Research Data Centre
of the Federal Statistical Office

Wiesbaden, 9 July 2014

Subject: Concept for Creating a fully anonymised CAMPUS File of the Diagnosis-Related Group Statistics 2010 (DRG statistics 2010)

I. Preliminary remarks

CAMPUS Files are fully anonymised microdata which have been generated by the Research Data Centres of the Federal Statistical Office and the Statistical Offices of the Länder for the purpose of academic teaching at universities. CAMPUS Files function to enrich practical statistical training with official microdata, providing universities with an effective tool for a high quality teaching.

Fully anonymised microdata fall under Article 16, para.1 No. 4 of the BStatG[1] and are excepted from the confidentiality provision. Before transmitting the data as a fully anonymised CAMPUS-File, allocating individual information to statistical units has to be made impossible. The following describes the measures taken to fully anonymise the microdata.

II. Original material

The DRG statistics show treatment cases of all hospitals that are obliged by Section 21 of the German Hospital Fees Act (KHEntgG) to report data. This includes all hospitals invoicing in line with the DRG reimbursement system and fall under Section 1 of the German Hospital Fees Act. Civil patient cases in hospitals of the Federal Armed Forces are also reported. Hospitals of employer's liability insurance associations contribute to the DRG statistics if the treatment costs are covered by the health insurance and not by the accident insurance. However, treatment cases from hospitals of the penal system or police hospitals are not included in the DRG statistics. Likewise, treatment cases from psychiatric and psychosomatic facilities that fall under Section 17b (1), first sentence, second partial sentence of the Hospital Financing Act (KHG) cannot be used for analysis.

Therefore, the DRG statistics enable analysis of all full inpatient treatment cases within the range of the German DRG system. The data cover socio-demographic characteristics of the patients (e.g. age, sex), type of illness (divided into primary and secondary diagnosis), surgeries and procedures, length of stay, the specialist department, as well as type and volume of the accounted flat rate per case.

[1] Law on Statistics for Federal Purposes (Federal Statistics Law – BStatG) of 22 January 1987 (BGBl. I p. 462, 565), as amended most recently by Article 13 of the Law of 25 July 2013 (BGBl. I p. 2749).

III. Anonymisation procedures

1. Age of the data set

The anonymised data set should have a certain age. This is a key requirement of an anonymisation concept, since the availability of additional information and the interest of re-identification for a potential data intruder decrease with increasing age of the microdata.[2] The current data set is from 2010, making it already outdated as several surveys have been conducted more recently. Therefore, the age of the data set is a significant obstacle for de-anonymisation.

2. Filtering/deleting individual cases

Only full inpatients in main departments (typ_fall = 1; typ_bereich = 1; typ_abt = 1) were left in the data set. Furthermore, only patients from Germany were left in the data set. Missing values (pat_land = ohn, pat_land = unb) and cases with the value ”foreign” (pat_land = aus) were removed.

Cases with missing values for the variable “age” or the indication ”unknown” for the variable “sex” were also deleted.

Patients with cause of admission ”admission following treatment in a rehabilitation facility” were deleted. Also, cases with the cause of admission ”organ withdrawal” or ”re-admission because of complications” were removed from the data set because of a special risk of disclosure for these cases.

Patients with unknown specialist departments have been removed from the data set as well as cases with invalid diagnoses or procedures in ICD codes or OPS codes. An overview of the deleted statistical units can be seen in table 1.

[2] Cf. 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, Volume 5, Stuttgart: Kohlhammer, p. 148.

