Dieses Skript wird nach inhaltlicher Prüfung auf dem DaTraV-Volldatensatz ausgeführt. Bei syntaktischen oder inhaltlichen Auffälligkeiten kann das Skript in Abstimmung mit dem DIMDI nachträglich angepasst werden. Bei beiden Möglichkeiten erhält der Antragsteller im Rahmen eines Bescheids eine so genannte Ergebnismenge. Zur Wahrung des Identifikationsschutzes der Versicherten darf diese nur aggregierte Daten enthalten und muss faktisch anonymisiert sein. Der Antrag ist kostenpflichtig.

Zur Entwicklung von lauffähigen Auswertungsprogrammen sind möglichst realitätsnahe und umfängliche Testdaten sowie univariate Verteilungen der Variablen nicht nur wünschenswert, sondern in einem gewissen Umfang in der Praxis auch zwingend erforderlich. Insbesondere Nutzer, die nicht bereits über umfängliche Erfahrungen mit vergleichbaren Daten verfügen, müssen die Gelegenheit erhalten, sich einen ersten Eindruck vom Umfang und von der Häufigkeit der Dokumentation bestimmter Ereignisse in den Daten machen zu können (z. B. welche und wie viele Diagnosen und Arzneiverordnungen werden in etwa bei Versicherten im Alter ab 60 Jahren pro Jahr erfasst).

Für die Programmierung des Auswertungsskripts stellt das DIMDI aktuell einen Beispieldatensatz zur Verfügung. Der Datensatz wurde künstlich erzeugt und enthält in der aktuell vorliegenden Version beispielsweise für jeden ICD-Code die identische Anzahl an Versicherten, gleichverteilt über alle Geschlechts- und Altersgruppen. Damit kann ein bestehendes Skript auf formale syntaktische Korrektheit geprüft werden und der Abstimmungsbedarf zwischen den Mitarbeitern des Informationssystems Versorgungsdaten und den Forschern wird verringert. Eine inhaltlich-empirische Prüfung selbst von sehr grundlegenden Annahmen und Voraussetzungen ist jedoch ausgeschlossen. Damit wird aus Sicht der Forscher der zeitliche Abstimmungsbedarf nach Einreichung des ersten Skriptes nicht in dem Maße reduziert, wie es mit einem Beispieldatensatz möglich wäre, der nicht nur auf die Prüfung formal syntaktischer Korrektheit abzielt.

Dem Interesse von Forschern/Nutzern nach einer möglichst umfänglichen Bereitstellung entsprechender Testdaten stehen Überlegungen und Bestimmungen zum Datenschutz entgegen: Aufgrund der großen Zahl an denkbaren Merkmalskombinationen, insbesondere bei Daten über längere Zeiträume, lässt sich eine Re-Identifikation von einzelnen Personen in den DaTraV-Daten trotz der Pseudonymisierung nicht in jedem Fall ausschließen. Hier gilt es, eine Güterabwägung zwischen dem Recht auf Forschung und dem

Recht auf informationelle Selbstbestimmung vorzunehmen. Wie auch in anderen Projekten auf Basis von GKV-Routinedaten sollte dem Forscher diese Reidentifikationsmöglichkeit bewusst sein. So ist es für die Erreichung des Forschungsziels zuweilen notwendig, einzelne Versichertenverläufe oder niedrig aggregierte Zwischentabellen zu erzeugen und diese im Rahmen von Plausibilisierungsschritten zu sichten und zu prüfen. Im Rahmen von Publikationen werden allerdings ausschließlich höher aggregierte Zahlen veröffentlicht. Damit würde den Forschungsinteressen einerseits und der Wahrung auf Identität der Versicherten andererseits ausgewogen Rechnung getragen.

Anforderungen an einen Beispieldatensatz

Unabdingbar für die Entwicklung von Auswertungsprogrammen erscheint, dass die grundlegende Struktur der Gesamtdaten auch in den Testdaten abgebildet wird, also grundsätzlich alle Datentabellen mit allen Merkmalen/Variablen bereitgestellt werden und die relationalen Beziehungen der Testdaten denen der Gesamtdaten entsprechen, weshalb ein einfacher Ausschluss sensibler Merkmale bei der Bereitstellung von Testdaten nicht in Frage kommt.

