# Crash Involvement Studies Using Routine Accident and Exposure Data: A Case for Case-Control Designs

H. Hautzinger*

*Institute of Applied Transport and Tourism Research (IVT), Kreuzaeckerstr. 15, D-74081 Heilbronn, Germany

**Abstract –** Fortunately, accident involvement is a rare event: the chance of an individual road user trip to end up in a crash is close to zero. Thus, according to general epidemiological principles one can expect the case-control study design to be especially suitable for quantifying the relative risk (odds ratio) of accident involvement of road users with a certain risk factor as compared to road users that do not have this characteristic. Ideally, of course, the database for such a case-control study should be established by drawing two independent random samples of cases (accidental units) and controls (non-accidental units), respectively. If, however, special data collection is not an option, it is nevertheless possible to analyze routine accident and exposure data under a case-control design in order to fully exploit the information contained in already existing databases. As a prerequisite, accident and exposure data from different sources are to be combined in a single file of micro or grouped data in a way consistent with the case-control study design. Among other things, the proposed methodological approach offers the possibility to use in-depth data of the GIDAS type also in investigations of active vehicle safety by combining this data with appropriate vehicle trip data collected in mobility surveys.

## NOTATION

$\Psi$ population odds ratio
$\Lambda$ population relative risk
$\psi$ sample odds ratio
$\lambda$ sample relative risk

## INTRODUCTION

### Basic idea

As is well known, the case-control study design is useful for risk factor assessment in situations where the disease in question is rare. Accident-involvement is such a rare event: the chance of a road user trip to end up in a crash is close to zero. Thus, one can expect the case-control design to be efficient for quantifying the relative risk (odds ratio) of traffic accident involvement of road users with a certain risk factor as compared to road users that do not have this characteristic. A case-control design is characterised by a dataset of accident-involved road users ("cases") and a second independent dataset of road users not involved in an accident ("controls") belonging to the same general population. Ideally, of course, the database for such a case-control study should be established by drawing two independent random samples of cases and controls, respectively.

Quite often, however, such special data collection is not an option and the researcher is restricted to the use of already existing data (secondary or routine data). In this situation it may nevertheless be possible to analyze routine accident and exposure data available from external sources under a case-control design in order to fully exploit the information contained in these databases. The most crucial prerequisite for such an approach is that accident and exposure data from different sources can be combined in a single file of micro or grouped data in a way consistent with the case-control study design. The methodology presented in this paper has been developed under the TRACE project [1].

### Example

Accident data and vehicle registration data can be combined under a case-control study design in order to assess risk factors for accident involvement. Cases may, for instance, be accident-involved vehicles recorded in an in-depth study like GIDAS and controls could be vehicles randomly selected from the national vehicle register. If cases are vehicles involved in an accident during a specific study year,

controls should be vehicles registered in the country under consideration during the same year (the sample of controls may, for instance, be drawn from the mid-year vehicle stock).

In the above context, "vehicle-years" would normally be considered as units at risk and, consequently, the case-control study could be conducted at the vehicle-year level. In this situation, the population at risk consists of all vehicle-years coinciding with the study period (e.g. calendar year 2007). Obviously, this population can be considered as being decomposed into the subpopulations of "accidental" and "non-accidental" vehicle-years, respectively. Thus, accident-involved vehicles recorded in the in-depth study ("cases") may be interpreted as a sample from the subpopulation of all accidental units at risk. Similarly, vehicles drawn from the national vehicle register ("controls") may be considered as sampled from the subpopulation of non-accidental units at risk.

Clearly, any risk factor to be assessed must be recorded both for cases and controls. Thus, in studies using routine traffic accident and vehicle registration data, the assessment of risk factors for accident involvement is restricted to vehicle and vehicle-holder characteristics which are contained in both data sources. This, of course, limits the scope of purely "secondary" studies. Sometimes, however, it might be possible to "enrich" the data files of cases and controls. If, for instance, an appropriate vehicle identification number is contained in both files, one can augment the list of variables with various technical characteristics of the vehicle.

If vehicles with and without the risk factor of interest differ substantially with respect to possible confounding variables like vehicle mileage, simple group comparisons under the case-control design might be biased. As mileage information is frequently not available, one could, however, adjust relative accident involvement risk for variables known to be strongly associated with vehicle mileage (e.g. engine power and vehicle age).

