Statistical Journal of the United Nations ECE 21 (2004) 23–32
IOS Press

Record linkage of hospital discharge register with population register: Experiences at Statistics Netherlands

Agnes de Bruin*, Jan Kardaun, Fred Gast, Edie de Bruin, Marije van Sijl and Gerard Verweij
Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg, The Netherlands

Abstract. This paper describes the first-time linking by Statistics Netherlands of the hospital discharge register (HDR), held by Prismant, with the population register (PR). This linkage is the first step in building an integrated population-based health statistics dataset at individual level, using existing register data as much as possible. The linkage of the HDR with the PR has proved to be successful: 87.6% of the HDR records could be uniquely linked to a person record in the PR, 8.7% were multiply linked, and 3.6% of the records could not be linked. The multiple linkages are a natural result of the limited resolution power of the linkage key (date of birth, sex, postal code). The 3.6% non-linkages are also plausible, given a rate of approximately 2% administrative errors in both registers. The quality of the linkage was further studied by estimating the level of mismatches (false positive linkages): this rate was estimated to be in the order of magnitude of approximately 1%, and in any case less than 3%. It was concluded that the linkage of the Dutch HDR with the PR is of good quality and that it forms an adequate basis for statistical analyses.

Keywords: Record linkage, registers, hospital admissions, population data, methods, quality

1. Introduction

The development of a population-based health statistics dataset is a part of a strategic research project at Statistics Netherlands. The aim is to build a system of coherent information on use of medical services and health status by linking the available national data sources, i.e. medical registers, registers with socio-economic data, and survey data. Making maximum use of existing data sources reduces the need for primary data collection by means of surveys. Moreover, it is expected that linking the available sources will create an important asset, as it enables integrated analysis of (professional) health data and socio-economic data. In addition it is possible to do follow-up studies of population groups in time, and thus produce longitudinal statistics. Statistics Netherlands is entitled by law to use and link micro data from registers, but only for statistical purposes and with stringent obligations to ensure data confidentiality.

The Dutch population register (PR) is used as the backbone of the linkages with other person-related data sources. The PR contains information on all residents of the Netherlands, and has been available electronically since 1995. Statistics Netherlands receives micro data (at person level) from the PR for a set of demographic variables, such as birth, death, registered address, country of birth, and familial

*Corresponding author: Agnes de Bruin. Tel.: +31 70 337 5299; Fax: +31 70 337 5987; E-mail: abun@cbs.nl.

relations [1]. For the purpose of linkage, Statistics Netherlands includes these data with the individual changes in a longitudinal database.

With regard to the health data sources, Statistics Netherlands has kept the causes of death register (since 1901) and has conducted the national health interview survey since 1981. Both these sources are linked to the PR from 1995 onwards. However, with a view to building a more comprehensive health statistics database, Statistics Netherlands has started to explore external data sources that may be used for this purpose. The national hospital discharge register (HDR) (*Landelijke Medische Registratie*), maintained by Prismant in Utrecht is the first external register we explored, because of the economic importance of this aspect of health care and because the register has a high coverage rate. The HDR contains data on hospital admissions, covering all general and university hospitals and most specialised hospitals. Information is collected on both in-patients and day patients. The information concerns administrative patient data, admission and discharge data, diagnoses, surgical procedures, and the medical specialties concerned [2]. The term 'hospital discharge register' refers to the fact that the patients who have been *discharged* are registered, i.e. in the HDR 1995 admissions of patients who were discharged from hospital in 1995 are registered.

This paper describes in short the methodological aspects of the linkage of the PR with the HDR; a more extensive description has been published elsewhere [3]. The data are linked at the micro-level, i.e. the level of individual records. The applied linkage procedures are described, as well as the results and quality of the linkage.