Der Beispieldatensatz muss idealerweise auch die im Originaldatensatz enthaltenen Inkonsistenzen und Implausibilitäten enthalten, damit diese durch geeignete Methoden bei der Skripterstellung programmtechnisch aufgefangen werden können. Inkonsistenzen und Implausibilitäten sind bei einer Routinedatenbasis wie den GKV-Routinedaten immanent enthalten, da die Erhebung der Daten nicht, wie beispielsweise in klinischen Studien üblich, auf Basis eines Studienprotokolls, sondern im Rahmen von administrativen Vorgaben erfolgt und primär Abrechnungszwecken dient. Eine Routinedatenbasis, hier die DaTraV-Daten, ist immer ein so genannten Rohdatensatz, aus dem im Rahmen eines Projekts durch Aufbereitungs- und Validierungsschritte erst ein projektspezifischer Auswertungsdatensatz entsteht. Die Aufbereitungsschritte fokussieren hierbei i. d. R. nur auf solche Variablen und Abhängigkeiten, die für die Beantwortung der Fragestellung relevant sind. Inhaltliche Implausibilitäten mit Relevanz für einzelne Fragestellungen sind beispielsweise Schwangerschaften bei Männern oder geriatrische Diagnosen bei Kindern. Ebenso finden sich datentechnische Inkonsistenzen, wie fehlende Variablen oder implausible Ausprägungen von Variablen, z. B. Angabe von mehr als 10.000 abgegebenen Packungen pro Verordnung.

Ein Beispieldatensatz darf also kein fehlerfreier aufbereiteter valider Datensatz sein, sondern muss ganz im Gegenteil idealerweise alle Schwächen des Originaldatensatzes abbilden. Dabei muss allerdings auch vermieden werden, dass bei der Erstellung des Testdatensatzes Auffälligkeiten artifiziell erzeugt werden, die sich im Originaldatensatz nicht finden.

Generierung eines Testdatensatzes

Ein erster Schritt bei der Bereitstellung von Testdaten zur Einschränkung von Re-Identifikationsmöglichkeiten sollte in der Ziehung einer Zufallsstichprobe aus der Grundgesamtheit aller in DaTraV-Daten erfassten Personen bestehen. Eine solche Zufallsstichprobe aus dem Originaldatensatz kann die oben genannten Anforderungen weitgehend erfüllen. Zudem kann eine solche Zufallsstichrobe mit verhältnismäßig geringem Aufwand gezogen werden, beispielsweise durch Bildung eines Hashcodes aus dem überjährigen Versichertenpseudonym. Dieses Verfahren wurde einem Methodenforschungsprojekt des Statistischen Bundesamtes entwickelt [Ihle P, Köster I, Schubert I, et al: GKV-Versichertenstichprobe. Wirtschaft und Statistik, Heft 9: 1999; 742-749] und in der Versichertenstichprobe AOK Hessen / KV Hessen [Ihle P, Köster I, Herholz H, Rambow-Bertram P, Schardt T, Schubert I: Versichertenstichprobe AOK Hessen/KV Hessen - Konzeption und Umsetzung einer personenbezogenen Datenbasis aus der Gesetzlichen Krankenversicherung. Das Gesundheitswesen 2005; 67: 638-645] angewendet. Bei der Ziehung ist eine Besonderheit der DaTraV-Daten zu beachten: Für Versicherte, die lediglich ein jahresbezogenes Pseudonym

besitzen, wird der Hashcode ersatzweise aus diesem jahresbezogenen Pseudonym gebildet, damit auch diese Versichertengruppe Eingang in den Beispieldatensatz findet. Bei einer späteren Ergänzung der Daten durch weitere Beobachtungsjahre oder Versichertenpopulationen (z. B. die Gruppe der Versterbenden eines Jahres) wären jeweils gleichartige Ziehungen von Zufallsstichproben ausschließlich aus der Gruppe der neu hinzukommenden Personen notwendig.