## ASSESSMENT OF RISK FACTORS FOR ACCIDENT INVOLVEMENT

### Preparation of the case-control database

Under the approach outlined above one may, for instance, assess the effect of a certain in-vehicle safety system like ESP on the risk of accident involvement. In order to obtain the desired case-control database, vehicles recorded routinely in an in-depth accident study or in national road traffic accident statistics are considered as a random sample from the subpopulation of all accident-involved vehicles. These accident-involved cars (more precisely, accidental vehicle-years) are considered as „cases". Similarly, vehicles contained in the national vehicle register are considered as a random sample of cars that have not been involved in an accident during the specified time period (possibly screening to eliminate accident-involved cars). These cars are considered as „controls". Both for cases and controls it is to be ascertained whether or not the corresponding car is equipped with the device to be assessed.

The routine accident and exposure data thus obtained may be displayed in a 2 × 2 contingency table showing the joint frequency distribution of accident involvement status (rows) and risk factor status (columns):

| | equipped | not equipped |
|---|---|---|
| accident-involved | *a* | *b* |
| not involved | *c* | *d* |

The above table contains sample data. The corresponding population values of the cell frequencies may be denoted by capital letters *A*, *B*, *C* and *D*.

## Measuring comparative chance of accident involvement

Since the sampling fractions $f$ and $g$ for cases („accident-involved“) and controls („not involved“), respectively, will normally be different, the expected values in the sample are given by the following products

[1] $fA, fB, gC$ and $gD$ .

In case-control studies where the sampling fractions $f$ and $g$ are not equal (in our context $f$ will normally be considerably larger than $g$) only the odds ratio can be estimated, but not risk, relative risk or odds. The expected value of the sample odds ratio $\psi$ = (a/c)/(b/d) equals the population odds ratio:

[2] $(fA/gC) / (fB/gD) = (A/C) / (B/D) = \Psi$.

Thus, the odds ratio is the appropriate measure of comparative chance of traffic accident involvement of equipped (“exposed”) vehicles as compared to those not equipped (“not exposed”).

As accident-involvement is a very rare event, the odds $A/C$ are approximately equal to the empirical risk $R_1 = A/(A+C)$ and the odds $B/D$ will differ only slightly from the empirical risk $R_0 = B/(B+D)$. Thus, the odds ratio $\Psi$ is a good approximation to the relative accident-involvement risk

[3] $\Lambda = R_1 / R_0$

of cars equipped with the device as compared to cars without the safety system of interest.

Consequently, both the population odds ratio $\Psi$ and the relative risk $\Lambda$ may be estimated by the sample odds ratio

[4] $\psi = (a/c) / (b/d) = (ad)/(bc)$ .

Clearly, the above measure $\psi$ of comparative chance of accident-involvement can also be calculated for subgroups of vehicles. If in addition to point estimates of the population odds ratio also confidence intervals are to be calculated standard statistical theory can be applied [2].

## Controlling for confounding variables

Accident-involvement is, of course, not only affected by the dichotomous risk factor „equipment with safety device of interest“ (actually, equipment will be a protective factor rather than a risk factor). Cell frequencies in the above 2 × 2 table of accident involvement counts will, for instance, also depend on car mileage. If average annual mileage differs between cars with and without the safety device under consideration the above comparison is biased.

In order to account for structural differences between cases and controls one can use multiple logistic regression models to analyse the case-control sample data. In these models the accident involvement or case-control status of a sample unit (involved / not involved in accident during study period) is the binary outcome variable whereas risk factor status (equipped yes/no) and vehicle mileage (kilometres driven during study period) are explanatory variables. Such an approach requires mileage data to be ascertained for the sample vehicles. In principle, this could be accomplished by interviewing the holders and/or drivers of the cars in the study. If such a retrospective vehicle mileage survey cannot be conducted, one could alternatively use vehicle characteristics known to be correlated with mileage and car use (e.g. vehicle age, engine power, car make and model etc.) as additional explanatory variables in the logistic regression model.

## PRACTICAL APPLICATION OF THE CASE-CONTROL APPROACH

### Description of the routine accident and exposure data sets used

In order to illustrate the approach using real-world data, a case-control study has been carried out based on routine data from German road traffic accident statistics 2002 (for cases) and from the German mobility survey MiD 2002[1] (for controls), respectively. In this study the effect of the individual's age and gender on accident involvement risk of car drivers was investigated. According to the nature and content of the two independent routine databases, the case-control study was conducted at the trip level [1].