2. Methods

2.1. Preliminary processing of HDR data

1995–2001 data were selected from the HDR. These files were first adjusted to a uniform population of hospitals and admissions. Admissions included all registered admissions, with the exception of day patient care for childbirth, psychiatric treatment and rehabilitation treatment, as these categories are not supposed to be included in the HDR. The hospitals included all general and university hospitals, and specialised hospitals with the exception of epilepsy clinics and long-stay centres for rehabilitation and asthma treatment. Not all the hospitals included submitted complete micro data; the coverage is presented in Table 1. The relatively low coverage of the admissions in specialised hospitals is mostly caused by 1–2 (number depends on the year) cancer clinics that did not register records at the micro-level in the HDR at all. If these clinics are excluded the average coverage in 1995–2001 of the specialised hospitals is over 97%.

Furthermore, the records of patients admitted to Dutch hospitals but not resident in the Netherlands were excluded from the HDR data, as Statistics Netherlands' population statistics are based on the resident population, registered in the Dutch population register.

2.2. Preliminary processing of PR data

For the purpose of record linkage, Statistics Netherlands has created a *central linkage file of persons* (CLFP), a longitudinal file containing persons registered in the PR. The CLFP starts on 1 January 1995 and is updated to the year before the current calendar year. In the CLFP the person identifier of the PR is replaced by a meaningless, but unique, number, the *person record identification number* (PRIN). Changes in demographic attributes (e.g. address, marital status, emigration) are processed by including

Table 1
Percentage of total admissions of which HDR micro-data were available, by hospital category

	General hospitals	University hospitals	Specialised hospitals[1)]
1995	99.7	96.0	56.7
1996	99.5	97.7	58.6
1997	100.0	97.8	55.8
1998	99.8	97.1	68.9
1999	99.9	100.0	66.7
2000	99.6	100.0	57.2
2001	99.5	100.0	68.6
Total number of hospitals (2001)	8	97	10

1) Excluding rehabilitation, epilepsy and asthma clinics.

additional records of the person concerned in the CLFP, indicating the PRIN, the serial number of the record, period of validity, and the nature of the change.

For linkage with the HDR, the CLFP valid from 1995/01/01 to 2002/01/01 was used.[1]

2.3. Linkage procedures

Because the HDR was originally designed to be anonymous, only limited linkage variables were available: date of birth, sex, and truncated postal code (4 digits). For a quarter of all cases (admitted patients) the full postal code (6 positions: 4 digits and two letters) was available, however. The individual HDR records were linked step by step to the PR on the primary linkage key date of admission, date of birth, sex and postal code (full code in step I and truncated code in step II). On the multiply linked records of the first two steps, a third linkage step (III) was conducted on date of death, for the patients who died in hospital.

A range of 30 days both ways was tolerated for the date of admission, in order to allow for administrative differences between the registers in recording dates of address changes.

In the linkage n HDR records ($n \geqslant 1$) were linked to m PR records ($m \geqslant 0$). When $m = 0$, no linkage could be made, when $m = 1$ a unique linkage was made, and when $m > 1$ there were multiple linkages. In the case of multiple linkages the respective n HDR records were copied m times, and the data of the m PR records were added to each of the HDR records. The resulting file after linkage thus consisted of $n \times m$ records.

When a unique or multiple linkage was made, several PR data were added to the records, such as the PRIN, the 'distance' of the date of admission and the validity date of the PR record (because of the $+/-$ 30 days tolerance), the value of n and m, etc.

Apart from the three-step linkage procedure described above, a single-step linkage was also performed on the truncated postal code (first 4 positions only), date of birth and sex. This was done because for some analyses it may be preferable to use data that are obtained from a uniform linkage process.

3. Linkage results

Table 2 shows the results of the three-step linkage of the HDR with the PR for 1995–2001.

[1]For the sake of convenience the CLFP is further referred to as the PR, although it is in fact a derivative and selection of the population register.