Als minimal nutzbare Stichprobengröße wird von den Autoren eine 0,1%-Stichprobe angesehen. Entsprechende Daten würden innerhalb eines Beobachtungsjahres lediglich Informationen zu rund 70.000 Personen mit Versicherung in der GKV enthalten. Für 99,9% aller GKV-Versicherten finden sich in einer entsprechenden Stichprobe folglich keinerlei Informationen. Die Bereitstellung einer Stichprobe mit kleinem Auswahlsatz scheint damit auch unter datenschutzrechtlichen Aspekten möglich zu sein. Zusätzliche flankierende Auflagen zur Nutzung dieses Beispieldatensatzes werden weiter unten beschrieben.

Grundsätzlich ist nicht davon auszugehen, dass in einer entsprechend kleinen Stichprobe alle in den Gesamtdaten vorhandenen Merkmalsausprägungen vorkommen. Um dennoch alle potenziell im Rahmen einer speziellen Auswertung denkbaren Merkmalsausprägungen bei der Programmierung berücksichtigen zu können, ist neben der Stichprobe auch die Bereitstellung von Informationen zu Häufigkeiten aller Merkmalsausprägungen aller versichertenbezogenen Merkmalen aus allen Datentabellen obligat erforderlich. Auf die Bereitstellung der Merkmalsausprägung der Versichertenpseudonyme kann hierbei verzichtet werden. Aus Datenschutzgründen könnten die Häufigkeitskategorien bei kleinen Zellbesetzungen (zwischen 1 und 29) lediglich die Angabe „29“ enthalten, entsprechend der Information „Häufigkeit <30“. Inwieweit auch weitere Kategorisierungen vorgenommen werden können, beispielsweise in logarithmusähnlichen Gruppen wie 30 bis 99, 100 bis 299, 300 bis 999, 1000 bis 2999, usw., ist zu diskutieren. Zu diskutieren ist, ob auch diese aggregierten Angaben dem Antragsteller lediglich über ein projektspezifisches Webportal zur Verfügung gestellt werden sollte. Öffentlich verfügbare Verteilungen Angaben könnten dann höher aggregiert ausfallen.

Die vorgeschlagene Stichprobengröße mit einem Auswalsatz von 0,1% dürfte aus Erfahrung bisheriger Sekundärdatenanalysen i. d. R. nicht ausreichen, um Ergebnismengen mit hinreichend gut besetzten Zellen zu liefern. Mit steigendem Komplexitätsgrad der Auswertungen werden die Zellbesetzungen erwartungsgemäß immer kleiner. Dennoch dürfte eine solche Stichprobe erste Hinweise liefern, inwieweit auch bei Auswertung des Volldatensatzes Beschränkungen bestehen bleiben und daher eine Änderung des methodischen Vorgehens sinnvoll sein könnte.

Vorschläge für Zusammenfassungen von Merkmalsausprägungen

Aus datenschutzrechtlicher Sicht müssen bei der Bereitstellung von versichertenbezogenen Daten Maßnahmen zur Sicherstellung der Anonymität erfolgen Nachfolgend findet sich Liste von Empfehlungen, welche der Versichertenmerkmale im Beispieldatensatz verändert werden können, um das Reidentifikationsrisiko eines Versicherten weiter zu senken, bzw. auszuschließen. Gleichzeitig wurde darauf geachtet, die Nutzbarkeit des Beispieldatensatzes für die Programmierung von Auswertungsskripten möglichst wenig einzuschränken und die vorgeschlagenen Modifikationen von Merkmalsausprägungen stringent und damit überschaubar zu gestalten. Durch die vorgeschlagenen Modifikationen werden alle direkten und indirekten Informationen zu Zeitpunkten von Ereignissen auf eine Monatsgenauigkeit beschränkt. Zudem werden Angaben zu Ausgaben auf maximal zwei signifikante Stellen beschränkt. Merkmale zur regionalen Zuordnung erlauben nach Modifikation lediglich eine Zuordnung auf der Ebene von Bundesländern. Die vorgeschlagene Modifikation gewährleistet des Weiteren, dass bei ursprünglich gültigen Merkmalsausprägungen auch nach Modifikation weiterhin gültige Merkmalsausprägungen im ursprünglichen Format resultieren.

