Forschungsdatenzentrum
des Statistischen Bundesamtes

Wiesbaden, den 9. Juli 2014

Betreff: Konzept zur absoluten Anonymisierung der Fallpauschalenbezogenen Krankenhausstatistik 2010 (DRG-Statistik 2010) zur Erstellung eines CAMPUS-Files

I. Vorbemerkungen

CAMPUS-Files sind absolut anonymisierte Mikrodaten, die von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder speziell für die Lehre an Hochschulen entwickelt wurden. Ihre Funktion besteht darin, die praktische Statistikausbildung mit amtlichen Einzeldaten anzureichern und damit den Hochschulen ein effektives Werkzeug für eine qualitativ hochwertige Lehre zu liefern.
Absolut anonymisierte Mikrodaten fallen unter §16 Abs. 1 Pkt. 4 des BStatG[1] und sind vom Geheimhaltungsgebot ausgenommen. Vor einer Weitergabe der Daten als CAMPUS-File in absolut anonymisierter Form muss sichergestellt werden, dass eine Zuordnung der Einzelangaben zu den Merkmalsträgern nicht mehr möglich ist. Im Folgenden werden die vorgenommenen Anonymisierungsmaßnahmen beschrieben, die zur absoluten Anonymität der Einzeldaten führen.

II. Ausgangsmaterial

Die DRG-Statistik bildet die Behandlungsfälle aller Krankenhäuser ab, die nach § 21 Krankenhausentgeltgesetz (KHEntgG) zur Meldung verpflichtet sind. Dies beinhaltet alle Krankenhäuser, die nach dem DRG-Vergütungssystem abrechnen und dem Anwendungsbereich des § 1 KHEntgG unterliegen. Unter anderem werden auch Behandlungsfälle von Krankenhäusern der Bundeswehr erfasst, sofern es sich um die Behandlung von Zivilpersonen handelt. Krankenhäuser der Berufsgenossenschaften tragen ebenfalls zur DRG-Statistik bei, wenn die Behandlungskosten durch die Krankenversicherung und nicht durch die Unfallversicherung getragen werden. In der DRG-Statistik sind jedoch die Behandlungsfälle von Krankenhäusern des Straf- und Maßregelvollzugs sowie von Polizeikrankenhäusern nicht enthalten. Behandlungsfälle von psychiatrischen und psychosomatischen Einrichtungen nach § 17b Abs. 1 Satz 1 zweiter Halbsatz des Krankenhausfinanzierungsgesetzes (KHG) können für Analysen ebenfalls nicht genutzt werden.

Mit der DRG-Statistik stehen dem Datennutzer somit alle vollstationären Krankenhausbehandlungen im DRG-Entgeltbereich in Deutschland für Analysen zur Verfügung. Neben soziodemographischen Merkmalen der Patientinnen und Patienten (z. B. Alter, Geschlecht) werden insbesondere die Erkrankungsart nach Haupt- und Nebendiagnosen, Operationen und Prozeduren, Verweildauer und Fachabteilung sowie Art und Umfang der abgerechneten Fallpauschalen erhoben.

[1] Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 3 des Gesetzes vom 7. September 2007 (BGBl. I S. 2246).

III. Anonymisierungsmaßnahmen

1. Alter des Datensatzes

Eine zentrale Anforderung eines Anonymisierungskonzepts besteht darin, dass der zu anonymisierende Datensatz ein bestimmtes Alter aufweisen sollte, da die Verfügbarkeit von Zusatzinformationen und das Re-Identifikationsinteresse eines potenziellen Datenangreifers mit steigendem Alter der Mikrodaten abnehmen.[2] Der vorliegende Datensatz stammt aus dem Jahr 2010 und ist damit schon mehrfach durch neuere Erhebungen überholt. Das Alter der Daten stellt also ein erhebliches Anonymisierungshemmnis dar.

2. Filtersetzung/Löschen einzelner Fälle

Es wurden nur vollstationäre Patienten in Hauptabteilungen (typ_fall = 1; typ_bereich=1; typ_abt =1) im Datensatz belassen. Des Weiteren wurden nur Patienten aus dem Inland im Datensatz belassen. Fehlende Angaben (pat_land = ohn, pat_land = unb) sowie Fälle mit der Merkmalsausprägung „Ausland“ (pat_land = aus) wurden aus dem Material entfernt.

Fälle mit fehlenden Werten bei der Variable Alter oder der Angabe „unbekannt“ bei der Variable Geschlecht wurden ebenfalls gelöscht.

