The determination of the total number of hospitalized traffic victims by comparison of police and hospital reports.

Peter Polak and Siem Oppe
SWOV, September 1997.

1. Introduction.

This paper is completely based on research carried out by Dr. P.H. Polak, as reported in the SWOV-report R-97-15: Registratiegraad van in ziekenhuizen opgenomen verkeersslachtoffers (in Dutch only).
This paper describes only the methods used and the technical outcomes. Results regarding the implications of the outcomes in terms of safety effects, the representativity of national safety statistics and the extent of under-reporting will be reported separately.

In the Netherlands, as in many other countries, the number of fatalities is well known, but the number of hospitalized road traffic victims is not. A recent estimate of the degree of completeness of the police reporting for this group in the Netherlands is 60%.
In order to check this global indication, a comparison has been made between the records in two well-established databanks: the AVV/BG- databank of police-reported traffic victims, which is the official Dutch database for traffic accidents, and the SIG-databank, which is a database from the Ministry of Health. The first one covers only road traffic victims, but more than hospitalized ones (our target group), the second covers only hospitalized patients, but more than road traffic victims. The first one is incomplete, the second (almost) complete. However, it is not possible to earmark all traffic victims in the second database. Therefore, a study has been carried out, aiming at the following four subgroups of our target group:
the hospitalized road traffic victims that are reported in both databanks, the remaining victims in each databank that belong to the target group and those victims that are missing in both databanks. The possibilities are given in Table 1.

Hospitalized victims	In SIG-database	Not in SIG-database	No target population
In AVV/BG-database	In both databases	only in AVV/BG-database	Not hospitalized
Not in AVV/BG-database	Only in SIG-database	Not in one of the two databases	
No target population	No traffic accident		

Table 1. *Distribution of hospitalized victims, according to presence in databases and belonging to the target population.*

To make this classification, records of victims in both databanks have to be matched on a record by record base. This asks for a unique characteristic that is present in both databanks. In some situations this can be done directly. E.g. in Denmark a personal ID-number is registered in comparable databanks. In the Netherlands such a key-number is not recorded. Matching is therefore based on a (small) number of key-variables that have been registered in both databanks. These key-variables as such are not unique, but it is assumed that the combination (almost always) is. Because there are errors in both databanks, matching is not easy. Therefore, a distance has been defined between records on the basis of the selected matching-keys. If all key-values agree completely, the distance between the records is zero. If some values differ, than the distance is larger than zero; the larger the disagreement, the larger the distance. Distances are computed for all records from the first database with all records in the second database.
Two values are used for the matching process: the distance and a measure of selectivity. If a pair of two records has a certain distance and each of the two records has a considerably higher distance to all records in the other database, than the selectivity for that pair of records is high. Using this matching strategy, matched records can be graded, according to their probability of correctness.
The total number of matches for the two databases was estimated from the matched pairs, using the probability of correct matching. Together with information from the other subgroups of Table 1. an estimate has been computed, not only for the amount of under-reporting of hospitalized road traffic victims in total, but also of the amount of bias regarding factors such as travel mode and region. It was shown that this bias was in some cases substantial and should lead to corrected statistics.

2. Definitions.

As said before, the major aim was to estimate the total number of hospitalized traffic victims. A secondary aim was, to combine the information from both databanks, in order to answer questions that could not be answered on the basis of the separate databanks. The AVV-BG databank has only scarce information about the injuries, while the SIG-databank has only scarce information about the accident. Both aims lead to different matching criteria. For the first aim completeness of the matched outcome set is important, while for the second aim unbiasedness is is of special interest. If the set of matched records is biased, at least information about the extent of it with regard to the most important variables is necessary. The quality of matching is therefore an important characteristic next to the estimated total number of victims.

In order to define a unique key for matching, the following four characteristics have been used:

- date of birth
- sex
- time and date of accident resp. hospitalization
- hospital

In previous research this combination turned out to be a good basis for matching. Direct matching on the basis of this information is not optimal, because of registration imprecision or errors and uncertainty about the time between the accident and hospitalization. Furthermore, some data is missing.

