D.1.15. Final Report on *Task 1.5*

Contract No: **TREN-04-FP6TR-SI2.395465/506723 “SafetyNet”**
Acronym: **SafetyNet**
Title: **Building the European Road Safety Observatory**

Integrated Project, Thematic Priority 6.2 “Sustainable Surface Transport”

Project Co-ordinator:
Professor Pete Thomas
Vehicle Safety Research Centre
Ergonomics and Safety Research Institute
Loughborough University
Holywell Building
Holywell Way
Loughborough
LE11 3UZ

Organisation name of lead contractor for this deliverable: TRL

Due Date of Deliverable: 31/12/2007

Submission Date: 20/03/2008

**Report Authors: Jeremy Broughton, Emmanuelle Amoros, Niels Bos,
Petros Evgenikos, Stefan Hoeglinger, Péter Holló,
Catherine Pérez, Jan Tecl**

Project Start Date: 1st May 2004 **Duration: 4 years**

Project co-funded by the European Commission within the Sixth Framework Programme (2002 -2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (inc. Commission Services)	
RE	Restricted to group specified by consortium (inc. Commission Services)	
CO	Confidential only for members of the consortium (inc. Commission Services)	

Transport
Project co-financed by the European Commission, Directorate-General Transport and Energy

Table of Contents

Transport

Project co-financed by the European Commission, Directorate-General Transport and Energy

1. Executive Summary

The objective of Task 1.5 of the SafetyNet IP has been to estimate the actual numbers of road accident casualties in Europe from the CARE database by addressing two issues:

- the under-reporting in national accident databases and
- the differences between countries of the definitions used to classify injury severity.

Currently, the only comparable measurement units available in CARE are the numbers of fatal accidents and of people killed, where the degree of under-reporting is acceptably small in most EU Member States and there is a common definition. The same is not true, however, of non-fatal accidents and of casualties who are not killed. As a result, at present the numbers of non-fatal accidents and of people seriously and slightly injured cannot be compared in different Member States. In addition, the definition of injury severity differs among member states, so that a casualty which would be recorded in one country might not be recorded in another. Equally, a casualty which might be recorded as 'seriously' injured in one country might be recorded as 'slightly' injured in another.

As a result of this lack of comparability, international comparisons of road safety focus entirely on fatal accidents and fatalities, which form only a small minority of the totals. It is highly desirable to extend these comparisons to include the full range of injury severities. The objective of Task 1.5 has been to allow this to happen.

In order to overcome the inconsistencies in the reporting of non-fatal casualties, this Task has:

1. estimated the under-reporting level for non-fatal casualties by developing a uniform methodology and applying it in several EU countries,
2. estimated the number of serious casualties per country according to a new common measurement unit.

This report documents the results that have been achieved:

- The study began by agreeing a common methodology that would be applied by all partners in Task 1.5 for their studies.
- Studies were carried out in 8 countries according to this methodology, and the report contains detailed descriptions of the individual studies.
- In each study, files of police and hospital records were assembled for the road accidents that occurred in a common area. These files were compared to identify matching records, i.e. those casualties who were present in both files. For these matching records, certain medical details were added to the police records: length of stay in hospital and injury severity (specifically the Maximum Abbreviated Injury Score MAIS, an internationally accepted summary measure of injury severity).

- Two matrices were then prepared to summarise the outcome of each study, one based on injury severity and the other on length of stay.
- These matrices were brought together for analysis, and conversion factors for each study were estimated in a consistent way. These factors allow the actual number of serious casualties in each country to be estimated consistently from police accident statistics.
- The new common measurement unit for counting serious casualties could be based on either injury severity or length of stay. It is concluded that the most robust definition is of a non-fatal casualty with MAIS>=3 (inclusive).
- Initial comparisons have been made of casualty data adjusted by the conversion factors estimated by the national studies.

The various national studies encountered a range of problems concerning access to the hospital data and content of the data. In general these were overcome successfully, although there were some implications for the results that could finally be achieved.

The coverage of the studies varied widely, influenced to some extent by whether hospital data had to be collected directly (as in the Czech Republic and Hungary) or were already available from files that had been compiled by national or regional authorities. The size of the datasets varies widely, depending on the size of the study area and the period included. The studies are summarised below.

Country	Study area	Period
Austria	National	2001
Czech Republic	Local (Kromeriz)	2003 - 2005
France	Regional (Département of the Rhône)	1996 - 2003
Greece	Regional (Corfu)	1996 - 2003
Hungary	Local (part of Budapest)	Aug 2004 - Jan 2006
The Netherlands	National	1997 - 2003
Spain	Regional (Castilla y Leon)	July - Dec 2005
United Kingdom	Regional (Scotland)	1997 - 2005

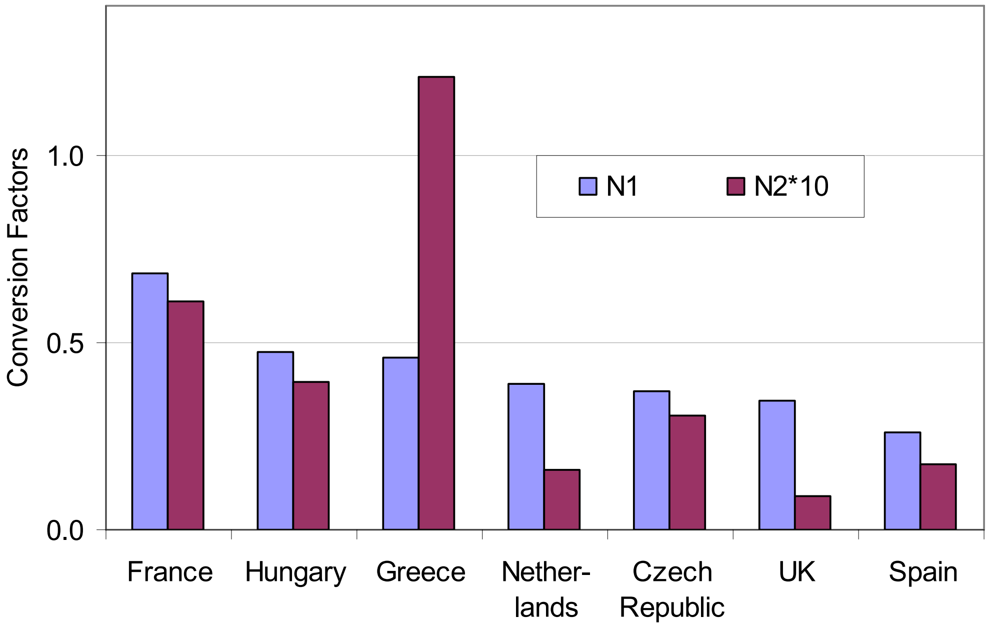
Ideally, these studies would have covered complete countries and so been truly national. Only 2 studies were truly national, so the question arises in the remaining 6 countries of whether conversion factors estimated from sub-national studies can be generalised to the national data. The answer must vary from country to country, but in general the larger the study area the more likely the conversion factors are to be nationally representative.