Patienten mit dem Aufnahmeanlass „Aufnahme nach vorausgehender Behandlung in einer Rehabilitationseinrichtung“ wurden gelöscht. Außerdem wurden Fälle mit dem Aufnahmegrund „Organentnahme“ oder “Wiederaufnahme wegen Komplikationen“ ebenfalls aus dem Datensatz entfernt, da hier ein besonderes Deanonymisierungsrisiko besteht.

Patienten mit unbekannten Fachabteilungen wurden gleichermaßen entfernt. Ebenso wurden Fälle mit ungültigen Diagnosen oder Prozeduren bei ICD-Codes bzw. OPS-Codes gelöscht. Eine Übersicht der gelöschten Merkmalsträger entnehmen sie Tabelle 1.

[2] Vgl. Südfeld, Erwin (1987): Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt, in: Statistisches Bundesamt (Hrsg.): Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Forum der Bundesstatistik, Band 5, Stuttgart: Kohlhammer, S. 148.

Tabelle 1: Entfernen einzelner Fälle aus dem Datenmaterial

Merkmal	zu löschende Fälle
typ_fall	Typ Fall ≠ 1
typ_abt	≠ 1
pat_land	Land des Patienten = ‚aus' (Ausland) Land des Patienten = ‚ohn' (ohne) Land des Patienten = ‚unb' (unbekannt)
alter	Alter = 999
sex	Geschlecht = ‚u' (unbekannt)
aufn_anl	Aufnahmeanlass = ‚R' (Aufnahme nach vorausgehender Behandlung in einer Rehabilitationseinrichtung)
fab_max	Fachabteilung mit längster Verweildauer = 91 (unbekannt) Fachabteilung mit längster Verweildauer = 99 (unbekannt)
icd_hd3	Hauptdiagnose = ‚AAA' (unbekannt)
icd_nd1 – icd_nd10	Nebendiagnose = ‚AAA' (unbekannt)
ops_ko1 – ops_ko10	Prozedur = ‚AAA' oder ‚BBB' (unbekannt)

3. Entfernen von Variablen:

Um die Anonymität der Einzelangaben zu gewährleisten, wurden einige Variablen des Originalfiles nicht in den absolut anonymisierten Datensatz übernommen:

Regionalangaben

Die Variablen *Regierungsbezirk des Instituts* (kh_rb), *Kreis des Instituts* (kh_kreis), *Gemeinde des Instituts* (kh_gem), *PLZ des Instituts* (kh_plz), *Patienten-Regierungsbezirk* (pat_rb), *Patienten-Kreis* (pat_kreis), *Patienten-Gemeinde* (pat_gem) wurden nicht in den CAMPUS-File aufgenommen, da die Gefahr der Deanonymisierung bei dieser feinen regionalen Aufgliederung zu groß wäre.

Nebendiagnosen und Prozeduren

Bei den Variablen *ICD-Code Nebendiagnose* (icd_nd1-icd_nd89) und *OPS-Code* (ops_ko1-ops_ko100) bleiben nur die ersten zehn Nebendiagnosen bzw. Prozeduren (icd_nd1-icd_nd10 bzw. ops_ko1-ops_ko10) im Datensatz enthalten. Alle anderen Diagnosen bzw. Prozeduren wurden entfernt.

Fachabteilungen

Die Variablen *Fachabteilungen* (fab1-fab100) und *Verweildauer Fachabteilung* (tage_fa1-tage_fa100) wurden gelöscht, da die detaillierte Auflistung der Verlegungsketten der Krankenhausfälle innerhalb eines Instituts ein zu hohes Risiko einer Deanonymisierung einzelner Patienten darstellen könnte.

Sonstige Variablen

Ebenso wurden die Variablen *Beteiligung Belegoperateure* (z_bel_oper), *Beteiligung Beleganästhesisten* (z_bel_an), *Beteiligung Beleghebammen* (z_bel_heb), *Abteilungsart* (abt_art1-abt_art100), *DRG-Code* (drg1-drg30) und *entlassender Standort* (*entl_ort)* entfernt. Die Variable *Stundenfall* (std_fall) wurde entfernt, da dieselbe Information in der neu gebildeten gruppierten Verweildauer (typ_vwd) enthalten ist. Das *Institutionenkennzeichen* (ik) wurde ebenso aus den Mikrodaten entfernt sowie die *Fallnummer* (fall_nr) aus den Originaldaten. Die in dem CAMPUS-File enthaltene Variable *Systemfreie Fallnummer* (fall_nr) wurde zufällig generiert. Der im Datensatz verbliebene Merkmalskranz findet sich in dem entsprechenden Schlüsselverzeichnis (siehe Anhang).