*Cases* are accident-involved car drivers selected from the records of German traffic accident statistics (year 2002, all accident-involved car drivers). The number of cases is 455886. It is easy to see that every accident-involved road user corresponds to an *accidental trip*. Thus, the cases are a 100 percent sample from the actual and finite population of accidental car driver trips in Germany 2002. Clearly, this population is a subpopulation of all car driver trips of the year 2002 which is to be considered as the population at risk.

*Controls* are car driver trips sampled under the above mentioned mobility survey MiD 2002, where representative trip data covering the year 2002 have been collected using the trip diary technique. Just as with all mobility surveys, the MiD survey has been conducted under a cluster sampling design (households as clusters of persons and trips). The number of car driver trips in the MiD survey amounts to 69443. For the purpose of this example we can assume that all these trips are non-accidental, i.e. controls. As the annual total number of car driver trips for Germany 2002 is estimated at $41561 \times 10^6$, the sampling fraction for controls is very small ($1.67 \times 10^{-6}$); on average, information is available only for less than 2 trips out of 1 million car driver trips.

As usual, the method of data analysis depends on the scaling of the risk factor.

### Assessing a dichotomous risk factor

In order to assess the effect of the dichotomous risk factor driver gender on accident involvement risk, the sample data are presented in the following 2 × 2 table:

| **Risk factor status** Driver gender | **Accident involvement status** **cases** accidental trips | **controls** non-accidental trips |
|---|---|---|
| - male | 293002 | 38688 |
| - female | 162885 | 30755 |
| **Total** | 455886 | 69443 |

From the sample data shown in this table one may estimate the population odds ratio $\Psi$ for accident involvement (male as compared to female drivers) as follows:

[5] $$\psi = (293002 \times 30755) / (38688 \times 162885) = 1.430.$$

[1] MiD is an acronym for „Mobilität in Deutschland" (=mobility in Germany).

The approximate standard error of the log of the sample odds ratio[2] is calculated to be

[6] $\sqrt{[1/293002 + 1/162885 + 1/30755 + 1/38688]} = 0.00824.$

Thus, approximate 95 percent confidence limits for the population odds ratio $\Psi$ are

[7] $\exp\{\log_e 1.43 \pm 1.96 \times 0.00824\}$

that is, (1.407, 1.453).

Consequently, being a male car driver increases the chance of accident involvement by a factor of around 1.43 (male car drivers have 143% of the involvement risk of female car drivers). We are 95 percent sure that the interval from 1.407 to 1.453 contains the true odds ratio $\Psi$ (which is a good approximation to the population relative risk $\Lambda$).

Under a case-control design the chi-square test (or where necessary Fisher's exact test) may be used without modification to test the null hypothesis of no association between risk factor status (gender) and case-control status (accident involvement yes/no).

As with any kind of study, the results obtained for a single risk factor may be compromised by confounding or interaction with other variables. In addition to the Mantel-Haenszel method logistic regression models and other more complex generalised linear models may be used to adjust for confounding or to deal with interaction.

An example is presented in a subsequent sub-section.

## Assessing a polytomous risk factor

When the risk factor is a polytomous attribute, one level or category of the risk factor is chosen as a base level and all other levels are compared to this base. This comparison to the base is made level by level ignoring at a time all other levels. Consequently, level-specific odds ratios and confidence intervals can be calculated as previously described. We consider "driver age class" as an example:

| **Risk factor status** Driver age | **Accident involvement status** **cases** accidental trips | **controls** non-accidental trips | **Odds Ratio** |
|---|---|---|---|
| - 18-24 | 111661 | 7245 | 2.292 |
| - 25-44 | 201639 | 28661 | 1.000 |
| - 45-59 | 86376 | 21575 | 0.569 |
| - 60-64 | 21661 | 5465 | 0.563 |
| - 65+ | 34549 | 6488 | 0.757 |
| **Total** | 455886 | 69443 | |

[2] The standard error as calculated here is based on the assumption of two independent simple random samples of cases and controls. Actually, however, controls have been selected under a cluster sampling design. For simplicity, the corresponding design effect (variance of the estimate obtained from the more complex sample to the variance of the estimate obtained from a simple random sample of the same number of units) is ignored here.