The new common measurement unit is a non-fatal casualty with MAIS>=3. Most of these are recorded by the police as seriously injured, but the studies show that the police record some as slightly injured. Consequently, according to this definition the number of casualties C in a particular country is estimated as

C = N1 * police reported serious casualties +
N2 * police reported slight casualties

where N1 and N2 vary from country to country. The overall factors from 7 studies are shown below (they could not be estimated in Austria because of data limitations). N2 is considerably smaller than N1 and hence is multiplied by 10 in this figure.

Conversion Factors for MAIS>=3, all road users



It was originally envisaged that the conversion factors would be generalised to other countries, in order to increase the utility of the CARE database. However, the results have led to the conclusion that this would not provide reliable results. The only satisfactory approach would be to carry out comparable studies in as many countries as possible.

The results from the Dutch and UK studies have also shown that the conversion factors can change through time as police accident reporting practices evolve. Thus, studies need to be repeated regularly to update the factors.

In summary, the research that has been carried out in the course of SafetyNet Task 1.5 represents a significant step forward and allows for the first time the number of severely injured casualties to be compared meaningfully between countries. The goals of the research were ambitious, but the practical problems that were encountered have meant that some could not be achieved fully. The lessons that have been learnt will allow this type of study to be carried out more effectively in future.

The central problem of this type of study is of obtaining access to anonymised medical records. Access to these records for research purposes is often problematical. Modern linkage techniques such as those used in this study, however, make these data increasingly valuable. Ways need to be found to persuade the custodians of these data to allow them to be used for purposes that support the broader aims and welfare of society.

2. Introduction

The objective of Task 1.5 of the SafetyNet IP is to estimate the **actual** numbers of casualties in Europe from the CARE database by addressing the issue of under-reporting and the differences in national systems for injury classification. Currently, the only comparable measurement units available in CARE are the numbers of fatal accidents and of people killed, where the degree of under-reporting is acceptably small in most EU Member States. The same is not, however, true of non-fatal accidents and of injured people who do not die, so at present the numbers of non-fatal accidents and of people seriously and slightly injured cannot be compared in different Member States. In addition, the definition of injury severity differs among member states, so that an accident or casualty which would be recorded in one country might not be recorded in another, while an accident or casualty which might be recorded as 'serious' in one country might be recorded as 'slight' in another,

The result is that at present international comparisons of the level of road safety rely almost exclusively upon the analysis of data for fatal accidents and fatalities. This is unsatisfactory, for by most criteria non-fatal accidents and casualties impose a burden on society that is at least as great as fatal accidents and fatalities. For example, using the British Government's cost-benefit value of prevention of road accidents, in 2005 fatal road accidents accounted for 37% of the cost-benefit value of preventing accidents that involved personal injury, and this falls to 27% when accidents involving material damage only are included. Thus, while fatal accidents and casualties form a major part of the burden of road accidents in a country, they by no means represent the totality.

In order to overcome the inconsistencies of reporting and permit non-fatal accidents and casualties to be analysed meaningfully, this Task has attempted to:

1. estimate the under-reporting level for each casualty severity (seriously injured, slightly injured) by developing a uniform methodology and applying it in eight EU countries.
2. estimate in each country the number of casualties according to a new common measurement unit.

The results from this Task will expand the scope of CARE-based road accident analyses. By allowing consideration of road safety to extend beyond its current focus on fatal accidents, the increased size of the data sets available for analysis will reduce the effects of chance, thereby permitting more detailed analyses to be carried out.

The structure of the work of Task 1.5 has been as follows:

1.5.1 Development of common methodology

1.5.2 Execution of National studies on under-reporting; these will be conducted by WP1 partners in eight EU countries (Austria, Czech Republic, France, Greece, Hungary, Netherlands, Spain, United Kingdom)

1.5.3 Elaboration of National under-reporting coefficients
1.5.4 Adoption of the common definition of hospitalised persons

Work has now been completed on all subtasks and this report presents the full results. Details of Subtasks 1.5.1 and 1.5.2 were presented in the Interim Progress Report (Deliverable D1.6), and these are summarised in the remainder of this section for completeness. The common methodology specified that the final data from the National studies should be provided in a certain common format, and Section 3 presents results prepared using a common format for all studies. These results allow Subtasks 1.5.3 and 1.5.4 to be completed in Section 4. Section 5 discusses the conclusions that can be drawn from this research and presents recommendations for future work in this area.

Each of the national studies represents a substantial body of work, with details varying between countries because of variations in the type of data that are available. Appendix A presents reports from the national studies, including results of more detailed analyses carried out in individual countries.

The national studies have been carried out for a specific purpose within the SafetyNet project, but it should be recognised that the linked data sets have a potential value that extends well beyond this purpose. In particular, clinical details from the original medical records are now available for large numbers of casualties reported by the police, and this enhanced information has considerable research potential – and indeed would be very difficult and expensive to collect in a specific research project. These potential applications will not be discussed farther in this report, but should be borne in mind when assessing the costs and benefits of this type of study.

Terminology

The CARE database uses three categories of injury severity: fatal, serious and slight. As it combines various national databases, the precise meaning of these terms varies by country, and indeed this is one of the main reasons for carrying out this research. Further, many studies have found that police officers sometimes do not apply the definition current in their country correctly when reporting an accident. Arguably, one might use quotation marks as a reminder of the possibility of mis-reporting in CARE, for example using “serious” to refer to casualties recorded in CARE as being seriously injured but including some who were actually fatally or slightly injured.

The possibility of mis-reporting of injury severity in police accident data is widely recognised. Most readers of this report will be aware of this issue so it seems unnecessary to use quotation marks.

It is useful at this point to mention one adaptation of the standard terminology that will be made in Section 4.5. A criterion is developed in Section 4 for defining a serious casualty in a uniform way in different countries, and this category will be termed serious* to differentiate it.