SA151 – Versichertenstamm:

- Merkmal: SA151_VERSICHERTENTAGE
  Da für einen weit überwiegenden Teil der Verstorbenen von einer durchgängigen Versicherung ab Jahresbeginn bis zum Todestag ausgegangen werden kann, sollte die Zahl der Versicherungstage bei Verstorbenen derart verändert werden, dass der (zumindest vermeintlich) aus den angegebenen Versichertentagen ableitbare Todestag immer auf den 15. des ursprünglich abgeleiteten Monats fällt. Analoges würde auch für Angaben zu Versichertentagen bei Geburt im Berichtsjahr gelten.
- Merkmal: SA151_GEBURTSJAHR
  Die alleinige Angabe des Geburtsjahrs stellt bereits eine Vergröberung des tagesgenauen Geburtstages dar und sollte daher unverändert in den Beispieldatensatz übernommen werden, um projektspezifische Altersklassen bilden zu können.

SA152 – Versichertenstamm

- unverändert

SA153 - Extrakorporale Blutreinigung

- unverändert

SA451 - Ambulante Arzneimittel:

- Merkmal: SA451_VERORDNUNGSDATUM
  Alle Tageswerte werden auf den 15. des jeweils angegebenen Monats gesetzt, womit Arzneiverordnungen nur noch einem bestimmten Monat zugeordnet werden können. Diese Vergröberung der Datumsangabe erlaubt weiterhin Auswertungen in zeitlicher Relation zu den monatsgenauen stationären Diagnosen.

SA651 - Ambulante Diagnosen

- unverändert

SA751 – Leistungsausgaben:

- Merkmale (n=7): SA751_AERZTE bis SA751_KRANKENGELD
  Alle Angaben unterhalb von 10 Euro werden auf eine signifikante Stelle gerundet, alle Angaben ab 10 Euro werden auf zwei signifikante Stellen gerundet. Beispiele: statt 9,67: 10,00; statt 8,31: 8,00; statt 12,85: 13,00; statt 5638,77: 5700,00; statt 127456,45: 130000,00 Euro (die Beispielangaben erfolgen hier ausschließlich zur besseren Lesbarkeit in Euro und sind auf die in den Daten ausgewiesenen Cent-Beträge zu übertragen).

SA551 - Stationäre Diagnosen

- unverändert

SA951 – Krankenkassenzugehörigkeit

- unverändert

SA999 – Amtlicher Gemeindeschlüssel

- Merkmal: SA999_GS
  Alle gültigen 5-stelligen Kreisschlüssel aus einem Bundesland (BL) werden auf die Kennung des Kreises mit der höchsten Einwohnerzahl innerhalb des jeweiligen Bundeslandes im Jahr 2011 gesetzt (womit letztendlich eine BL-Kennung resultiert, jedoch zugleich immer ein gültiger Kreisschlüssel in den Daten erfasst ist).
- Merkmal: SA999_GS_LAND bleibt unverändert

- Merkmal: SA999_GS_RB (Herleitung wie bei SA999_GS)
- Merkmal: SA999_GS_KREIS (Herleitung wie bei SA999_GS)
  Anmerkung: PLZ ließen sich ggf. mit einem analogen Vorgehen auf „gültige“ PLZ mit ein oder zwei signifikanten Stellen reduzieren.

Neben pseudonymisierten versichertenbezogenen Daten werden in der DaTraV-Umgebung eine Reihe zusätzlicher Informationen ohne datenschutzrechtliche Relevanz bereitgestellt, die, sofern es sich um frei verfügbare Daten handelt, möglichst gemeinsam mit den Testdaten bereitgestellt werden sollten. Problemlos sollte dies beispielsweise bei Informationen zum ICD-10-GM möglich sein. Im Hinblick auf Arzneimittel sollte geklärt werden, inwiefern wenigstens Angaben zu ATC-Codes sowie DDD für die in der Stichprobe enthaltenen PZN bereitgestellt werden können.