Drivers aged 25 to 44 years were chosen as the base group because they are the largest group in number, and thus most accurately measured. Obviously, the risk of car drivers aged 18 to 24 years to be involved in a traffic accident is more than twice as high as the involvement risk of drivers aged 25 to 44 years ($\psi_{18\text{-}24|25\text{-}44} = 2.292$). The standard error of the log of the odds ratio is estimated at

[8] $$\sqrt{[1/111661 + 1/7245 + 1/201639 + 1/28661]} = 0.01367.$$

Consequently, approximate 95 percent confidence limits for the population odds ratio $\Psi_{18\text{-}24|25\text{-}44}$ are

[9] $$\exp\{\log_e 2.292 \pm 1.96 \times 0.01367\}$$

that is, (2.231, 2.354). As stated above, this confidence interval might be somewhat too narrow because the design effect has been neglected. For the remaining three age groups the odds ratio can be estimated analogously. According to the above table, there is some relationship between odds ratio and age class. If this relationship is to be analysed, one can use logistic regression models for categorical or ordinal risk factors (dependent variable is case-control status of car driver trip).

## Assessing several risk factors simultaneously

A multiple logistic model can be applied to assess the joint effects of driver age group and driver gender on car driver accident involvement risk. The variables of the model are specified as follows:

| | |
|---|---|
| Y: case-control status | (response variable coded 1 for cases and 0 for controls) |
| A: age group | (explanatory variable, 5 classes) |
| G: gender | (explanatory variable, 2 classes) |

The data are supplied to the computer package (SAS) in grouped form. As there are $2 \times 5 \times 2 = 20$ combinations of the outcomes of the three variables, the data matrix consist of 20 rows. The first 3 columns of the data matrix correspond to the 3 variables Y, A and G. Column 4 contains the frequency counts for all combinations; these counts are used as weights in the regression analysis.

| case-control status (Y) | age group (A) | gender (G) | count |
|---|---|---|---|
| 1 | 18-24 years | male | 71506 |
| 1 | 18-24 years | female | 40155 |
| 1 | 25-44 years | male | 122787 |
| 1 | 25-44 years | female | 78852 |
| 1 | 45-59 years | male | 56435 |
| 1 | 45-59 years | female | 29941 |
| 1 | 60-64 years | male | 15864 |
| 1 | 60-64 years | female | 5797 |
| 1 | 65+ | male | 26410 |
| 1 | 65+ | female | 8139 |
| 0 | 18-24 years | male | 3992 |
| 0 | 18-24 years | female | 3253 |
| 0 | 25-44 years | male | 13436 |
| 0 | 25-44 years | female | 15225 |
| 0 | 45-59 years | male | 12288 |
| 0 | 45-59 years | female | 9287 |
| 0 | 60-64 years | male | 3852 |
| 0 | 60-64 years | female | 1613 |
| 0 | 65+ | male | 5114 |
| 0 | 65+ | female | 1374 |

The total number of units in the database is 525320 (cases: 455886; controls: 69434).

The logistic model can be formulated as follows:

[10] $P_{ij} = \exp(u_{ij})/[1+\exp(u_{ij})] = 1/[1+\exp(-u_{ij})]$.

where $P_{ij}$ denotes the probability for a unit (car driver trip) to be a "case" given age class $i$ and gender category $j$ and $u_{ij}$ is defined as

[11] $u_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$.

In the logistic model the effects are centred, i.e. the coefficients $\alpha_i$ and $\beta_j$ sum up to zero, respectively. Analogously, the interaction effects $\gamma_{ij}$ sum up to zero for each row $i$ and column $j$ in the $5 \times 2$ table corresponding to the combinations of A and G. The logistic model can easily be extended to consider more than two risk factors.

The main elements of the output of the SAS procedure CATMOD[3] are shown in the following display:

```
The SAS System

The CATMOD Procedure

                      Data Summary

Response           ccs             Response Levels       2
Weight Variable    COUNT           Populations          10
Data Set           CASECONTROL     Total Frequency  525320
Frequency Missing  0               Observations         20


             Population Profiles

Sample    AGECLASS       GENDER    Sample Size
-----------------------------------------------
    1     18-24 years    female          43408
    2     18-24 years    male            75498
    3     25-44 years    female          94077
    4     25-44 years    male           136223
    5     45-59 years    female          39228
    6     45-59 years    male            68723
    7     60-64 years    female           7410
    8     60-64 years    male            19716
    9     65+            female           9513
   10     65+            male            31524

               Response Profiles

Response    ccs  case-control status
-------------------------------------------
    1       0    case
    2       1    control (reference category)


                  Maximum Likelihood Analysis

           Maximum likelihood computations converged.


     Maximum Likelihood Analysis of Variance

Source               DF   Chi-Square    Pr > ChiSq
-------------------------------------------------
Intercept             1     102090.0        <.0001
AGECLASS              4     10004.60        <.0001
GENDER                1       523.00        <.0001
AGECLASS*GENDER       4       513.48        <.0001

Likelihood Ratio      0            .             .
Analysis of Maximum Likelihood Estimates

                                              Standard        Chi-
Parameter                       Estimate         Error      Square    Pr > ChiSq
-------------------------------------------------------------------------------
Intercept                         1.8066       0.00565    102090.0        <.0001
```