2.1 Background

As explained above, road accident reporting systems and standards differ among Member States, so it has been necessary to identify a common international standard to provide a benchmark with which to compare accident data from each country. The benchmark is achieved by developing the method that has been used several times in several countries to study the level of under-reporting. This consists of comparing:

- those road accident victims who have been recorded by the police in the national accident database, with
- those who have been recorded in medical records maintained by hospitals.

Fortunately, medical authorities are farther advanced than transport authorities with establishing international recording systems, in particular the International Classification of Diseases (ICD) and Abbreviated Injury Scale (AIS) coding systems. Hence, the basic approach adopted for this Task consists of

- linking road accident and medical databases, to investigate the level of under-reporting and, equally importantly, to copy details from the medical record of each linked casualty to the corresponding record in the accident database (Task 1.5.2)
- comparing the distributions of linked and unlinked casualties from the national studies by Maximum AIS (MAIS) and length of stay in hospital (Task 1.5.3)
- defining a new injury classification based on the most appropriate medical variable(s) and calculating national coefficients to estimate 'true' casualty totals from the numbers recorded in CARE (Task 1.5.4)

This has been an ambitious programme of work, and perhaps the greatest potential obstacle that was faced has been the difficulty of achieving access to suitable databases of medical information. These databases should ideally cover distinct, geographically well-defined regions, so that one may be confident that any road accident casualty recorded in the medical databases should – in the absence of under-reporting - also be recorded in the accident database for that region. There may also be ethical problems, since the databases may contain details that allow individuals to be identified. On the other hand, the data needed for the study are anonymous, the only personal details being age and sex which are needed to link to the accident records.

The linkage of road accident and medical databases is a relatively demanding task, involving a degree of judgement when specifying the differences that may be tolerated when deciding whether a pair of records actually refer to the same casualty. The level of prior experience of this type of work varied widely among the partners.

Similar studies had already been carried out in a number of other countries, such as the series of studies that was carried out in Western Australia (e.g.

Lopez et al, 1999). Indeed, in the USA this type of linkage has been carried out routinely for over a decade as part of the Crash Outcome Data Evaluation System (CODES) on State-wide data sets from 29 States (see http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/CODES.html). Thus, this type of linkage has been applied in a wide range of contexts, although the technical details can vary between studies. The SafetyNet application is thought to be the first time that the technique has been used to develop a common benchmark for comparing accident data from different countries.

Road accident victims with only slight injuries may not require significant medical treatment and hence would not be recorded in any medical database. The common methodology does not cover such cases, as it uses only official records. In the longer term it may be possible to develop survey-based methods to include them.

Ideally, under-reporting studies would be carried out in all countries that supply data to the CARE database, so the fact that the national studies will only involve 8 countries also presents difficulties. The pragmatic solution originally envisaged had been to use information provided by Governmental Experts (including the existence of corresponding data or studies in their countries) to draw up a system of analogues, i.e.

to identify, for each country *without* a national study, which of the countries *with* a national study it most closely resembles in terms of its national accident reporting system,

to generalise the coefficients estimated for each country *with* a national study to all analogous countries.

The information provided by Governmental Experts has not been sufficient for this to be done. Section 4 will review the data collected by the national studies and consider whether an alternative solution may exist.

Even if such a solution could be identified, any approach which generalises from studies in a minority of EU Member States is not fully satisfactory, and the long-term aim should be to conduct comparable studies in all Member States.

The common methodology

The broad problem that is addressed by Task 1.5 was described in the previous section, which also gave a general account of the approach that has been adopted. The common methodology that was finalised in May 2005 will now be presented. Problems experienced when applying the methodology are recorded in the reports in the Appendix and summarised in Section 4.

There are broadly two alternative methods for collecting the information from hospitals to be compared with accident data recorded by the local police:

1. Regional or national medical authorities may routinely assemble databases of records from which suitably detailed medical data can be extracted,

2. Medical data can be collected specifically for this project.

In order to carry out option 2, a representative sample of hospitals that receive accident victims would be selected, and the approvals needed for the data collection obtained from the medical authorities. During the period of data collection, project staff would regularly visit the hospital departments that receive accident victims. They would sift the records held in these departments to identify those people whose presenting history indicates that they had been injured in a road accident. A range of details would be recorded for each of these people, and subsequently entered in the project database.

The number of hospitals and the length of the data collection period would depend upon the funding available. In view of the limited budget allocated to Task 1.5, the collection of data in the volumes required to derive statistically reliable results may well not be affordable. In this case, only option 1 would be feasible.

With either option, the medical records are cross-checked regularly with the police accident records. The checking takes account of the catchment[1] area of each hospital, comparing the hospital records with police accident records only for that area. The aim is to identify all cases where the same person is present in both sets of records. Personal names are unlikely to be available in both data sets, in which case this process must be based on factors common to both data sets such as the casualty's age, sex and mode of travel, together with accident circumstances such as date, time and location.

The outcome is a combined set of police and medical data in which these matched cases are marked. The matching process should make allowance for minor errors in the recording of personal details, for example small discrepancies in age between the two sources. The process used may need to vary in detail from country to country to allow for local data and facilities.

Once the cross-checking of medical and police records has been completed, the proportion of accident casualties that has been reported by the police can be calculated. This will provide the level of under-reporting of casualties in the police data. This is likely to vary with type of accident (e.g. in relation to the number and type of vehicles involved), which should also be examined.

The medical data collected under either option must include those details that are needed to cross-check with police records. The principal extra item of data is whether or not the casualty was admitted to hospital, and if so for how long. Medical details should also be collected, provided ethical approval has been given by the medical authorities. The aim will be to record the maximum AIS[2] score for each body region. Under option 2 it will be possible to do this directly

[1] The area around the hospital from which accident victims are normally brought to the hospital for initial treatment

[2] Abbreviated Injury Scale, ranging from 1 for minor injuries to 6 for injuries that are currently untreatable

from the hospital records if the data collection staff have sufficient expertise, or it may be necessary to transcribe sufficient details for more expert staff to assess the AIS scores subsequently.

The combined police and medical data set are used to compile two 3-dimensional matrices of casualty counts. Matrix 1 is based on a casualty's length of stay in hospital, Matrix 2 is based on the severity of their injuries as summarised by the MAIS score (the maximum of the AIS scores per body region, as described in the following section). Road user type is identified from police data, as it is often poorly recorded in medical records.