Table 1: Removal of individual cases from the data material

Variable	Cases to be deleted
typ_fall	type case ≠ 1
typ_abt	≠ 1
pat_land	land of the patient = ‚aus' (foreign) land of the patient = ‚ohn' (none) land of the patient = ‚unb' (unknown)
alter	age = 999
sex	sex = ‚u' (unknown)
aufn_anl	cause of admission = ‚R' (admission following treatment in a rehabilitation facility)
fab_max	department with the longest duration of stay = 91 (unknown) department with the longest duration of stay = 99 (unknown)
icd_hd3	primary diagnosis = ‚AAA' (unknown)
icd_nd1- icd_nd10	secondary diagnosis = ‚AAA' (unknown)
ops_ko1- ops_ko10	procedure = ‚AAA' or ‚BBB' (unknown)

3. Removal of variables:

To guarantee the anonymity of individual data, some variables of the original file were not taken into the fully anonymised data set:

Regional data

The variables *administrative region of the institute* (kh_rb), *administrative district of the institute* (kh_kreis), *municipality of the institute* (kh_gem), *postcode of the institute* (kh_plz), *administrative region of the patients* (pat_rb), *administrative district of the patients* (pat_kreis) and *municipality of the patients* (pat_gem) were not included in the CAMPUS File because the risk of de-anonymisation would be too great for this detailed regional breakdown.

Secondary diagnoses and procedures

Concerning the variables *ICD code secondary diagnosis* (icd_nd1-icd_nd89) and *procedures* (ops_ko1-ops_ko100) only the first ten secondary diagnoses or procedures (icd_nd1-icd_nd10 and ops_ko1-ops_ko10) remain in the data set. All other diagnoses or procedures were removed.

Departments

The variables *specialist departments* (fab1-fab100) and *duration of stay in specialist department* (tage_fa1-tage_fa100) were deleted because a detailed listing of the transfer sequences of hospital cases within an institute could involve too high a risk of a disclosure of individual patients.

Other variables

Equally, the following variables were removed: *Participation/*performance *inpatient operator(s) (z_bel_oper)*, *Participation/*performance *inpatient anesthesist(s)* (z_bel_an), *Participation/*performance *inpatient midwife(s)* (z_bel_heb), *department category* (abt_art1-abt_art100), *DRG-Code* (drg1-drg30) and *discharging facility* (*entl_ort)*. The variable *day case* has been removed because this information is included in the newly generated grouped *period of hospitalization* (typ_vwd). The Hospital ID (ik) as well as the original case number (fall_nr) have been removed from the microdata. The anonymised case number (fall_nr) within the CAMPUS File is a newly generated randomised number. The range of variables remaining in the data set is found in the corresponding list of codes (see appendix).

4. Grouping the values of variables

To ensure a minimum number of cases in univariate and especially in multivariate frequency tables, variables have been categorised or values of categorised variables have been further grouped.

The aim of this grouping is to guarantee a multivariate minimum number of cases: As a rule, the combination of a variable with the region of the institute, the region of the patient and the grouped age of the patient should have a minimum number of cases of 5,000 per cell before sampling.

Land

The values of the variable *Land of the institute* (kh_land) and *Land of the patients* (pat_land) were grouped by three regional units for the variables kh_region and pat_region with:

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

Settlement structural types of areas

The settlement structural types of areas of the institutes (kh_typ_gem) and patients (pat_typ_gem), which originally consist of 17 categories, were grouped into three regional types (agglomeration areas, urbanised and rural areas). For more information see:
http://www.bbsr.bund.de/cln_032/nn_1067318/BBSR/DE/Raumbeobachtung/Raumabgrenzungen/SiedlungsstrukturelleG-ebietstypen/PDF__Download,templateId=raw,property=publicationFile.pdf/PDF_Download.pdf

Age and age groupes

The variable *age* (alter) is already contained categorised in the data set, grouped into divisions of 5 years each (*typ_alter*). Because of low numbers of cases in multidimensional tables, the existing age groups were further grouped and divided into the following classes:

- under 1 year old
- 1 to under 10 years old
- 10 to under 20 years old
- 20 to under 30 years old
- 30 to under 40 years old
- 40 to under 50 years old
- 50 to under 60 years old
- 60 to under 70 years old
- 70 to under 80 years old
- 80 years and older

Cause of admission and cause of discharge

For the variable *cause of admission* (aufn_anl) different categories have been grouped (see list of codes/appendix). The first two digits of the variable *cause of discharge* (entl_grd) have been maintained and values have further been grouped (see list of codes/appendix).

