

Leitfaden zur Erstellung eines Scientific-Use-Files (Off-Site) mit Informationen zu ambulanten Behandlungsfällen des Jahres 2002 aus den Stichprobendaten von Versicherten der gesetzlichen Krankenversicherung nach § 268 SGB V

1 Vorbemerkung

Im Jahr 1987 wurde mit § 16 Abs. 6 des Bundesstatistikgesetzes¹ der Wissenschaft ein privilegierter Zugang zu Mikrodaten der amtlichen Statistik eingeräumt. Demnach ist eine Übermittlung von Einzeldaten an die Wissenschaft erlaubt, sofern diese nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft reidentifiziert werden können. „Unverhältnismäßig“ bedeutet in diesem Zusammenhang, dass der Aufwand einer Reidentifikation deren Nutzen übersteigt (faktische Anonymität). Die Deanonymisierung von Einzelangaben in einem faktisch anonymen Datensatz kann nicht mit absoluter Sicherheit ausgeschlossen werden. Das Unverhältnismäßigkeitsgebot impliziert, dass eine Verletzung der Anonymität von Merkmalsträgern nur bei nutzbringenden Zuordnungen gegeben ist.² Damit wird vom Gesetzgeber keine absolute Anonymität mehr vorausgesetzt, sondern eine faktische als ausreichend erachtet. Da dies nur für „Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung“ gilt, wird diese Regelung auch als „Wissenschaftsprivileg“ bezeichnet.³ Das vorliegende Konzept beschreibt die Realisierung eines Scientific-Use-Files (SUF) zu ambulanten Behandlungsfällen von Stichprobenversicherten der gesetzlichen Krankenversicherung für das Jahr 2002.

2 Ausgangsmaterial

Die vorliegenden Stichprobendaten von Mitgliedern und Mitversicherten der gesetzlichen Krankenversicherung wurden speziell für die Analyse relevanter Modelle im Rahmen des morbiditätsorientierten Risikostrukturausgleichs zusammengetragen.⁴ An der umfangreichen Erhebung waren etwa 350 Krankenkassen, die 23 Kassenärztlichen Vereinigungen (KVen) und ihre Verbände, das Bundesversicherungsamt (BVA), die Bundesversicherungsanstalt für Angestellte (BfA) sowie das Deutsche Institut für medizinische Dokumentation und

¹ Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 16 des Gesetzes vom 21. August 2002 (BGBl. I S. 3322).

² Vgl. Höhne, Sturm, Vorgrimler, Konzept zur Schutzwirkung faktischer Anonymisierung, in *Wirtschaft und Statistik*, 4/2003, S. 287.

³ Zur Anonymisierung in der Bundesstatistik vgl. Köhler, S., Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, in: *Forum der Bundesstatistik* Band 31, 1999, S 133-150.

⁴ siehe Reschke u.a. (2005): *Klassifikationsmodelle für Versicherte im Risikostrukturausgleich*; Band 155 der Schriftenreihe des Bundesministeriums für Gesundheit und Soziale Sicherung; Nomos: Baden-Baden.

Information (DIMDI) beteiligt. Die Datenübermittlung erfolgte von den einzelnen Kassenärztlichen Vereinigungen an die Bundesverbände der Krankenkassen. Etwa 90 Prozent der deutschen Bevölkerung sind gesetzlich krankenversichert und bilden folglich die Grundgesamtheit. Die Stichprobenauswahl basiert auf einer Drei-Prozent-Zufallsstichprobe in Form einer Geburtstagsstichprobe. Jede (mit)versicherte Person, die am 11. eines beliebigen Monats eines beliebigen Jahres geboren wurde, mindestens an einem Tag im Erhebungszeitraum in einer der beteiligten gesetzlichen Krankenkassen versichert war und nicht als Auftragsfall geführt wurde, ist in die Analyse einbezogen worden. Die Originaldaten sind nach Satzarten strukturiert und setzen sich aus insgesamt 11 Dateien mit Informationen zu den Jahresdaten der Versicherten, ambulanten und stationären Behandlungsfällen, Arzneimitteln, Arbeitsunfähigkeit und Krankengeld zusammen. Der hier vorliegende Scientific-Use-File enthält jedoch nur Informationen zu den Jahresdaten und ambulanten Behandlungsfällen.