Matrix 1

road user type		Length of Stay		police coding
car occupant pedestrian pedal cyclist motorcyclist other	X	out-patient overnight 1-3 days >3 days not coded *(not matched in medical records)*	X	fatal serious slight not coded *(not matched in police records)*

Matrix 2

road user type		MAIS		police coding
car occupant pedestrian pedal cyclist motorcyclist other	X	1-6 not coded *(not matched in medical records)*	X	fatal serious slight not coded *(not matched in police records)*

The 'Common Methodology' defines the strategy to be followed by each national study and the numerical outputs, but not the details. It was recognised at the outset that the type and availability of data varies from country to country, so it would be impossible to be more prescriptive. Inevitably, a range of linkage methods were applied, and Section 4 considers the implications. Full details of the individual studies and their linkage methods are provided in the Appendix.

The definition of Length of Stay had to be tightened in the course of the national studies:

Out-patient (0 nights, 1 part-day)
Overnight (1 night, 2 part-days)
1-3 days (2-4 nights)
>3 days (>4 nights)

The Greek study could not apply this definition exactly because of limitations in the source data, as explained in Section 7.4.

2.3 Injury and severity coding

It is useful at this point to review the way in which details of injuries are coded, as this is the basis for summarising the overall severity of a casualty's injuries via the MAIS. The details have evolved over many years, but there is a well-established international coding system.

The Abbreviated Injury Scale (AIS) is a specialised trauma classification of injuries based mainly on anatomical descriptors of the tissue damage caused by the injury (EGISM, 2004). The AIS classification system was designed to distinguish between types of trauma of clinical importance as well as types of trauma of interest to vehicle designers and research engineers. It has been shown to provide a good basis for valid measurement of probability of death. The AIS has two components: (1) the injury descriptor (often referred to as the 'pre-dot' code) which is a unique numerical identifier for each injury description; and (2) the severity score (can be referred to as the 'post-dot' code). The severity score ranges from 1 (minor) to 6 (currently untreatable), and is assigned to each injury descriptor. The AIS is based on anatomical injury, and not on physiological parameters (of the person injured). It implies that there is only a single AIS severity score for each injury, for any one person. The AIS scores injuries and not the consequences of injuries; it is not a measure of impairments that result from the injury. The AIS severity code is not simply a ranking of expected mortality from injury; it is based on potential for mortality but also on the diagnostic certainty, rapidity, duration, complexity and expected effectiveness of resolution with or without existing therapy (AAAM, 1990). The MAIS is the maximum AIS of all injury diagnoses for a person

The MAIS can be estimated directly by trained staff, but alternatively it can be derived from other classifications. The International Classification of Diseases (ICD) is a system designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics (DHHS; NCHS, 2007; WHO, 1992). It provides a way to classify medical terms reported by physicians, medical examiners and coroners on death certificates, also data from physicians' offices and hospital inpatient and outpatient records, so that they can be grouped together for statistical purposes.

In the United States, as in many other countries, the ICD is used to code and classify mortality data from death certificates. The ICD Clinical Modification (CM) is used to code non-fatal-injury data from medical records (i.e., hospital records, emergency department records, and physician office data). From 1979 to 1998, injury-related fatalities were coded using the 9th revision of the ICD (ICD-9) external cause of injury codes, more commonly referred to as E codes. In 1999, the 10th revision of the ICD (ICD-10) was implemented for coding deaths (DHHS). There is as yet no Clinical Modification of ICD-10. Countries have decided individually if and when to migrate from ICD-9 to ICD-10.

Under ICD-9, the external cause of injury or death was assigned an E code ranging from E800.0 to E999.9 based on information documented on the death certificate. External cause of injury codes describe the circumstances, such as a motor vehicle crash, drowning, or suffocation, as well as the intent of the injury (i.e. unintentional, homicide, suicide, intent undetermined, or other). In 1999, along with ICD-10, the injury E codes for fatalities were replaced with V, W, X, and Y cause codes along with special U codes for terrorism.

The ICD is developed collaboratively between the World Health Organization (WHO) and 10 international centres, to ensure that medical terms reported on death certificates are internationally comparable and lend themselves to statistical analysis. The ICD has been revised approximately every 10 years since 1900. These revisions reflect advances in the medical field and changes in our understanding of disease mechanisms and terminology, and are designed to maximise the amount of information and flexibility a code can provide. ICD-10 more closely reflects current medical knowledge than ICD-9.

One approach for using the ICD for severity assessment has been to develop the ICDMAP software that translates ICD-9-CM coded discharge diagnoses into AIS pre-dot codes, injury descriptors and severity scores (MacKenzie et al. 1997). The mapping does result in some loss of information due to differences in the injury classification systems. Resulting severity scores referred to as ICD/AIS scores are considered to be conservative measures of injury severity.

Until recently there was no software to convert the current ICD-10 codes to AIS. In 2006, Dr. Maria Seguí-Gómez from the Universidad de Navarra (Spain) developed a program that allows AIS to be coded from ICD-10 (ECIP, 2006). There is as yet no validation study that compares the effect on severity trends of changing from ICD-9-CM to ICD-10. An empirical adjustment is derived in Section 4.4 for road accident casualties with MAIS>=3.

Another approach is to estimate the severity of an injury based on the estimated probability of surviving that injury, *P*(survival), known as ICISS (Osler et al. 1996). To calculate this, first survival risk ratios (SRRs) are calculated for each diagnosis code as the proportion of cases with that diagnosis code who did not die. SRRs are calculated by dividing the number of survivors among patients with a specific ICD by the total number of patients with that ICD code. Each case is then assigned an ICISS, which is the product of SRRs of all their diagnoses. The resulting ICISS are estimates of *P*(survival) which range from 0 (unsurvivable) to 1 (certain to survive) (Stephenson et al, 2005).

The ICISS has also problems that need to be addressed, such as the fact that it is in some measure database-specific, or depends upon other injuries in multiple trauma cases. Nonetheless, the development of this approach to injury severity assessment is on-going and shows great promise. It could provide a valid alternative to MAIS in future.

3. Results

Two matrices are prepared in each national study for the purpose described in the Introduction: to estimate Conversion Factors that can be applied to police accident statistics, e.g. as held in the CARE database, in order to estimate national casualty totals according to two criteria, Length of Stay or MAIS. The calculations are set out in detail in Section 3.1 using data from the UK national study. Certain assumptions are required for the calculations, and these are presented and discussed in the context of these data.