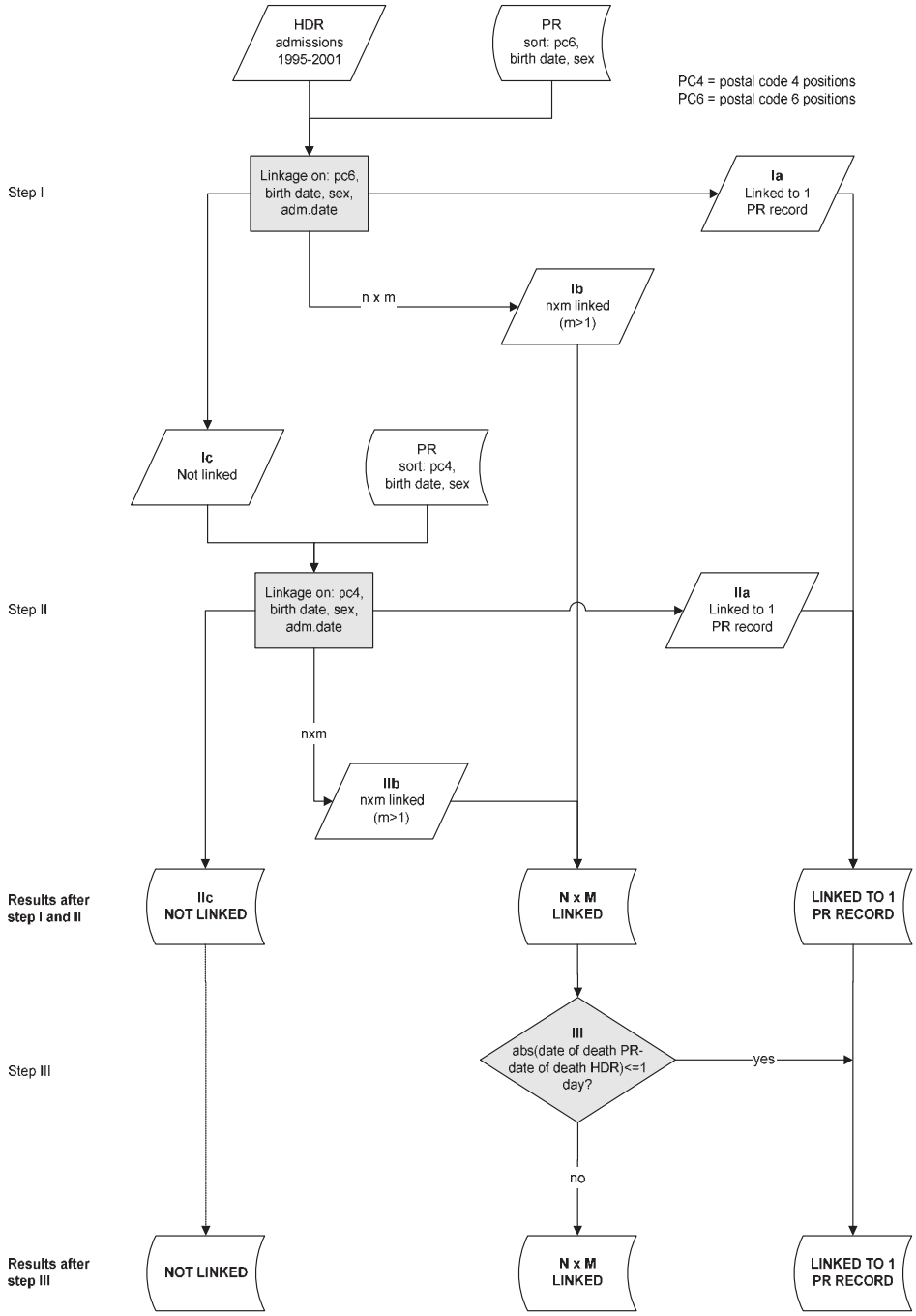

Fig. 1. presents an overview of the linkage process.