Not all inconsistencies, deletions or errors have the same importance. E.g., the number of response classes of a variable (two in case of sex, several in case of the hospital), the ease of registration and registration- and checking procedures play a role. Therefore, a weighted measure of discrepancy between the information on the key-variables in both sets is to be recommended. In many areas of research, e.g. in psychology, distance models are used to describe similarities and dissimilarities between sets of characteristics. Objects (persons, opinions on political parties, preferences for consumer goods etc.) and their relations are represented by points in a (multidimensional) space. If two objects are more similar to each other than two other objects, then their distance in that representation space should be smaller.

The characteristics need not be quantitative measures themselves, but are often quantified by the researchers. Apart from the weighting of each separate characteristic, also the classes of each characteristics must be quantified, in order to compute a distance. Some techniques are using the basic information (nominal, ordinal or metric in nature) and look for a joint quantification of weights and classes, such that a representation of the similarities or dissimilarities is found in a reduced representation space. In our case, we have used a quantification on the basis of heuristic rules, during a process of matching of a subset "by hand" and of an analysis of the errors in the data.

The distance function can be written as follows:

$$D = \sum_i c_i \; \delta(\alpha_i,\beta_i)$$

In this formula c_i is the weight for each pair of matching variables i. $\delta(\alpha_i,\beta_i)$ the distance between the two classes within the pair. The final distance D is the sum of the weighted distances δ_i.

Pairs that match completely should have a distance of zero. However, the match between time and date of the accident and of hospitalization are always different. Therefore, some relaxation of the criterion is used. (Narrow) margins are selected for a zero distance between these characteristics.

If α and β are both known and equal, δ is zero. If sufficiently different, then δ is one. If the difference is small, the value is between zero and one and depending on the difference. In case α or β is unknown also some value between zero and one is chosen.

The value of δ_i depends on information about errors and differences between values on the

variable i for pairs that were matched with certainty, during the matching by hand. Those pairs that agreed on all characteristics but one, were used to specify the δ_i-value. The sets investigated consisted of the data for one quarter of a year. For the AVV-BG databank 6700 records were used and for the SIG-databank 5679.
The value of c_i depends on:

- the probability of an error
- the resolution of the variable (the number of possible values)
- the distribution over the possible values

Table 2. shows the differences in the scores on the variable 'epoch' (the combination of time and day of the accident or hospitalization), for the matching 'by hand'.
In both databases the probability of a male is approximately 0.68. Therefore, the probability of a random match on sex is $0.68^2 + 0.32^2 = 0.56$. For hospital the probability of such a random match is 0.01.
As an example of the possible effects of random matching on the results, the following calculation that is based on these values of the four key-variables gives an insight. The estimated probability of a random (perfect) match on the basis of the four characteristics is 0.56 x 0.01 x 0.003 x 0.02 = 0.00000034. For the two test sets the expected number of random matches amounts 6700 x 5679 x 0.00000034= 12.9 matches.

Epoch-difference	Number
1-90 days negative	3
0-1 dag negative	27
0-1 hour positive	189
1-2 hour positive	545
2-3 hour positive	558
3-4 hour positive	419
4-5 hour positive	149
5-6 hour positive	73
6-7 hour positive	23
7-12 hour positive	39
12-24 hour positive	20
1-2 days positive	10
2-3 days positive	8
3-10 days positive	10
10-90 days positive	13

Table 2. *Differences in epoch for the first quarter of 1993.*

It can therefore be concluded that the number of 3 negative matches in the first row of Table 2 and the 23 discrepancies in the last two rows can indeed be caused by incorrect matching by use of the other three variables. For the 27 observations in row 2, it is more likely that a coding error has been made.

On the basis of these results, the value of δ is chosen one for all values not between -1 day and plus three days. Those pairs between -1/2 hour and three hours will get a value of zero. Small different positive values are given to pairs with differences within one day and between one and three days. Given the high value of c, all pairs with a value of $\delta=1$ will not be matched. The number of records that should have been matched, but that will not be matched for this reason is estimated to be some twenty records.
Similar procedures have been applied to the other variables. On the basis of these analyses, c-values and δ-values are determined.