Results from the other studies are presented in less detail in Section 3.2. More detailed results of interest from individual studies are presented in the Appendix, as part of the technical descriptions of the studies.

First, a general overview of the national studies is given in Table 1, while Figure 1 illustrates the eight countries involved.

Table 1: Summary details of studies

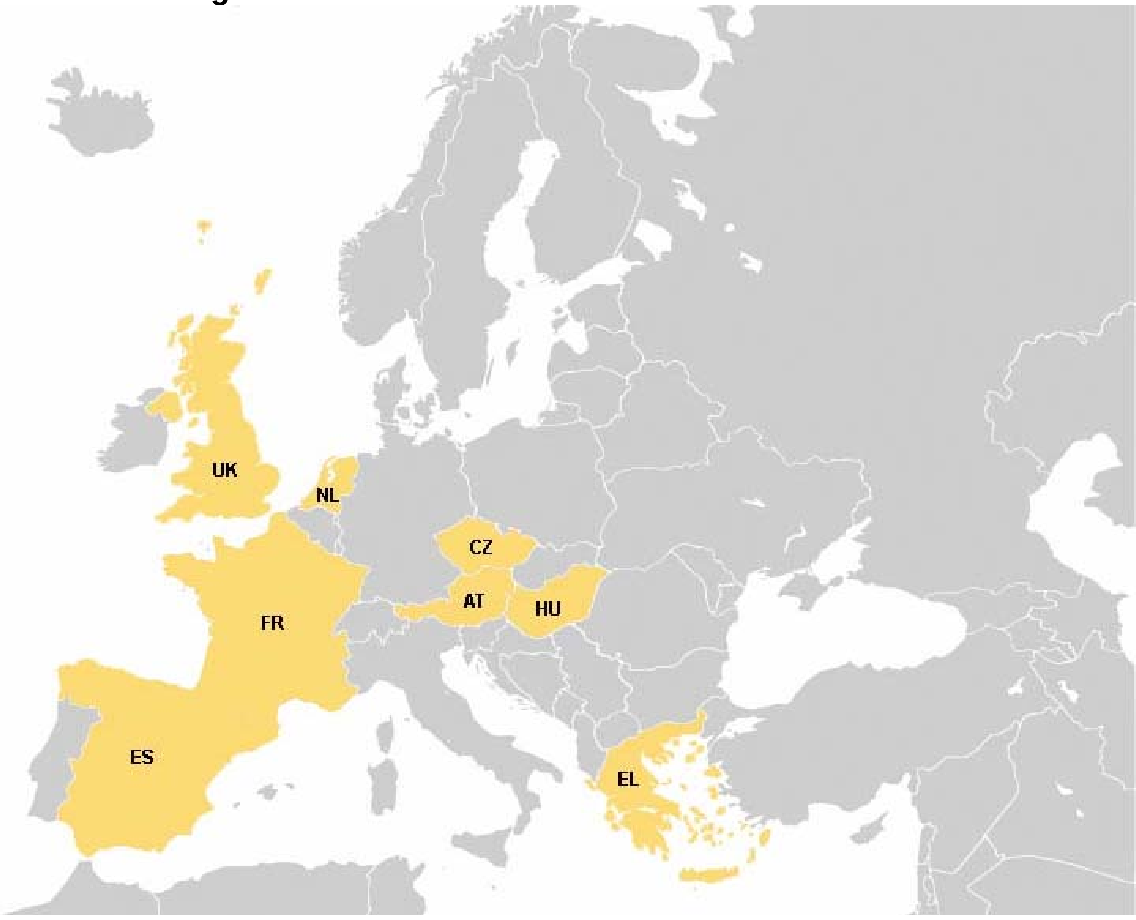
Country	Study area	Period	Coding of MAIS
Austria	National	2001	From ICD10
Czech Republic	Local (Kromeriz, central Moravia)	2003 - 2005	From ICD10
France	Regional (Département of the Rhône)	1996 - 2003	Coded directly
Greece	Regional (Corfu)	1996 - 2003	From ICD9
Hungary	Local (part of Budapest)	Aug 2004 - Jan 2006	Coded directly
Netherlands	National	1997 - 2003	From ICD9
Spain	Regional (Castilla y Leon)	July - Dec 2005	From ICD9
United Kingdom	Regional (Scotland)	1997 - 2005	From ICD10

The fundamental assumption

The fundamental assumption that underlies the calculations in the following Sections is that the medical and police data have been linked correctly, i.e. the links that have been made are valid, whereas records that have not been linked genuinely refer to different people. Clearly, the validity of this assumption depends upon the accuracy of the data in the two sets of records that are used for the linking process, but it is inescapable. This assumption needs to be borne in mind when reading the explanation of the calculations. Further assumptions are introduced as appropriate.

The accuracy of the linkage achieved could only be checked rigorously with access to the personal identifiers in the two sources of information for at least a subset of records. Such highly confidential information was not available to any of the national studies.

Figure 1: Countries where studies were carried out



3.1 Results from the UK study

The UK study linked records from the Scottish Hospital Inpatient System for 1997-2005 with STATS19 police accident records from Scotland. Full details are provided in Section 7.8. The results can be summarised as follows:

	Police	Not police
Hospital	26,198 casualties in SHIPS and STATS19	20,672 casualties in SHIPS but not in STATS19
Not hospital	151,165 casualties in STATS19 but not in SHIPS	unknown number of casualties neither in SHIPS nor in STATS19

The results of the linkage are summarised by road user type and police-reported casualty severity in Table 2.

Table 2: Linkage results, by road user type

		Police			Not	Grand
	Road user	Fatal	Serious	Slight	police	Total
Hospital	Car occupant	186	8,776	5,280	7,238	21,480
	Motorcyclist	26	1,979	620	2,736	5,361
	Pedal cyclist	18	697	422	4,018	5,155
	Pedestrian	180	5,255	2,008	2,888	10,331
	Unknown	0	0	0	1,229	1,229
	Other	17	727	434	2,563	3,741
	Subtotal	427	17,434	8,764	20,672	47,297
Not hospital	Car occupant	1,471	6,891	91,578		99,940
	Motorcyclist	334	1,393	5,225		6,952
	Pedal cyclist	79	709	6,480		7,268
	Pedestrian	516	2,578	20,822		23,916
	Other	144	1,260	14,229		15,633
	Subtotal	2,544	12,831	138,334		153,709
Grand Total		2,971	30,265	147,098	20,672	201,006

Length of Stay

The Length of Stay data will be analysed first. The overall results of the linkage are shown in Table 3, omitting fatal casualties. The proportion of casualties who were not reported by the police is lower among the more severely injured. The Length of Stay is reported for all SHIPS cases, so there are no unknown cases.