The relatively low percentage of unique linkages in the first step is a logical result of the fact that a complete postal code of 6 positions could be found in only approximately one quarter (4, 426, 463 out of 16, 115, 612) of the records. Therefore the non-linkages after step I are mostly (97.4%) HDR records

Table 2
Results of linkage of HDR to PR after linkage steps I, II and III

	number	%
Total number of HDR records (1995–2001)	16,115.612	100.0
Step I Linkage on postal code 6 pos.+birth date+sex+admission date[1)]		
Ia. linked to 1 PR person	4,086.747	25.4
Ib. linked to >1 PR persons	29.172	0.2
Ic. not linked (proceed to step II)	11,999.693	74.5
Step II Linkage of Ic on postal code 4 pos.+birth date+sex+admission date[1)]		
IIa. linked to 1 PR person	10,018.374	62.2
IIb. linked to >1 PR persons	1,394.213	8.7
IIc. not linked	587.106	3.6
Results after Step I and Step II		
linked to 1 PR person (uniquely linked)	14,105.121	87.5
linked to >1 PR persons (multiply linked)	1,423.385	8.8
not linked	587.106	3.6
Step III Linkage of Ib and IIb on date of death PR = date of death HDR $+/-$ 1 day		
Results after Step III		
linked to 1 PR person (uniquely linked)	14,121.787	87.6
linked to >1 PR persons (multiply linked)	1,406.719	8.7
not linked	587.106	3.6

1) with tolerance of $+/-$ 30 days

with a truncated postal code of 4 positions. These were forwarded to the linkage in step II. After two steps the proportion of unique linkages amounted to 87.5%. The additional linkage on date of death increased this number by only 0.1%. This is because most persons who died in hospital had already been linked in the first two steps; the linkage on date of death is only applied to the 8.8% multiple linkages after step II.

The final result after three steps is that 87.6% of the HDR records were uniquely linked, 8.7% were multiply linked and 3.6% could not be linked to the PR. The 8.7% multiple linkages are a natural result of the limited resolution power of the linkage key, as was verified by linking the PR with itself using the same key.

It was further found that 95.1% of the records with a full postal code were uniquely linked after two linkage steps. Therefore, if all hospitals were to register full postal codes in the HDR, an ultimate linkage result of approximately 95% could be achieved.

Table 3 presents the results of the linkage on the truncated postal code (4 positions), date of birth and sex. This basic linkage produced 84.1% unique linkages; a difference of 3.4 percent points with the previously presented results after steps I and II. The percentage of multiple linkages increased by the same amount: these are the records that link uniquely on the full postal code, but not on the truncated code. Obviously, the percentage of non-linkages remained the same.

4. Quality of the linkage

In studying the quality of linkages, it is important to focus both on the level of missed linkages or false negatives, i.e. the erroneously not-linked records, and on the level of false positive linkages, i.e. the erroneously linked (mismatched) records. Administrative errors in registers will mostly lead to missed linkages, but may sometimes also lead to false positive linkages. For example, in the HDR-PR linkage described here, it can be argued that there are approximately 85 (age) $\times$ 365 (days) $\times$ 10^4 (4 positions

Table 3
Results of linkage of HDR to PR on only truncated postal code (4 pos.), date of birth, and sex

	abs.	%
Total number of HDR records (1995–2001)	16,115.612	100.0
Linkage on postal code 4 pos.+birth date+sex+admission date[1)]		
linked to 1 PR person (uniquely linked)	13,561.014	84.1
linked to >1 PR persons (multiply linked)	1,967.492	12.2
not linked	587.106	3.6

1) with tolerance of $+/-$ 30 days

Table 4
Characteristics of non-linked HDR records compared with the other HDR records (1995–2001)

	Non-linked records	Other records
Number of records (total n = 16,115.612)	587.106	15,528.506
Average age (yr)	42.8	45.8
Men (%)	48.2	45.6
Day patients (%)	24.9	34.5
Average duration of all hospitalisations (days)	10.0	6.1
Average duration of in-patient hospitalisations (days)	13.0	8.8

postal code) $\times$ 2 (sex) = 620 $\times$ 10^6 combinations of the linkage key date of birth, truncated postal code, and sex. Given a population in the Netherlands of approximately 16 $\times$ 10^6, the probability of one random record forming a false positive linkage is about 1:40.