Weight at admission

The variable *weight at admission* (aufn_gew) has been divided into the following categories:

- under 3000 g
- 3000 g to under 4500 g
- 4500 g to under 7500 g
- 7500 g and more

Period of hospitalisation

The variable *period of hospitalization* (tage) was available in a categorised form (typ_vwd), but with a very detailed gradation. Due to low numbers of cases in multidimensional tables the categories were grouped as follows:

- day cases
- 1 to 3 days
- 4 to 7 days
- 8 to 10 days
- 11 to 14 days
- 15 to 21 days
- 22 days and more

The variable *longest period of hospitalisation* (tage_max) has been categorised according to the variable *period of hospitalisation*.

Specialist department with the longest period of hospitalization

Only the first two digits of the variable *specialist department with the longest period of hospitalization* (fab_max) have been maintained. Furthermore, the values for departments specialised in "psychiatric treatment" (28-31) have been grouped to one category (28). Departments specialised in "gynecology" and "obstetrics" (24-25) have been combined as well. The values „nuclear medicine" (32) and "radiation medicine" (33) were also grouped together (32).

Diagnosis and procedure codes

The variable *ICD code* (icd_hd3) has been rearranged into main groups according to the current ICD-10GM *classification. The ten ICD codes on the secondary diagnosis (icd_nd1-icd_nd10) have first been reduced to the first three digits and have then been grouped to main groups.*

The OPS codes (ops_ko1-ops_ko10) have been grouped analogously into main groups in accordance with the current OPS version.

Time of respiration

The originally metrical variable *time of respiration* (beatm) has been divided into the following categories:

- not respirated
- respirated for up to 12 hours
- respirated for 13 to 72 hours (= 3 days)
- respirated for 73 to 168 hours (= 7 days)
- respirated for more than 168 hours (= 7 days)

5. Hospital ID

To prevent an identification of individual institutes through the correspondent variable – Hospital ID – this variable is deleted.

6. Case-Mix-Revenue

For analysis purposes a continuous variable should be left in the CAMPUS File. Therefore, the variable *Case-Mix-Revenue* (cm_vol) has not been grouped. The Case-Mix Revenue is calculated by multiplying the effective cost weight by the relevant state-wide base rate of the hospital cases. Additional charges and full inpatient treatments, which are not remunerated by the DRG catalogue, are not included. Because there is neither information about the Land of the institute nor detailed information about the whole period of hospitalization, a disclosure can be ruled out.

7. Sampling

Due to sampling, a potential data intruder can no longer know for sure whether a certain statistical unit is still part of the data set.[3] Therefore, the risk of wrong re-identification is significantly increased in case of a data intrusion. The aim of sampling in this case is to protect the anonymity of both the hospitals and the patients. To achieve this, a two-stage sample was implemented with the hospitals and their cases of treatment.

First, a random sample of the hospitals, which is stratified by the Land of the institute, was selected with a selection probability of 0.5. Within the hospitals a random sample of cases was selected, which was stratified by the main groups of primary diagnosis and had a selection probability of 0.2.

Hence, it is guaranteed that small and large institutes of all Länder as well as all main groups of primary diagnosis occurring within an institute are represented in the sample.

IV. Conclusion

The procedures described in the concept (III. 1 – 7) to anonymise official microdata meet the standards of full anonymisation. Therefore the microdata of the CAMPUS File of the Diagnosis-Related Group Statistics 2010 (DRG Statistics 2010) is fully anonymised.

[3] Cf. Höhne, Jörg (2010): Verfahren zur Anonymisierung von Einzeldaten, in: Statistisches Bundesamt (Editor): Statistik und Wissenschaft, Volume 16, p. 25.