3 Anonymisierungsmaßnahmen

Bei der Erstellung sind Maßnahmen zur Einhaltung der faktischen Anonymität notwendig. Das Originaldatenmaterial besteht aus mehreren Satzarten, die für den SUF teilweise zusammengeführt wurden. Im Ergebnis besteht der SUF zu ambulanten Behandlungsfällen des Jahres 2002 aus insgesamt zwei Dateien – einer Datei ‚Jahresdaten der Stichprobenversicherten für das Berichtsjahr 2002‘, die sich aus Merkmalen der Leitdatei (Leitdatei des Originalmaterials) und der Jahresdaten der Versicherten für das Berichtsjahr 2002 (SA 411 des Originalmaterials) zusammensetzt sowie einer Datei ‚Ambulante Behandlungen‘ mit Informationen zu den ambulanten Behandlungsfällen (Satzart 311 des Originalmaterials) und den daraus resultierenden Diagnosen (Satzart 312 des Originalmaterials).

3.1 Allgemeine Maßnahmen

Löschen einzelner Versicherter aus dem Datenmaterial

Versicherte, die in einem oder mehreren Merkmalen Unplausibilitäten bzw. fehlende Werte aufweisen, werden komplett aus dem Datenmaterial entfernt (siehe Tabelle 1).

Tabelle 1		
Gründe für das Entfernen von einzelnen Versicherten aus dem gesamten vorliegenden Datenmaterial		
Satzart	Merkmal	zu löschende Fälle
311	ef4_311	- Fachgruppe des Arztes = 99 oder leer
	ef5_311	- rechnerischer Ausgabenbetrag = leer
312	ef4_312	- Diagnosenzähler > 50
	ef5_312	- Diagnose = AAA bzw. alle Fälle, die nicht ICD-10-Klassifikation SGB-V Version 1.3 entsprechen - Versicherte, die in Satzart 311, aber nicht in 312 vorkommen
412	ef6_412	- Pflage tage = leer
	ef7_412	- Ausgaben = leer
413	ef4_413	- Diagnosenzähler > 30
	ef5_413	- Diagnose = AAA bzw. alle Fälle, die nicht ICD-10-Klassifikation SGB-V Version 2.0 entsprechen - Versicherte, die in Satzart 412, aber nicht 413 vorkommen und umgekehrt
415	ef4_415	Operationszähler > 30
	ef6_415	Operation = AAA bzw. alle Fälle, die nicht OPS-301 Version 2.0 entsprechen
417	ef9_417	Diagnose = AAA bzw. alle Fälle, die nicht ICD-10-Klassifikation SGB-V Version 1.3 oder 2.0 entsprechen

3.2 Maßnahmen zur Erstellung der Datei ‚Jahresdaten der Stichprobenversicherten für das Berichtsjahr 2002‘

Das Modul ‚Jahresdaten der Stichprobenversicherten für das Berichtsjahr 2002‘ im SUF umfasst die Leitdatei sowie die Satzart 411 des Originalmaterials. Bevor diese beiden Dateien zusammengeführt werden konnten, waren folgende Maßnahmen erforderlich:

Leitdatei

In der Leitdatei befindet sich jeder Versicherte ein Mal. Kriterien für die Auswahl der Stichprobe im Originalmaterial waren:

- Die ausgewählten Versicherten hatten im Jahr 2002 an mindestens einem Tag Krankenversicherungsschutz durch eine der beteiligten Krankenkassen,
- Die Versicherten wurden nicht als Auftragsfall geführt,
- Sie haben ihren Geburtstag an einem 11. eines beliebigen Monats eines beliebigen Jahres.