4. Vergröberungen der Merkmalsausprägungen von Variablen

Um in univariaten und insbesondere multivariaten Häufigkeitstabellen Mindestfallzahlen gewährleisten zu können, wurden Variablen kategorisiert bzw. die Merkmalsausprägungen bereits kategorisierter Variablen weiter vergröbert.

Ziel der Vergröberungen ist es, multivariate Mindestfallzahlen zu gewährleisten: Die Kombination eines Merkmals mit der Region des Instituts, der Region des Patienten und dem gruppierten Alter des Patienten sollte in der Regel vor der Stichprobenziehung eine Mindestfallzahl pro Zelle von 5.000 aufweisen.

Bundesland

Die Merkmalsausprägungen der Variablen *Land des Instituts* (kh_land) und *Patienten-Land* (pat_land) wurden in den Variablen kh_region und pat_region in drei Gebietseinheiten unterteilt:

1. Nord: Schleswig-Holstein, Hamburg, Niedersachsen, Bremen, Nordrhein-Westfalen
2. Süd: Hessen, Rheinland-Pfalz, Baden-Württemberg, Bayern, Saarland
3. Ost: Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt, Thüringen

Siedlungsstrukturelle Gebietstypen

Die *Siedlungsstrukturellen Gebietstypen* (17er-Gliederung) *der Institute* (kh_typ_gem) *und Patienten* (pat_typ_gem) wurden in Regionstypen (3er-Gliederung in Agglomerationsräume, Verstädterte Räume, Ländliche Räume) umgewandelt. Nähere Informationen finden sich unter:
http://www.bbsr.bund.de/cln_032/nn_1067318/BBSR/DE/Raumbeobachtung/Raumabgrenzungen/SiedlungsstrukturelleGebietstypen/PDF__Download,templateId=raw,property=publicationFile.pdf/PDF_Download.pdf

Alter und Alter gruppiert

Die Variable *Alter* (alter) ist im Datensatz bereits kategorisiert (typ_alter) in 5er-Schritten enthalten. Aufgrund geringer Fallzahlen in mehrdimensionalen Tabellen wurden die bestehenden Altersgruppen weiter vergröbert und in folgende Klassen eingeteilt:

- unter 1 Jahr alt
- 1 bis unter 10 Jahre alt
- 10 bis unter 20 Jahre alt
- 20 bis unter 30 Jahre alt
- 30 bis unter 40 Jahre alt
- 40 bis unter 50 Jahre alt
- 50 bis unter 60 Jahre alt
- 60 bis unter 70 Jahre alt
- 70 bis unter 80 Jahre alt
- 80 Jahre unter älter

Aufnahmeanlass und Entlassungsgrund

Bei der Variablen *Aufnahmeanlass* (aufn_anl) wurden verschiedene Kategorien zusammengefasst (siehe Schlüsselverzeichnis/Anhang). Bei der Variablen *Entlassungsgrund* (entl_grd) blieben die ersten beiden Ziffern erhalten und Ausprägungen wurden ebenfalls weiter zusammengefasst (siehe Schlüsselverzeichnis/Anhang).

Aufnahmegewicht

Die Variable *Aufnahmegewicht* (aufn_gew) wurde in folgende Kategorien gegliedert:

- unter 3000 g
- 3000 g bis unter 4500 g
- 4500 g bis unter 7500 g
- 7500 g und mehr

Verweildauer

Die Variable *Verweildauer* (tage) lag bereits kategorisiert vor (typ_vwd), allerdings in sehr feiner Abstufung. Aufgrund geringer Fallzahlen in mehrdimensionalen Tabellen wurden die Kategorien wie folgt vergröbert:

- Stundenfall
- 1 bis 3 Tage
- 4 bis 7 Tage
- 8 bis 10 Tage
- 11 bis 14 Tage
- 15 bis 21 Tage
- 22 Tage und länger

Die Variable *Längste Verweildauer* (tage_max) wurde entsprechend der Variable *Verweildauer* kategorisiert.