The quality of the HDR-PR linkage was studied by estimating the possible level of false negative and false positive linkages on the basis of analyses of the linked data and the available literature.

4.1. False negative linkages

The level of missed linkages cannot be determined with certainty. However, a study of the quality of the address data in the PR [4] shows between 1% and 2% administrative errors in postal codes. Furthermore, for 0.4% of the persons recorded in the PR, the day and month of birth were unknown, which also leads to missed linkages. Assuming that the combined administrative error in the main linkage variables (postal code, date of birth and sex) is about 2% in the PR, and assuming that this level is the same in the HDR (no data available), then there could be approximately 4% missed linkages. This estimate corresponds reasonably well to the proportion of non-linkages actually found (3.6%). Moreover, there will also be cases in this 3.6% that justifiably do not link to the PR because these patients are not registered in the PR for some reason, for example because of their short stay or because they are illegal inhabitants. Therefore, the proportion of 3.6% non-linkages is certainly no higher than what could be expected on the basis of the level of administrative errors in the registers.

Because the proportion of non-linkages is reasonably low, their effect on the statistical analyses outcome will be generally small, unless the non-linked records are a very selective group and the statistical analysis focuses on variables related to this selectivity. In order to study the possible magnitude of this problem, the non-linked HDR records were compared with the other HDR records (uniquely or multiply linked) using a number of characteristics as available in the HDR. The results are given in Table 4.

The group of non-linked HDR records generally appeared to have a slightly lower average age (caused in fact by an overrepresentation of infants, young adults and the very old), a higher proportion of males, a lower percentage of day patients and a longer average duration of hospitalisation. Further analyses showed that these patients also more often lived in the urban areas in the west of the Netherlands. Also

Table 5
HDR-PR linkage results on full postal code compared to those on truncated postal code of HDR records (1995–2001) with a valid full postal code

	Linkage on truncated postal code (4 pos.), date of birth and sex							
	Not linked		Uniquely linked		Multiply linked		Total number of HDR records	
	number	(%)	number	(%)	number	(%)	number	(%)
Linkage on full postal code (6 pos.), date of birth and sex								
Not linked	166.863	(100)	113.642	(3)	19.280	(3)	299.785	(7)
Uniquely linked	x	(x)	3,542.549	(97)	544.103	(92)	4,086.652	(93)
Multiply linked	x	(x)	x	(x)	29.436	(5)	29.436	(1)
Total number of HDR records	166.863	(100)	3,676.191	(100)	592.819	(100)	4,415.873	(100)

'x': cannot occur on logical grounds.

there was some overrepresentation of psychiatric disorders and injuries from accidents. So it may be concluded that the non-linked group is indeed a somewhat selective group, but also that the differences are not very large. As the group of linked HDR records is much larger than the non-linked group, the differences are much smaller there when compared to all records. Thus, given the limited degree of selectivity and magnitude of the non-linked group, it is expected that the effect on most statistical analyses will be small, especially when the focus is on relations in the total group of linked records. However, biases may occur if the analyses focus on certain specific population groups.

4.2. False positive linkages

The possible magnitude of false positive linkages was studied in several ways. First, by comparing the linkage results of the HDR records containing a full postal code with the corresponding results when only the truncated postal code was used. This comparison allows an estimation of the number of false linkages as a result of errors in the postal code to be made, which is the most 'vulnerable' of the linkage variables used. Secondly, we looked at the extent to which the linked records had consistent mortality data in the HDR and PR. This could only be done with the mortality variable because only this variable was present in both registers and was not used in the primary linkage key (i.e. in steps I and II of the linkage). Thirdly, the extent of overlapping admissions of the same person in time was studied in the linked records, as this too may give an indication of false linkages.