Beim Merkmal Geburtsjahr (ef2_leit) wird aufgrund geringer Fallzahlen in den älteren Jahrgängen das Bottom-Coding als Anonymisierungsmaßnahme durchgeführt. Merkmalswerte, die unterhalb des Geburtsjahres 1913 liegen, werden auf diesen Wert festgesetzt.

Satzart 411: Jahresdaten der Versicherten

Da sich in der Satzart 411 Versicherte (n=1411) finden, die aufgrund eines Wechsels des Rechtskreises bzw. der Versichertengruppe mehr als ein Mal in den Daten enthalten sind, werden diese aus Gründen der Deanonymisierung aus dem SUF-Material entfernt, sodass im Scientific-Use-File jede versicherte Person nur einmal vorhanden ist. Ebenso weisen die Merkmale ‚Verstorben‘ (ef3_411), ‚Dialyse‘ (ef5_411) sowie Versichertengruppe im Risikostrukturausgleich (ef8_411) bereits im Originalmaterial eher geringe Fallzahlen auf und werden demzufolge nicht in den SUF einbezogen. Weiterhin werden aufgrund der großen Menge an sensiblen Merkmalen im Datenmaterial für die Gewährleistung der faktischen Anonymität der Daten Vergrößerungen durchgeführt. Dies betrifft in dieser Satzart zum einen die Variable ‚Ausgaben für sonstige Leistungen‘ (ef4_411). Es werden sieben Gruppen gebildet:

0:	0 Cent	4:	bis 250 000 Cent
1:	bis 10 000 Cent	5:	bis 500 000 Cent
2:	bis 50 000 Cent	6:	mehr als 500 000 Cent.
3:	bis 100 000 Cent		

3.3 Maßnahmen zur Erstellung der Datei ‚Ambulante Behandlungen‘

Im Originalmaterial liegen zu diesem Modul insgesamt die drei Satzarten ‚Ambulante Abrechnungen‘ (Satzart 311), ‚Diagnosen der ambulanten Behandlung‘ (Satzart 312) sowie ‚Gebührenpositionen der ambulanten Behandlung‘ (Satzart 313) vor. Die im SUF enthaltene Datei setzt sich zusammen aus den beiden erst genannten Satzarten.

Satzart 311: Ambulante Abrechnungen

Die Satzart 311 umfasst je ambulanten Behandlungsfall eines Versicherten einen Datensatz. Um die faktische Anonymisierung zu gewährleisten, werden die Merkmale entfernt, die aus den Erfahrungen in der kontrollierten Datenfernverarbeitung und am Gastwissenschaftsarbeitsplatz heraus für eine Vielzahl von Fragestellungen irrelevant sind bzw. das Risiko der Deanonymisierung erhöhen. Dies betrifft im Einzelnen die Merkmale Sachkosten der ambulanten Abrechnungen (ef6_311) und die Punktzahlsumme (ef8_311) (da beide Variablen Grundlage für die Generierung des Merkmals rechnerischer Ausgabenbetrag (ef5_311) sind) sowie die Variable Leistungsquartal (ef7_311). Einige Fachgruppen des Arztes (ef4_311) weisen zudem sehr geringe Fallzahlen in der univariaten Verteilung auf, so dass hier die Ausprägungen ‚Vorsorgemedizin‘ (21) und ‚Laboratoriumsmedizin‘ (09) zu einer Fachgruppe ‚Sonstiges‘ (26) zusammengefasst wurde. Der rechnerische Ausgabenbetrag (ef5_311) wird gruppiert in folgende Klassen:

- 0: 0 Cent
- 1: bis 1 000 Cent
- 2: bis 2 500 Cent
- 3: bis 5 000 Cent
- 4: bis 10 000 Cent
- 5: bis 50 000 Cent
- 6: mehr als 50 000 Cent.