[3] In order to obtain the SAS output in the form presented here the coding of cases and controls has to be reversed (i.e. 1 for controls and 0 for cases).

| | | | | | |
|---|---|---|---|---|---|
| AGECLASS | 18-24 years | 0.8927 | 0.0110 | 6559.47 | <.0001 |
| | 25-44 years | 0.1219 | 0.00749 | 265.02 | <.0001 |
| | 45-59 years | -0.4591 | 0.00825 | 3099.58 | <.0001 |
| | 60-64 years | -0.4593 | 0.0141 | 1058.72 | <.0001 |
| GENDER | female | -0.1293 | 0.00565 | 523.00 | <.0001 |
| AGECLASS*GENDER | 18-24 years female | -0.0569 | 0.0110 | 26.60 | <.0001 |
| | 25-44 years female | -0.1546 | 0.00749 | 426.13 | <.0001 |
| | 45-59 years female | -0.0476 | 0.00825 | 33.35 | <.0001 |
| | 60-64 years female | 0.0612 | 0.0141 | 18.80 | <.0001 |

Given the case-control database, the probability of a car driver trip to be a case, i.e. to be an accidental trip, given that the driver is aged 18-24 years ($i$=1) and male ($j$=1) is estimated at

[12] $P_{11} = 1/[1 + \exp(-1.8066 - 0.1293 - 0.8927 - 0.0569)] = 1/ [1 + \exp(-2.8855)] = 1/1.0558 = 0.9471.$

This quantity, of course, can *not* be used to describe the absolute risk of young male car drivers! The reason is that the database has not been created by drawing a random sample from the complete population at risk, i.e. from the population of all car driver trips made occurring in Germany 2002. Rather, two independent samples with extremely different sampling fractions have been drawn from the subpopulations of accidental and non-accidental units, respectively. As can be seen, the probability $P_{11}$ = 0.9471exactly corresponds to the empirical proportion of cases in the subgroup of male drivers aged 18-24 years. According to the above data matrix this proportion equals 71506/(71506+3992) = 0.9471 = 94.71%.

The above model estimation results can be interpreted as follows:

- According to the case-control design of the study one can only make statements on the *relative* risk of accident involvement (comparisons between the different combinations of age group and gender).
- The model constant 1.8066 simply reflects the fact that in the database used the number of cases is by far larger than the number of controls. The quantity exp(1.8066)/(1+exp(1.8066)) = 0.859 approximately equals the empirical proportion of cases in the database (which is 86.8%).
- Age class of driver is a highly significant explanatory variable for traffic accident involvement (Chi-square 10004.60; 4 degrees of freedom).
- The effect of driver age class on accident involvement risk is nonlinear (U-shaped) with highest risk for young drivers (18 to 24 years) and lowest risk for drivers aged 45 to 64 years. The estimate for the parameter $\alpha_5$ (age class 65+) is not shown in the SAS display and must be calculated by hand. As the parameters for the five age classes must sum up to zero, one obtains the estimate -0.0962 indicating that accident involvement risk increases again once driver's age exceeds 64 years. The parameters associated with the different age classes are to be interpreted as "partial" regression coefficients.
- Driver gender also determines accident involvement risk significantly. As compared to driver age class, the effect of gender, however, is less important (Chi-square 523.00; 1 degree of freedom). The coefficients associated with the two categories (male and female, respectively) are showing the partial effect of gender. As before, the estimate for parameter $\beta_1$ (male) has to be calculated by hand; here, one simply has to reverse the sign of the parameter for the female category. The positive sign of the parameter estimate for the male category (0.1293) indicates that male drivers are at higher risk as compared to female drivers.
- In addition to the two main effects (age class and gender, respectively), the two-way interaction effect is also significant (Chi-square 513.48; (5-1)(2-1)=4 degrees of freedom). Significance of the two-way interaction means that the effect of driver gender on accident involvement risk is not the same for all age groups. Generally, there is higher risk for male

drivers as compared to female drivers; for specific age groups, however, this effect may even be reversed.

