Leitfaden zur Erstellung eines CAMPUS-Files aus den Stichprobendaten von Versicherten der gesetzlichen Krankenversicherung nach § 268 SGB V für das Jahr 2002

1 Vorbemerkung

Im Jahr 1987 wurde mit § 16 Abs. 6 des Bundesstatistikgesetzes[1] 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.

Der CAMPUS-File hingegen ist ein absolut anonymisierter Public-Use-File, der speziell für die Lehre an Hochschulen erstellt wird. Seine Funktion besteht darin, die praktische Statistikausbildung mit amtlichen Einzeldaten anzureichern und damit den Hochschulen ein effektives Werkzeug für eine qualitativ hochwertige Lehre zu liefern. Die Schutzmaßnahmen für den absolut anonymisierten File müssen somit weitaus höher angesetzt werden. Das vorliegende Konzept beschreibt die Realisierung eines CAMPUS-Files 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.[2] 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 Information (DIMDI) beteiligt. Die Datenübermittlung erfolgte von den einzelnen Kassenärztlichen Vereini-

[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] Siehe Reschke u.a. (2005): Klassifikationsmodelle für Versicherte im Risikostrukturausgleich; Band 155 der Schriften-reihe des Bundesministeriums für Gesundheit und Soziale Sicherung; Nomos: Baden-Baden.

gungen 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 CAMPUS-File basiert, in Anlehnung an den bereits erstellten und veröffentlichten Scientific-Use-File, auf den Informationen zu den Jahresdaten und ambulanten Behandlungsfällen.

3 Anonymisierungsmaßnahmen

Das Originaldatenmaterial besteht aus mehreren Satzarten, die für den CAMPUS-File teilweise zusammengeführt wurden. Im Ergebnis besteht der CF 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 (Satzart 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	- Pflegetage = 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 CAMPUS-File 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.

Um eine Reidentifikation der Versicherten auszuschließen, wird unter anderem das Merkmal Geburtsjahr (ef2_leit) in folgende Klassen gruppiert:

01: 2002-1998	06: 1977-1973	11: 1952-1948	16: 1927-1923
02: 1997-1993	07: 1972-1968	12: 1947-1943	17: 1922-1918
03: 1992-1988	08: 1967-1963	13: 1942-1938	18: 1917-1913
04: 1987-1983	09: 1962-1958	14: 1937-1933	19: vor 1913
05: 1982-1978	10: 1957-1953	15: 1932-1928	

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 CAMPUS-File entfernt, sodass 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 CF einbezogen. Da der CAMPUS-File speziell für die Hochschulausbildung zur Verfügung gestellt wird, sind für statistisch-methodische Verfahren auch stetige Merkmale notwendig. Aus diesem Grund wird das Merkmal ‚Ausgaben für sonstige Leistungen' (ef4_411) hier nicht gruppiert, sondern mit einem Zufallsfehler überlagert. Hierbei können „...zu den metrischen Variablen eines Datensatzes Zufallszahlen addiert oder die Merkmalswerte mit Zufallszahlen multipliziert werden, so dass die Originalwerte durch die überlagerten Werte ersetzt werden".[3] In diesem Material wurden jeder Beobachtung mit der gleichen Wahrscheinlichkeit Multiplikatoren zwischen 0,91 und 1,09 als feste Größen zugespielt; sowohl der asymptotische Mittelwert als auch die Varianz bleiben somit

[3] Ronning, G./ Sturm, R./ Höhne, J./ Lenz, R./ Rosemann, M./ Scheffler, M./ Vorgrimler, D. (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. Band 4 der Reihe Statistik und Wissenschaft. Wiesbaden: Statistisches Bundesamt, S. 67.

erhalten. Auch die Zahl der Versichertentage (ef7_411) wird dahingehend verändert, dass diese Variable - anders als im SUF – in folgende fünf Gruppen zusammengefasst wird:

1: 1 - 91 Tage
2: 92 - 182 Tage
3: 183 - 273 Tage
4: 274 - 364 Tage
5: 365 Tage.

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 CAMPUS-File enthaltene Datei setzt sich zusammen aus den beiden erst genannten Satzarten.

Satzart 311: Ambulante Abrechnungen

Die Satzart 311 umfasst je ambulantem 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. Die Werte des rechnerischen Ausgabenbetrags (ef5_311) werden - wie das Merkmal ‚Ausgaben für sonstige Leistungen' aus der Datei ‚Jahresdaten der Versicherten für das Berichtsjahr 2002' - in der gleichen methodischen Vorgehensweise mit Zufallszahlen multipliziert. Aufgrund der Einzelfälle im höheren Wertebereich dieses Merkmals werden zuvor alle Fälle mit einem rechnerischen Ausgabenbetrag von mehr als 100.000 Cent aus dem Material entfernt.

Satzart 312: Diagnosen der ambulanten Abrechnung

In dieser Satzart befindet sich im Originalmaterial je Diagnose eines ambulanten Abrechnungsfalls aus der Satzart 311 ein Datensatz. Für den CAMPUS-File 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 bildet der CAMPUS-File jeweils nur die erste Diagnose eines jeden ambulanten Behandlungsfalls ab. Diese wird zusätzlich, in Anlehnung an die ICD-10-Klassifikation, in einer gröberen Klasse ausgewiesen.

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 Teilnahmekenntnis . Nach Bereinigung aller Satzarten verbleiben 1 631 224 Versicherte im Material, dies entspricht etwa 71 % des Ausgangsmaterials. Um eine absolute Anonymität des CAMPUS-Files gewährleisten zu können, wird als zusätzliche Maßnahme eine 0,7%-Substichprobe auf der Basis des Schlussziffernverfahrens gezogen. Zunächst wird das Datenmaterial nach Rechtskreis, Geschlecht und Geburtsjahr sortiert und dann jeder Versicherte mit einer laufenden Nummer versehen. Bei der Ziehung der Substichprobe werden die letzten drei Endziffern verwendet. Die Auswahlwahrscheinlichkeit beträgt 7 aus 1000 oder 1 aus 1000/7. Zunächst wird im Intervall zwischen 0 und 1000/7 eine Zahl Z zufällig ausgewählt. Ausgehend von diesem Startwert Z werden 7 Werte X_i im Intervall von 0 bis 999 nach der Formel:

$$X_i = runden\left[Z + i * \frac{1000}{7}\right], \text{ mit } i = 0,1,2,...6$$

ermittelt. Alle Versicherten mit den Endziffernkombinationen X_i (d.h. 7 aus 1000) werden in die Substichprobe aufgenommen. Insgesamt enthält die Substichprobe im CAMPUS-File 11.419 Versicherte.

4 Fazit

Die beschriebenen Anonymisierungsmaßnahmen führen dazu, dass das Datenmaterial im CAMPUS-File absolut anonymisiert und eine Reidentifikation nicht mehr möglich ist. Neben den durchgeführten Maßnahmen haben die Mikrodaten der gesetzlichen Krankenversicherung zudem ein recht hohes Alter - sie wurden bereits im Jahr 2002 erhoben. Schon die Aufbereitung des Datenmaterials im Jahr 2007 wies Schwierigkeiten in der Beschaffung von zusätzlichen Informationen wie z.B. Schlüsselverzeichnissen auf.

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.