4.2.1. Possible false linkages because of postal code errors

The linkage results of the HDR records with a full postal code linked on this full postal code (and sex and date of birth) were compared with the results when the same records were linked on the truncated postal code. Under the assumption that the records with a full postal code (6 positions: 4 digits and 2 letters) are representative of those with only a truncated code (only 4 digits), the effect of neglecting the 2 letters of the postal code can be assessed. The results are given in Table 5. For this comparison only the records with a potential valid full postal code, consisting of 4 digits and 2 letters were, selected.

The column percentages of Table 5 shows that 3% of the records that were uniquely linked on the truncated postal code were not linked on the full code. These records are considered to be uniquely linked, while they are not according to the full postal code. In cases where there is an error in the 2 letters of the postal code, but the 4 digits of the (truncated) code are correct, the resulting linkages can still be considered to be correct. However, in other cases these records are likely to have been falsely linked. As the ratio of the falsely and correctly linked records is not known, it can be concluded on the basis of the postal code data that 3% is the upper estimate for false linkages and that the actual proportion will be less than 3%.

Table 6
Differences in date of death between HDR and PR of uniquely linked records (1995–2001) in which death was registered in the HDR

	number	%
No death according to PR	3.385	1.15
− 366 days or more	99	0.03
− 91 to −365 days	32	0.01
− 61 to −90 days	20	0.01
− 31 to −60 days	39	0.01
− 15 to −30 days	47	0.02
− 8 to −14 days	63	0.02
− 7 days	14	0.00
− 6 days	17	0.01
− 5 days	30	0.01
− 4 days	76	0.03
− 3 days	128	0.04
− 2 days	257	0.09
− 1 day	1.788	0.61
no difference	285.183	96.77
+ 1 day	3.353	1.14
+ 2 days	63	0.02
+ 3 days	30	0.01
+ 4 days	8	0.00
+ 5 days	1	0.00
+ 6 days	2	0.00
+ 7 days	4	0.00
+ 8 to 14 days	21	0.01
+ 15 to 30 days	28	0.01
+ 31 to 60 days	12	0.00
+ 61 to 90 days	3	0.00
+ 91 to 365 days	5	0.00
Total	294.708	100.00

4.2.2. Consistency of mortality data

The linkage results of steps I and II (linkage on postal code, date of birth and sex) were used to compare the mortality data. According to the HDR data, the admitted patient had died in hospital in 2.1% of the total number of uniquely linked records. For these cases the date of death according to the HDR was compared with the date of death of the linked person in the PR. The results are presented in Table 6.

When a patient dies during a stay in hospital, the date of death in the HDR corresponds with the date of discharge. As the linkage with the PR is timed at the date of admission, with a tolerance of $+/-$ 30 days, it is only possible to have an earlier date of death in the PR than in the HDR in the time range of the 'admission date minus 30 days' up to the 'discharge date'. As a consequence, it is logical that in Table 6 extreme differences in date of death between HDR and PR occur less frequently in the case of a positive difference (earlier death in the PR than in the HDR) than in the case of a negative difference (death is on a much later date in the PR than in the HDR).

From Table 6 it can be concluded that in 96.8% of the cases the date of death in the HDR corresponds exactly with that in the PR. If we tolerate a difference of $+/-$ 1 day, which seems reasonable in connection with possible differences in administrative notification of the same death in the two registers, the correspondence rises to 98.5%. It can therefore be concluded that there was a substantial difference ($>$ 1 day) in the date of death between HDR and PR in only 1.5% of the cases. In 1.2% of the cases the person had not been recorded as dead in the PR in the studied time frame. The 1.5% of cases with

Table 7
Overlapping admissions in time of uniquely linked HDR records (1995–2001)

Total numer of HDR records	14,121.787	(linked to 6,335.793 persons in PR)
Overlapping admissions; of which	17.024	(linked to 8.427 persons in PR)
2 overlapping admissions	16.532	
3 or more overlapping admissions	492	
Proportion of overlapping admissions	0.12%	

a substantial difference in date of death may be caused by typing errors in the registration of the date of death and/or by false linkages. In the first case there may still be a correct HDR-PR link. However, these different causes cannot be further quantified on the basis of the available data. Thus, the proportion of 1.5% that is indicative for false linkages on the basis of the mortality criterion, should again be seen as an upper limit.