In order to quantify the relative risk of traffic accident involvement for certain subgroups of car driver trips (defined by age class and gender of driver), the odds of accident involvement given an arbitrary risk factor status combination ($i$, $j$) has to be related to the corresponding odds for a certain base or reference combination ($r$, $s$). Under the above logistic model with main and interaction effects, the odds ratio (relative chance of accident involvement given risk factor status combination ($i$, $j$) as compared to risk factor status combination ($r$, $s$) may be written as

[13] $\Psi_{ij|rs} = [P_{ij}/(1- P_{ij})] / [P_{rs}/(1- P_{rs})] = \exp(u_{ij}) / \exp(u_{rs}) = \exp[(\alpha_i - \alpha_r) + (\beta_j - \beta_s) + (\gamma_{ij} - \gamma_{rs})]$.

As before, for instance, age class "25-44 years" and gender category "female" may be considered as the reference categories ($r$ and $s$, respectively) of the two risk factor status variables.

Due to the significance of the two-way interaction, the odds ratio for male drivers ($j$) as compared to female drivers ($s$) is not constant. Rather, this measure of relative risk of traffic accident involvement varies over driver age classes $i$:

[14] $\Psi_{ij|is} = [P_{ij}/(1- P_{ij})] / [P_{is}/(1- P_{is})] = \exp(u_{ij}) / \exp(u_{is}) = \exp[(\beta_j - \beta_s) + (\gamma_{ij} - \gamma_{is})]$

The following estimated odds ratios are obtained:

| Driver age class ($i$) | Estimated odds ratio $\psi_{ij\|is}$ (male vs. female drivers) |
|---|---|
| 18-24 | exp[(0.1293 - (-0.1293)) + (0.0569 - (-0.0569))] = exp(0.3724) = 1.45 |
| 25-44 | exp[(0.1293 - (-0.1293)) + (0.1546 - (-0.1546))] = exp(0.5678) = 1.76 |
| 45-59 | exp[(0.1293 - (-0.1293)) + (0.0476 - (-0.0476))] = exp(0.3538) = 1.42 |
| 60-64 | exp[(0.1293 - (-0.1293)) + (-0.0612 - 0.0612)] = exp(0.1362) = 1.15 |
| 65+ | exp[(0.1293 - (-0.1293)) + (-0.1979 - 0.1979)] = exp(-0.1372) = 0.87 |

As long as driver age does not exceed 64 years, the chance of a car trip to end up in an accident is 15 up to 76% higher if the driver is male. Among car trips made by elderly drivers (65 years and over), however, trips made by female drivers are more prone to accident involvement than trips made by male drivers.

Similarly, it appears that the effect of driver age class on the risk of traffic accident involvement may be different for trips made by male and female drivers, respectively:

| Driver gender ($j$) | Estimated odds ratio $\psi_{ij\|rj}$ (driver age class 18-24 vs. age class 25-44) |
|---|---|
| male | exp[(0.8927 - 0.1219) + (0.0569 - 0.1546)] = exp(0.6731) = 1.96 |
| female | exp[(0.8927 - 0.1219) + (-0.0569 - (-0.1546))] = exp(0.8685) = 2.38 |

As can be seen, being a novice driver is a risk factor for accident involvement (involvement risk is roughly doubled as compared to drivers aged 25-44 years); this is especially true for female beginners.

Clustering of cases[4] and controls[5] has not been accounted for in this analysis. Random effects models could be used for this purpose.

## CONCLUDING REMARKS

### Usage of routine data versus special data collection

Empirical studies on traffic accident involvement risk may be carried out under different research designs: Surveys, cohort studies and case-control studies appear to be the most relevant. Ideally, under

[4] Two or more car drivers can be involved in the same accident. Therefore, accidents are clusters of road users involved.

[5] The set of trips made by a specific person on a given day is also to be considered as a cluster.

a given study design special data on traffic participation and accident involvement should be collected in order to answer the research questions. According to basic epidemiological principles, "special data collection" means sampling from the population at risk.