3. The matching procedure.

The matching procedure could also be called a mating procedure. A possible mate should have an acceptable distance. The 'best' mate for a record in one data set is the record in the other data set that has the smallest distance. But that possible mate might itself have a more favourable mate in the other data set. Furthermore, distances (even zero distances) need not be unique.
To solve these problems and to avoid unrealistic (time consuming) procedures based on calculation of all possibilities, a strategy has been selected that assures the possibility of an acceptable match and meets certain computation time criteria that makes application to data sets of the size at hand realistic.
In this procedure, the records in each data set are ordered according to epoch and given a rank number.
Furthermore, during the matching procedure each record gets a matching status indicator (undecided, matched, not matchable), two pointers to records in the other data set, two distances to the records pointed to and a selectivity value. At the beginning the status indicator is initiated at the value 'undecided' and the distances at a maximal value.

In principle all records in each set should be compared with all records in the other set. However, ordering according to epoch makes it possible to reduce the total number of comparisons with a factor 100 if records outside a certain range are ignored. At each comparison, the distance is computed and recorded in one of the two distance fields, if this distance is smaller than a previous value. The corresponding pointer field is filled with the rank number of the record from the other set. This is done in both data sets if applicable.
At the end of this procedure, for each record in each set there is a pointer to the most favourable mate, together with the associated distance and a second pointer and distance to the next favourable mate.

The next step is to check for each record in each set whether the smallest distance is acceptable. If not, then that record gets the status 'not matchable'.
The actual matching procedure is as follows:

a) start with first record in set 1 that is undecided;
b) look at the first mate (record with smallest distance in the other set);
c) check if the indicator of that mate is 'undecided'; if 'yes', then d1), else d2)
d1) if the first mate of the record in the other set is the starting record, then the records will be matched, else make this mate the starting record and goto b);
d2) if the second mate is not 'undecided', then the (start)record is 'not matchable'; if 'undecided', then goto d1).

It is proven that the change in the choice of the starting record in d1) is not circular and therefore ends in a finite number of steps.
For each record of a matched pair, the difference between the distance to the first and second mate is computed. The selectivity measure for a matched pair is subsequently computed as the minimum of these differences and added to both records, together with

their distance and the pointer to the other record.

4. Results.

Results of the automated procedure are compared to the outcomes of the matching by hand. Coefficients were adjusted and some programme bugs eliminated. The final results were satisfactory. Only a small proportion of the records matched by hand (in which also additional heuristic rules are used) were missing. The method was then applied to the data for 1992 and 1993. The results for both years were rather similar and in agreement with the outcomes of the quarter that had been used for correction of the procedure.
A number of checks was carried out to evaluate the outcomes in more detail.

4.1 Checks on the results.

In an ideal case, matched pairs should have a minimum distance and a maximum value on the selectivity measure, while the reverse should be the case with the records that were not matched.

	S=0-39	40-79	80-119	120-159	160+	Total
D = 0	23	86	2302	2708	940	6059
1-40	11	246	880	619	108	1864
41-65	76	690	1049	286	17	2118
66-100	193	336	118	8	1	656
101-130	399	180	30	2	0	611
131-200	2819	206	7	1	0	3033
200+	78	18	0	0	0	96
Total	3599	1762	4386	3624	1066	14437

Table 3. *Distance- and selectivity categories for matched records from 1993.*

Table 3. gives an overview of the relation between distance (D) and selectivity (S) for one year. It contains all matched records, ranging from very likely correctly matched till very unlikely correctly matched (distance 200+). Most of these records (6059) have distance zero, and 98% of these have a selectivity of 80 or more. For 86 % in the distance class 1-40 the selectivity is still higher than 80. For the classes with a higher distance we see on average a decrease in selectivity. Almost all records above distance 100 have low selectivity. On the basis of this analysis records got a 'matching quality measure' (rank A through F), depending on the combination of distance and selectivity.
The highest rank A (almost certain) is given when D=0 and S>79; rank B (very likely)

when D=0 and S between 40 and 79 or D between 1 and 40 and S>39; rank C (likely) when D between 41 and 65 and S>39; rank D (reasonable) when D between 66 and 100 and S>39; rank E (doubtful) when D between 101 and 130 and S>39; all other records got rank F (almost certainly not correct).
In all further work, this distinction in quality has been taken into account, in order to find the most likely estimates (e.g. of under-reporting or bias in databanks).
Other checks regarded the coding of accidents and the severity code in the SIG-databank.
Finally a comparison was made with the results of a similar pilot-matching from 1987.