Alternativen zum Stichprobenverfahren

Unter den Autoren wurde als Alternative zu einer Zufallsstichprobe auch die Generierung eines modellbasierten Beispieldatensatzes diskutiert. Hierfür werden aus dem Originaldatensatz Regeln generiert, mit Hilfe derer anschließend künstliche Modellversicherte erzeugt werden. Ein artifizieller Testdatensatz sollte vor diesem Hintergrund als datenschutzrechtlich unbedenklich gelten, da durch die Zuordnungsregeln nur Daten künstlicher Personen generiert werden.

Allerdings erscheint die Erstellung der Regeln nur mit erheblichem zeitlichem und aktuell nicht abschätzbarem Aufwand möglich zu sein, insbesondere dann, wenn der regelbasierten Modelldatensatz ein breites Spektrum möglicher Forschungsfragen abdecken soll.

Flankierende Maßnahmen

Weitere vertragliche und technische Maßnahmen können die intendierte Nutzung des Beispieldatensatzes unterstützen.

Der Beispieldatensatz wird dem Antragsteller im Rahmen eines bereits gestellten Antrags als Scientific Use File zur Verfügung gestellt werden. Vertraglich wird geregelt, dass der Datensatz ausschließlich zum Zwecke der Skripterstellung und -prüfung genutzt werden darf und anschließend zu löschen ist. Veröffentlichungen auf Basis des Testdatensatz sind prinzipiell ausgeschlossen und nur im Einzelfall in Abstimmung mit dem DIMDI erlaubt.

Durch den Einsatz geeigneter technischer Hilfsmittel, wie beispielsweise eines personalisierten Token-Generators für einen VPN-gesicherten Remotedesktopzugriff, könnte ergänzend sichergestellt werden, dass der Zugriff auf den Beispieldatensatz nur durch vertraglich benannte und autorisierte Personen erfolgt. Auch dieser Zugang würde ausschließlich im Rahmen eines Antragsverfahrens erfolgen. Bei dieser Form der Bereitstellung könnte der Datenzugang durch das DIMDI jederzeit entzogen werden, wenn der Verdacht auf nicht vereinbarte Nutzung besteht. Ein entsprechender Zugang würde jedoch die Verfügbarkeit einer geeigneten technischen Infrastruktur beim DIMDI voraussetzen.

Der hier skizzierte Ansatz deckt sich in wesentlichen Punkten mit dem Anonymisierungskonzept [3-4] für die CAMPUS-Files der fallpauschalenbezogenen Krankenhausstatistik (DRG-Statistik) (siehe hierzu [Forschungsdatenzentrum des Statistischen Bundesamtes: Konzept zur absoluten Anonymisierung der fallpauschalenbezogenen Krankenhausstatistik 2010 (DRG-Statistik 2010) zur Erstellung eines CAMPUS Files, Wiesbaden, 2014. Online:
http://www.forschungsdatenzentren.de/bestand/drg/cf/2010/fdz_drg_cf_2010_anonymisierungskonzept.pdf] sowie [Höhne J: Verfahren zur Anonymisierung von Einzeldaten. Statistik und Wissenschaft, Band 16. Statistisches Bundesamt, Wiesbaden, 2010. Online:

https://www.destatis.de/DE/Publikationen/StatistikWissenschaft/Band16_AnonymisierungEinzeldaten_1030816109004.pdf?__blob=publicationFile]).

Fazit und Ausblick

Zum jetzigen Zeitpunkt scheint eine kleine Zufallsstichprobe der sinnvollste Weg zu sein, einen Beispieldatensatz für die Generierung der Auswertungsskripte im Rahmen eines Antragsverfahrens bereitzustellen. Flankierende technische Maßnahmen (personalisierter VPN-Zugang) und vertraglich vereinbarte Auflagen für den Antragsteller werden empfohlen.

Korrespondenzadresse

Peter Ihle
PMV forschungsgruppe
Universitätsklinikum Köln
Herderstrasse 52
50931 Köln
Peter.Ihle@uk-koeln.de
FON +49 221 478 85532
FAX +49 221 478 142 6548