Table 3: The linkage results, by Length of Stay

	Length of Stay	Police		Not police	% not reported
		Serious	Slight		by police
Hospital	Outpatient	1,152	1,179	3,596	61%
	Overnight	4,434	4,336	7,219	45%
	1-3 days	4,690	2,132	4,880	42%
	>3 days	7,158	1,117	4,977	38%
Not hospital		12,831	138,334		
Grand Total		30,265	147,098	20,672	

The ‘police, not hospital’ casualties did not attend hospital, assuming that the record-linkage is perfect, so they are shown in Table 4 as ‘not in hospital’. The ‘hospital, not police’ casualties must be divided between the serious and slight categories so as to simulate the severity coding that the police would have used if they had been aware of these accidents, rather than the actual coding of serious. This is achieved by distributing the casualties for each Length of Stay (LoS) pro rata between the serious and slight categories. The calculation is as follows:

Let:

ser(i) = number of serious casualties reported by police with LoS=i

sli(i) = number of slight casualties reported by police with LoS=i

hnp(i) = number of casualties in hospital with LoS=i but not reported by police

then estimated total of serious(i) = ser(i)*[ser(i)+sli(i)+hnp(i)] / [ser(i)+sli(i)]

estimated total of slight (i) = sli(i)*[ser(i)+sli(i)+hnp(i)] / [ser(i)+sli(i)]

Table 4 presents the results. Capture/recapture methods have been used in some studies to estimate the number in the blank cell, i.e. casualties that were neither reported by police nor hospital. This has not been attempted here, but once a method was agreed it would be simple to update the calculation.

Table 4: Estimated results, by Length of Stay

Length of Stay	Police		Not police	Estimated total	
	Serious	Slight		Serious	Slight
Outpatient	1,152	1,179	3,596	2,929	2,998
Overnight	4,434	4,336	7,219	8,084	7,905
1-3 days	4,690	2,132	4,880	8,045	3,657
>3 days	7,158	1,117	4,977	11,463	1,789
Not in hospital	12,831	138,334		12,831	138,334
Total	30,265	147,098	20,672	43,352	154,683

The results show that, corresponding to each STATS19 serious casualty, 11,463/30,265=0.38 casualties were in hospital for more than 3 days, and corresponding to each slight casualty another 1,789/147,098=0.012 casualties were in hospital for more than 3 days. Such conversion factors can be used to estimate casualty totals from STATS19 casualty totals. For example, if serious casualties were to be defined as those staying more than 3 days in hospital then the actual total could be estimated as:

N = 0.38 x number of serious casualties reported by the police +

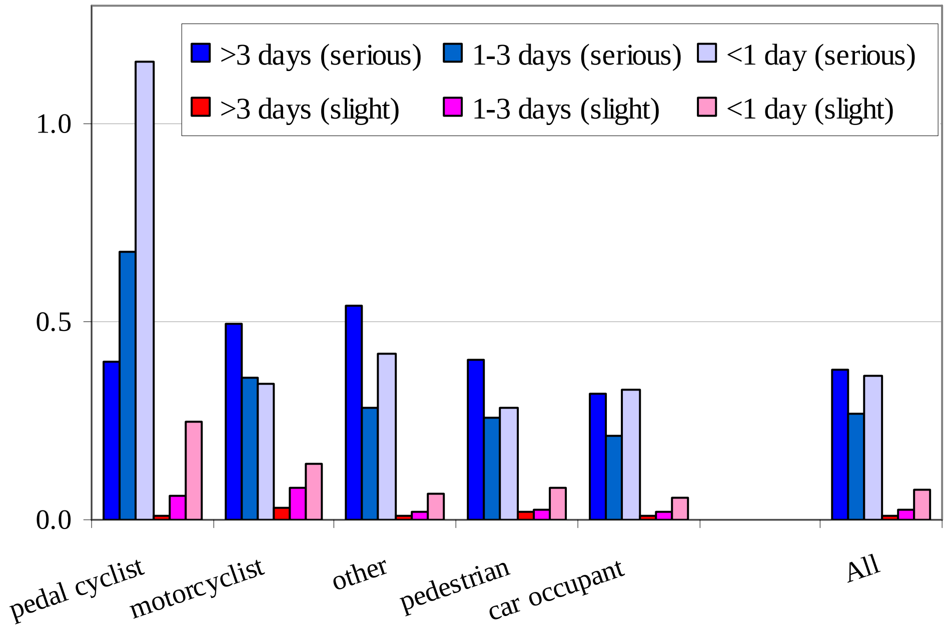
0.012 x number of slight casualties reported by the police

Note, however, that the calculation of the factors depends upon the period of the casualty data, as changes over time in hospital procedures are likely to affect the Length of Stay for any particular casualty. The conversion factors calculated by road user type are presented in Table 5 and illustrated in Figure 2. Table 5 is the first of the standard tables that will be used to compare results from the national studies.

Table 5: Conversion Factors based on Length of Stay

Length of Stay	Car Occupant		Motorcyclist		Pedal Cyclist		Pedestrian		Other		All	
	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
Outpatient/ overnight	0.33	0.06	0.34	0.14	1.16	0.25	0.28	0.08	0.42	0.06	0.36	0.074
1-3 days	0.21	0.02	0.36	0.08	0.68	0.06	0.26	0.03	0.29	0.02	0.27	0.025
>3 days	0.32	0.01	0.49	0.03	0.40	0.01	0.41	0.02	0.54	0.01	0.38	0.012
All	0.86	0.09	1.20	0.25	2.24	0.33	0.95	0.13	1.25	0.10	1.01	0.111
>=1 day	0.53	0.03	0.85	0.11	1.08	0.08	0.66	0.05	0.82	0.03	0.64	0.037

Figure 2: Conversion Factors based on Length of Stay



These Factors are averages for the 1997-2005 period. In order to estimate national totals with data from, say 2003-05, then slightly different values would be appropriate. This is considered further in Section 4.