As a low cost alternative to special traffic participation and accident involvement data collection, the use of "routine" accident and exposure data for scientific purposes is of importance. As can be expected, traffic accident statistics on the one hand and household mobility surveys or vehicle mileage surveys on the other hand play a dominant role in this context. Studies based on routine data are generally not especially useful for demonstrating causality, but are useful for descriptive purposes ([2], p. 18-22). In studies on accident involvement risk the potential of routine data is further limited due to the reasons described below.

## Limitations of routine data in risk studies at the trip level

Whereas the annual number of accidental trips $Y_A$ is quite well documented in official traffic accident statistics, the total annual number $Y$ of all road user trips - and thus the size of the population at risk - is never known from a complete census. Rather, this number (usually called "total trip volume") can only be estimated from sufficiently large sample surveys on individual travel behaviour. As large-scale mobility surveys are costly, they are conducted in most countries only every 5 or 10 years.

## Limitations of routine data in risk studies at the person-year level

The number $N_A$ of accident-involved road users is not known from statistical sources. As, however, multiple accident involvement of individuals is rare, the annual number of accidental trips $Y_A$ (which is recorded routinely by police) will be only slightly larger than the number $N_A$ of road users involved in an accident in the course of the calendar year under consideration. Thus, $N_A$ may be approximated sufficiently precise by $Y_A$.

In contrast to this, the total number $N$ of trip makers under risk is extremely difficult to estimate for longer study periods (e.g. one year) as in most mobility surveys the respondents are reporting their trips only for a single day of the year. Thus, for instance, the number $N_{bicycle}$ of persons participating in traffic as cyclists (at least one bicycle trip per year) is simply unknown and could only be estimated from a specifically designed mobility survey where the reporting period of the sample units corresponds to one calendar year. In such a survey the interviewee had to be asked whether or not he or she has used the bicycle as a travel mode during the last twelve months.

## Individual versus grouped routine data

Clearly, generic data on individual units at risk offer the best basis for risk analysis. Routine data on accident involvement, however, are quite often only available in grouped form, i.e. as tables where accident involvement counts are broken down by one or more characteristic of the accident or the accident-involved road users. Fortunately, if appropriate exposure quantities are available at the same level of aggregation, grouping does not unduly restrict the possibilities of statistical risk analysis.

## Sources of routine data on accident involvement

The most important sources of data on traffic accident involvement and accident causation are

- official road traffic accident statistics (police-recorded data),
- in-depth traffic accident studies, and
- vehicle insurance data files.

However, also hospital data may be used [3]. As compared to other fields of epidemiological research, routine data from national traffic accident statistics already offer a wide variety of possibilities for

analysis. This is especially true if the accident records contain sufficiently detailed information on the accident-involved vehicles.

## Sources of routine data on exposure to accident involvement risk

Exposure data contain information on the number and characteristics of the units at risk (irrespective of traffic accident involvement). Depending on the analysis level, the corresponding data can be obtained from different routine sources.

Typical data sources for accident involvement risk studies at the trip level are mobility surveys (trip diaries). Sources for risk studies at the person- or vehicle-year level are (*i*) population census data, (*ii*) vehicle registration data and (*iii*) vehicle mileage surveys.

## Problems of combining accident and exposure data from different sources

In situations where special data collection is not an option, the analyst has to combine routine accident and exposure data from different sources. While doing so, one regularly is faced with the problem of harmonizing the data (e.g. definition of variables and variable values) which can be an extremely cumbersome task.

Summarising, it can be said that accident involvement risk studies should be based on accident and exposure data. The so-called quasi-induced exposure method where only accident data are analysed is normally a less-than-ideal solution. As for reasons of economy the collection of special data on accidental and non-accidental units is frequently not possible, researchers are restricted to the use of routine accident and exposure data in many situations. If the combined data set is prepared in a way consistent with the case-control design, the potential of epidemiological methods for this type of study can be exploited.

## REFERENCES


1 H Hautzinger, C Pastor, M Pfeiffer and J Schmidt, *Analysis Methods for Accident and Injury Risk Studies*, Deliverable 7.3 EU Project No. 027763 – TRACE (Traffic Accident Causation in Europe), Heilbronn/Germany, IVT, 2007.

2 M Woodward, *Epidemiology – Study Design and Data Analysis*, Second Edition, Boca Raton/London, Chapman & Hall/CRC, 2004

3 D Böhning and S N A Rampai, A case-control study of non-fatal accidents on hospital patients in Bangkok metropolis, *Sozial- und Präventivmedizin* 42: 1997, p. 351-356