4.2.3. Overlapping admissions in time of the same person

The last aspect we studied to get an indication of the number of false linkages was the occurrence of overlapping HDR admissions in time, linked to the same person in the PR. Only admissions that overlapped by more than 1 day were taken into account, as administrative rules and practices in hospitals may legitimately result in a person being for example discharged from one hospital and admitted to another on the same day, or having day treatment in hospital immediately followed by an in-patient admission on the same day. In such cases there is an administrative overlap in admission periods of 1 day.

From Table 7 it can be concluded that only 0.12% of the uniquely linked HDR records overlapped by more than 1 day with another HDR-record linked to the same person. Moreover, further analysis of these overlaps showed that a considerable number did not directly point in the direction of false linkages. This was deduced among other things from the substantial overrepresentation of admissions with diagnosed cardiovascular diseases (+ 24%) and of diseases of the pesinatal period (+ 15%) compared with the distribution of diagnoses in all uniquely linked HDR records. These are typically diseases where transfers of patients from one hospital (department) to another are relatively frequent. One would assume that overlapping admissions indicative of false linkages would have no relation to the diagnosis. This is apparently not the case in up to roughly 40% of the overlapping records. If these records are excluded from the 0.12% overlapping admissions, the remaining percentage of overlapping admissions that may indeed point to false linkages drops to well below 0.1%.

4.2.4. Conclusion false positive linkages

Reviewing the estimates of false positive linkages found on the basis of postal code ($< 3\%$), mortality data ($< 1.5\%$), and overlapping admissions ($< 0.1\%$), it can be concluded that the overall proportion of false positives will be probably in the order of magnitude of 1%, and in any case less than 3%.

5. Conclusion

The linkage of the Dutch HDR with the PR has shown to be successful. 87.6 per cent of all the HDR records could be uniquely linked to a person record in the PR, which is a satisfactory result given the resolution power of the available linkage variables. If all records were registered in the HDR with a full postal code (three-quarters of the records now only have a truncated code), then approximately 95% unique linkages would be possible.

Given the low estimates of the possible false negative and false positive linkages, it is concluded that the linkage is of good quality and forms an adequate basis for statistical analysis.

Acknowledgements

Statistics Netherlands thanks the organisation Prismant (Utrecht) for making available the HDR data and for their cooperation in this study.

References

[1] C.J.M. Prins, Dutch population statistics based on population register data, *Monthly bulletin of population statistics* (*Statistics Netherlands*) **48**(2) (2000), 9–15.
[2] A.C.M. Blankendaal, E. van Deyl and M.A. Dutrée, *New Healthcare Information presents enormous challenge*, Health Information in the Netherlands (Dutch Association for Medical Administration), November 2003, pp. 62–65.
[3] A. de Bruin, E.I. de Bruin, A. Gast, J.W.P.F. Kardaun, M. van Sijl and G.C.G. Verweij, *Record linkage of data from Hospital Discharge Register and Population Register*, Methods, results and quality, Projectteam for development of a population-based Health Statistics Dataset, Statistics Netherlands, Voorburg/Heerlen, 2003.
[4] Agency for basic administration of personal data and travel documents (BPR), The quality of the address data in the municipal basic registration of population data, Dutch Ministry for the Interior, 's-Gravenhage, September 2002.

Agnes de Bruin is statistical researcher at the Department of Personal Data Registers, Division of Social and Spatial Statistics, Statistics Netherlands. She is currently working as a project manager for the development of a Health Statistics Dataset based on the linkage of medical registers with population data. Her former work experience at Statistics Netherlands is also in the field of health statistics, in particular the international harmonization of national health interview surveys.