MAIS

The overall results of matching the STATS19 casualty records from Scotland from 1997-2005 with the SHIPS data are shown in Table 6, omitting fatal casualties. The column '% not reported by police' is based on the three others, e.g. at MAIS 1 46%=6,294/(3,823+3,642+6,294). As with the data in Table 3 that were based on Length of Stay, the proportion of casualties who were not reported by the police is lower among the more severely injured.

Table 6: The linkage results, by MAIS

		Police		Not police	% not reported by police
	MAIS	Serious	Slight		
Hospital	1	3,823	3,642	6,294	46%
	2	8,336	2,473	8,050	43%
	3	3,139	412	2,227	39%
	4	226	33	220	46%
	5	75	1	44	37%
	6	197	19	134	38%
	9	1,638	2,184	3,703	49%
	1-9	17,434	8,764	20,672	
Not hospital		12,831	138,334		

The MAIS scores have been assigned from the ICD10 injury codes for each case. MAIS 9 is a code generated by the mapping algorithm that represents not known, i.e. the ICD10 codes were not sufficiently detailed to assign an MAIS score, e.g. ‘head injury’. 9.4% of serious casualties have MAIS 9, and 24.9% of slight casualties. These casualties appear on the whole to have relatively minor injuries, for example the incidence of MAIS 9 is lower among serious casualties than among slight and the % not reported by police is greater than for MAIS 1. The discussion in Section 2.3 suggests that the lack of a Clinical Modification of ICD-10 may explain the incidence of MAIS 9 casualties.

Excluding these cases would introduce one type of bias, tending to raise the apparent reporting level, while treating them all as MAIS 1 would introduce another type as they may well include cases with an actual MAIS of at least 3. On the whole, it appears preferable to treat the MAIS 9 cases as MAIS 1, so in the remainder of this report MAIS 9 has been grouped with MAIS 1.

The estimation process is more complex for MAIS than for Length of Stay in another respect. It was known that ‘police, not hospital’ cases had zero Length of Stay (assuming the record-linkage is perfect), but the MAIS of these cases must be estimated. These casualties have not required in-patient treatment, so it seems unlikely that MAIS will have exceeded 3. On the other hand, some MAIS 2 casualties could well be treated as outpatients, or in local doctors’ surgeries. It is reasonable to assume that all of these casualties had MAIS 1 or 2, but that they cannot be distributed reliably between 1 and 2. As with the calculation for Length of Stay, casualties that were not reported by the police are then distributed *pro rata* at each MAIS level to simulate the police severity coding, with the results shown in Table 7. Note that the calculation is carried for each MAIS value separately, then the results for values of 1, 2 and 9 are summed to form the first row. As noted previously, the SHIPS data cannot estimate reliably the actual number of slight casualties with MAIS 1 or 2.

Table 7: Estimated results, by MAIS

MAIS	Reported by police		Not reported by police	Estimated distri-bution of police, not hospital		Estimated total	
	Serious	Slight		Serious	Slight	Serious	Slight
1 or 2	13,797	8,299	18,047	12,831	138,334	37,647[1]	153,661[1]
3	3,139	412	2,227	0	0	5,108	670
4	226	33	220	0	0	418	61
5	75	1	44	0	0	118	2
6	197	19	134	0	0	319	31
All	17,434	8,764	20,672			43,610[1]	154,425[1]

[1] *likely to be underestimated*

Thus, it is estimated that for each serious casualty in the STATS19 records there are actually 5,963/30,265=0.20 casualties with MAIS>=3. Further, a small proportion of slight casualties in the SHIPS records actually had MAIS>=3, 764/147,098=0.005 per slight casualty. If serious casualties were to be defined as those with MAIS>=3 then the actual total (all types of road user) could be estimated as:

N_1 = 0.20 x number of serious casualties reported by the police +

0.005 x number of slight casualties reported by the police

Again, these values are averages for the 1997-2005 period, and slightly different values would apply for, say, the 2003-05, period.

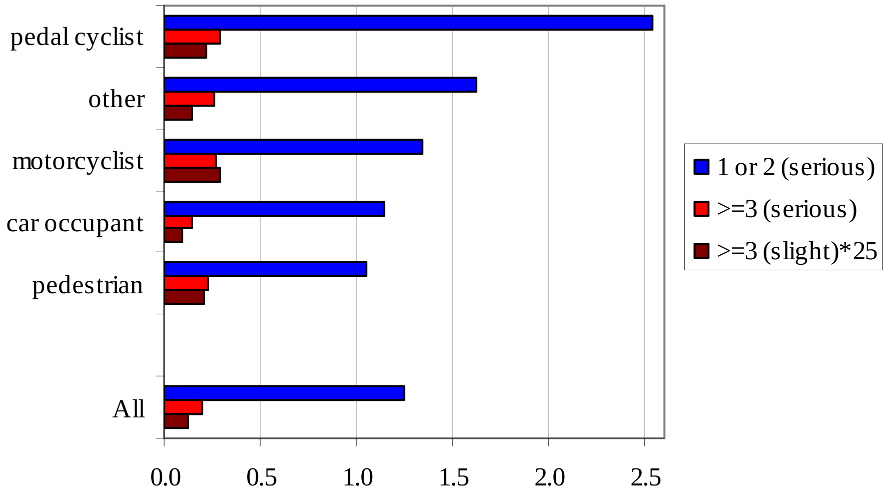
Table 8 presents the conversion factors by road user type. The results are illustrated in Figure 3, grouping MAIS>=3 together (factors for slight casualties are multiplied by 25 to facilitate visual comparison). The Figure emphasises that car occupants and pedestrians are recorded more fully in the Scottish STATS19 data than pedal cyclists. This is the other standard table that will be used to compare results from the national studies. Note, however, that certain of these results will need to be adjusted for reasons that are explained in Section 4.3.

Table 8: Conversion Factors based on MAIS

MAIS	Car Occupant		Motorcyclist		Pedal Cyclist		Pedestrian		Other		All	
	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
1 or 2[1]	1.15	1.03	1.34	1.13	2.54	1.24	1.05	1.03	1.62	1.06	1.24	1.04
3	0.13	0.00	0.25	0.01	0.26	0.01	0.18	0.01	0.23	0.01	0.17	0.00
4	0.01	0.00	0.01	0.00	0.02	0.00	0.03	0.00	0.01	0.00	0.01	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
6	0.01	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.01	0.00	0.01	0.00
All[1]	1.30	1.03	1.61	1.14	2.83	1.25	1.28	1.04	1.88	1.07	1.44	1.05
>=3	0.15	0.00	0.27	0.01	0.29	0.01	0.23	0.01	0.26	0.01	0.20	0.01

[1] *factors are likely to be underestimated*

Figure 3: Conversion Factors based on MAIS



3.2 Results from the other studies

This Section brings together results from the other national studies. The methods of calculation are identical to those used in the UK study, so the results are presented in less detail. The same structure is used in each case, but variations in the precise content of the data received mean that there are differences in detail.