Satzart 312: Diagnosen der ambulanten Abrechnung

In dieser Satzart befindet sich im Originalmaterial je Diagnose eines ambulanten Abrechnungsfalles aus der Satzart 311 ein Datensatz. Für den SUF wird die Struktur des Materials dahingehend verändert, dass – wie in Satzart 311 – jeder Behandlungsfall einem Datensatz entspricht und die einzelnen Diagnosen eines ambulanten Arztbesuches somit in einer Zeile aufgeführt sind. Dieses Vorgehen ist notwendig, um die beiden Satzarten 311 und 312 aneinanderzuspielen. Es erübrigt zudem die Notwendigkeit der Variable Diagnosenzähler (ef4_312), die daher aus dem Material gelöscht wird. Aus Gründen der Geheimhaltung werden weiterhin für den Scientific-Use-File aus dieser Satzart alle Personen entfernt, die mehr als 50 Diagnosen pro ambulanten Behandlungsfall aufweisen. Dies umfasst 0,03 Prozent aller Fälle des Originalmaterials. Da zudem Informationen über den Gesundheitszustand einer Person als sehr sensibel gelten, wird die Variable Diagnose (ef5_312) in Anlehnung an die ICD-10-Klassifikation in gröbere Klassen eingeteilt.

3.4 Weitere Maßnahmen

Die Satzart 313 „Gebührenpositionen der ambulanten Behandlung“ wird komplett aus dem Material gelöscht. Die hierin enthaltenen Informationen bergen ein hohes Reidentifikationsrisiko und sind erfahrungsgemäß für einen Großteil der Auswertungen von geringer Bedeutung.

3.5 Substichprobenziehung

Da bereits das Ausgangsmaterial mit einer Größe von etwa 2,3 Mio. Versicherten als Stichprobe erhoben wurde und im Rahmen der Anonymisierungsmaßnahmen weitere Personen entfernt wurden, führt dies zu einer Verringerung der Teilnahmekennntnis . Nach Bereinigung aller Satzarten verbleiben 1 634 208 Versicherte im Material, dies entspricht 71 % des Ausgangsmaterials. Aufgrund der Sensibilität der Daten wird dennoch als zusätzliche Maßnahme eine 70-prozentige Substichprobe gezogen. Die verringerte Teilnahmekennntnis wird durch die Ziehung der Substichprobe noch deutlich erhöht.

Um die 70%-Stichprobe zu erhalten, werden 30 Prozent aus dem Material entfernt. Hierbei wird wie folgt verfahren: Zunächst wird das Datenmaterial nach Rechtskreis, Geschlecht und Geburtsjahr sortiert und dann jeder Versicherte mit einer laufenden Nummer versehen. Bei der Ziehung der Stichprobe wird die letzte Ziffer der laufenden Nummer verwendet, die in einem Intervall von 0 bis 9 liegt. Die Auswahlwahrscheinlichkeit beträgt 3 aus 10, sodass im Intervall zwischen 0 und 9 drei Zahlen zufällig ausgewählt werden. Alle Personen, deren Endziffern der laufenden Nummer den ausgewählten Zahlen entsprechen, werden aus dem Material entfernt. Insgesamt enthält die Substichprobe 1 143 945 Versicherte.

4 Fazit

Die durchgeführten Anonymisierungsmaßnahmen führen dazu, dass eine Reidentifikation nur mit unverhältnismäßig hohem Aufwand möglich und mit einer sehr großen Unsicherheit für den Datenangreifer behaftet ist. Das Datenmaterial des Scientific-Use-Files ist faktisch anonym und eine Weitergabe an die Wissenschaft somit unbedenklich. Auch wenn bei der Anonymisierung größter Wert auf den Erhalt des Analysepotenzials gelegt wurde, sind nicht alle Fragestellungen der Wissenschaft exakt mit den Daten analysierbar. Für diese Fälle sei auf die alternativen Zugangswege zu Mikrodaten, die von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder angeboten werden, verwiesen.