Fachabteilung mit längster Verweildauer

Bei der Variablen *Fachabteilung mit längster Verweildauer* (fab_max) wurden nur die ersten beiden Ziffern behalten. Darüber hinaus wurden die Ausprägungen für Fachabteilungen mit psychiatrischem Behandlungsschwerpunk (28-31) zu einer Kategorie (28) zusammengefasst. Ebenso wurden die Fachabteilungen zur „Frauenheilkunde“ und „Geburtshilfe“ (24-25) unter der Ausprägung gebündelt. „Nuklearmedizin“ (32) sowie „Strahlenheilkunde“ (33) finden sich nunmehr in einer Kategorie (32) wieder.

Diagnose- und Prozedur-Codes

Die Variable *ICD-Code* (icd_hd3) wurde in Anlehnung an die gültige ICD-10GM-Version in Obergruppen zurücktransformiert, ebenso wurden die im Datensatz verbliebenen zehn *ICD-Code Nebendiagnosen* (icd_nd1-icd_nd10) zunächst auf die ersten drei Stellen zugeschnitten und anschließend in Obergruppen zusammengefasst.

Die *OPS-Codes* (ops_ko1-ops_ko10) wurden anhand der gültigen OPS-Version analog ebenfalls auf Basis des Dreistellers in Obergruppen zusammengefasst.

Beatmungszeit

Die ursprünglich metrische Variable *Beatmungszeit in Stunden* (beatm) wurde in folgende Kategorien eingeteilt:

- nicht beatmet
- bis 12 Stunden beatmet
- 13 bis 72 Stunden (= 3 Tage) beatmet
- 73 bis 168 Stunden (= 7 Tage) beatmet
- über 168 Stunden (= 7 Tage) beatmet

5. Institutionskennzeichen

Um eine Identifikation von einzelnen Instituten über die entsprechende Variable – Institutionskennzeichen – vorzubeugen, wird dieses gelöscht.

6. Erlösvolumen

Zu Analysezwecken soll möglichst auch ein stetiges Merkmal im CAMPUS-File belassen werden. Daher wurde beim Merkmal *Case Mix Erlösvolumen* (cm_vol) auf eine Kategorisierung verzichtet. Das bewertete Erlösvolumen wird ermittelt aus dem Produkt der effektiven Bewertungsrelation und dem jeweiligen Landesbasisfallwert (mit Angleichungsbetrag) der behandelten Krankenhausfälle. Zusatzentgelte und nicht mit dem Fallpauschalenkatalog vergütete vollstationäre Leistungen sind in der Berechnung nicht eingeschlossen. Da der Merkmalsträger jedoch keine Informationen zum Bundesland sowie detaillierte Informationen zum gesamten Krankenhausaufenthalt liefert, kann eine Reidentifikation ausgeschlossen werden.

7. Stichprobenziehung

Durch die Stichprobenziehung kann ein potenzieller Datenangreifer nicht mehr sicher sein, ob ein bestimmter Merkmalsträger noch Teil des Datensatzes ist.[3] Somit wird das Risiko einer falschen Re-Identifizierung im Falle eines Datenangriffs deutlich erhöht. Ziel der Stichprobenziehung im vorliegenden Fall ist es, sowohl die Anonymität der Krankenhäuser als auch der Patienten zu wahren. Dazu wurde eine zweistufige Stichprobenziehung unter den Krankenhäusern und deren Behandlungsfällen durchgeführt.

Auf der ersten Stufe wurde eine nach dem Bundesland der Institute geschichtete Zufallsstichprobe unter den Krankenhäusern mit einer Auswahlwahrscheinlichkeit von 0,5 gezogen. Innerhalb der Krankenhäuser wurde dann eine nach der Hauptdiagnosen-Obergruppe geschichtete Zufallsstrichprobe der Krankenhausfälle mit einer Auswahlwahrscheinlichkeit von 0,2 gezogen.

Somit ist gewährleistet, dass aus allen Bundesländern große und kleine Institute ebenso in der Stichprobe vertreten sind wie alle Hauptdiagnosen-Obergruppen die innerhalb eines Instituts auftreten.

IV. Fazit

Die im Konzept beschriebenen Maßnahmen (III. 1 – 7) zur Anonymisierung amtlicher Mikrodaten erfüllen somit das Maß der absoluten Anonymität. In Folge dessen sind die Mikrodaten im CAMPUS-File der Fallpauschalenbezogenen Krankenhausstatistik (DRG-Statistik) des Jahres 2010 absolut anonym.

[3] Vgl. Höhne, Jörg (2010): Verfahren zur Anonymisierung von Einzeldaten, in: Statistisches Bundesamt (Hrsg.): Statistik und Wissenschaft, Band 16, S. 25.