4.2 The ' footprint' method.

Furthermore, estimation of the number of records for the cells of Table 1. took place. This was done, not only using the records that were not matched or matched with high probability, but also the records that were matched with low probably, correcting for the over-estimation of these numbers.
Classes of pairs, matched with higher and lower certainty, were treated separately in the analysis. For the class with a perfect match and therefore very high probability of correctness, the mismatch on other variables than the key-variables was used to measure the probability of inconsistencies between scores. The cross-table of scores on such a variable in both databanks for only the correctly matched pairs was used to get a profile of similarities and dissimilarities in both scores that were not caused by random matching. This profile was called a 'footprint'. The footprint was further used to estimate the number of correct matches in the groups with lower certainty. The method was applied to the data from 1992 and 1993 combined, using information about the mode of transportation of the victim, although the categories of this extra variable are partly different in both databanks.

The main procedure is as follows.
For the pairs with maximal likelihood of matching (rank A), a comparison is made between the codes for traffic mode in the two records. It is assumed that, if these codes differ, this is caused by random error, by deletion (unknown) or systematic differences in coding instruction. Because it regards (almost) surely correct pairs, we call this matrix the 'footprint' of the table.
For each of the sets with less likely matchings the same can be done. However, here random matches are expected to disturb the footprint. On the assumption of 'complete random matching' (and therefore independent row and column values), we can compute the expected cell-values from the given marginals of such a table. It is assumed that the observed table (T) is a combination of the footprint (F) and the random table (R): $T = a \cdot F + (n-a) \cdot R$, where n is the total number of pairs, and a is the number of correct pairs. The most likely combination (in terms of the minimum Chi-square) of the footprint table and the random table results in the respective estimates of the correct matches (a) and incorrect matches (n-a).

Hospitalized victims	A = 0	10 - 40	44 - 65	66 - 100	101 - 130	131 - 200
Motorized	11302	3237	3069	785	649	2342
Perc. correct	100%	100%	100%	79%	40%	14%
Non-motorized	839	247	310	180	394	2052
Perc. correct	100%	100%	100%	51%	15%	1,6%

Table 4. *Estimated number of correctly matched records in two main SIG-categories of victims, as a function of distance, for 1992 and 1993 together.*

Of course, using this method it is not possible to detect which records are correct and which are not, but only to estimate the number of correct matchings for each quality class. The estimated number of random matches in Class B and C turned out to be negligible. 73% of Class D was correctly matched and 30% of class E.
Table 4 gives an indication of the outcomes of this procedure for aggregated traffic modes. It can be seen that the estimated number of correctly matched pairs for the records with small distances, up to 65 is 100% for motorized as well as non-motorized victims. For the other categories these percentages are lower and decreasing with distance. Especially or the non-motorized victims this is the case. Only 16% has a distance larger than 100. For the motorized victims still 40% has a distance between 100 and 130.

5. Conclusions.

Procedures has been developed that can effectively be used to estimate the total number of hospitalized victims by comparing the information from the official road safety statistics and additional information from hospital records. These procedures can be generalized to all kinds of situations where matching of databanks is necessary and matching by hand is too time consuming.
Although there was no unique key that matches the records and makes estimation of under-reporting possible from the information of the (almost complete) database from the hospitals directly possible, it was still possible to estimate this under-reporting, using a technique called the ‘footprint’ method. Also this procedure has a wider range of possible applications.
The outcomes of the study have been used in the Netherlands to revise the reporting of hospitalized victims. The result was that considerably higher numbers of victims are now officially reported. As a consequence of this policy, also the trends in these numbers will be less favourable over the last years, compared with the originally published figures.
On the basis of the outcomes of the study it was also possible to estimate the bias in the database of the official road statistics with regard to characteristics such as traffic mode and region. In principal this is also possible for other characteristics. The outcomes for safety itself will be reported in a separate publication.