The overall results from the various national studies are compared in Section 4.

Austrian study

The Austrian study was carried out with national data for the road accidents that occurred in the year 2001. The medical information comes from the national hospital discharge database, which combines administrative and medical data for all in-patients in the 270 Austrian hospitals. No out-patients are recorded in this database.

The medical records contain no information about the road user type of a casualty, so it is only possible to calculate overall conversion factors, i.e. not by road user type. Moreover, the records contain only one ICD code per casualty, whereas the software used to estimate MAIS uses up to 3 ICD codes. Consequently it has only been possible to calculate conversion factors by Length of Stay.

Table 9: The linkage results, Austria

	Police				Not police	Grand Total
	Fatally Injured	Severity unknown	Seriously Injured	Slightly Injured		
Hospital	104	1,382	2,882	1,689	12,010	18,067
Not hospital	854	5,187	5,325	39,800		51,166
Grand Total	958	6,569	8,207	41,489	12,010	69,233

Table 10: Conversion Factors based on Length of Stay, Austria

Length of Stay	Severity unknown	Seriously Injured	Slightly Injured
Overnight	0.20	0.17	0.051
1 - 3 days	0.26	0.33	0.051
>3 days	0.18	0.53	0.025
All	0.64	1.03	0.127
>=1 day	0.45	0.86	0.076

Czech study

The Czech study was carried out for the district of Kromeriz in the years 2003 – 2005. The town lies in central Moravia, about 70 km from Brno, and has a population of 30,000 inhabitants. There is one hospital, which is the source of the medical data.

Table 11: The linkage results, by road user type, Czech Republic

		Police		Not police	Grand Total
		Seriously injured	Slightly injured		
Hospital	Car occupant	18	126	76	220
	Motorcyclist	4	14	15	33
	Pedal cyclist	10	49	422	481
	Pedestrian	8	18	62	88
	Other	1	10	0	11
	Subtotal	41	217	575	833
Not hospital	Car occupant	64	441		505
	Motorcyclist	18	54		72
	Pedal cyclist	20	110		130
	Pedestrian	20	44		64
	Other	7	38		45
	Subtotal	129	687		816
Grand Total		170	904	575	1,649

Table 12: Conversion Factors based on Length of Stay, Czech Republic

Length of Stay	Car occupant		Motorcyclist		Pedal cyclist		Pedestrian		Other		All	
	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
Overnight	0.10	0.28	0.19	0.34	0.71	2.57	0.19	0.98	0.00	0.19	0.23	0.72
1-3 days	0.09	0.06	0.00	0.04	0.30	0.15	0.07	0.16	0.13	0.02	0.11	0.08
>3 days	0.11	0.01	0.09	0.01	0.50	0.02	0.31	0.02	0.00	0.00	0.19	0.02
All	0.29	0.35	0.28	0.39	1.51	2.74	0.57	1.16	0.13	0.21	0.53	0.82
>=1 day	0.19	0.07	0.09	0.06	0.80	0.17	0.38	0.18	0.13	0.02	0.30	0.09

Table 13: Conversion Factors based on MAIS, Czech Republic

MAIS	Car occupant		Motorcyclist		Pedal cyclist		Pedestrian		Other		All	
	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
1 or 2[1]	0.97	1.11	1.03	1.17	1.11	3.50	1.05	1.77	0.88	1.00	1.07	1.56
3	0.07	0.01	0.05	0.01	0.30	0.04	0.31	0.04	0.13	0.00	0.15	0.02
4	0.01	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.03	0.00
5	0.02	0.00	0.05	0.00	0.03	0.00	0.04	0.00	0.00	0.00	0.03	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
All[1]	1.08	1.12	1.12	1.18	1.61	3.54	1.40	1.80	1.00	1.00	1.28	1.58
>=3	0.11	0.01	0.09	0.01	0.50	0.04	0.35	0.04	0.13	0.00	0.21	0.02

[1] *factors are likely to be underestimated*

French study

The French study was carried out with data for the eight years 1996-2003 from the département of the Rhône, an area of 1.6 million inhabitants consisting of the city of Lyon, its suburbs and a rural area to the north. A road trauma registry has operated in the département since 1995, covering all road accident casualties who seek medical attention in health facilities. These data were linked with police records, and the results of the linkage are summarised by road user type in Table 14.

Table 14: The linkage results, by road user type, France

		Police		Not police	Grand Total
	Road user	Seriously injured	Slightly injured		
Hospital	Car occupant	1,703	11,195	28,212	41,110
	Motorcyclist	983	2,814	11,702	15,499
	Pedal cyclist	153	591	9,982	10,726
	Pedestrian	734	2,334	4,333	7,401
	Other	74	682	2,250	3,006
	Subtotal	3,647	17,616	56,479	77,742
Not hospital	Car occupant	544	7,509		8,053
	Motorcyclist	304	1,718		2,022
	Pedal cyclist	59	318		377
	Pedestrian	279	1,430		1,709
	Other	29	525		554
	Subtotal	1,215	11,500		12,715
Grand Total		4,862	29,116	56,479	90,457

Table 15: Conversion Factors based on Length of Stay, France

Length of Stay	Car occupant		Motorcyclist		Pedal cyclist		Pedestrian		Other		All	
	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
Overnight	1.12	1.91	1.28	2.61	4.82	9.82	0.79	1.44	1.16	2.18	1.21	2.23
1-3 days	0.18	0.04	0.19	0.11	0.58	0.31	0.14	0.07	0.41	0.05	0.20	0.06
>3 days	0.48	0.03	0.67	0.09	1.22	0.13	0.48	0.08	0.90	0.05	0.57	0.05
All	1.78	1.98	2.14	2.81	6.62	10.26	1.41	1.59	2.47	2.28	1.98	2.34
>=1 day	0.66	0.07	0.86	0.20	1.80	0.44	0.62	0.14	1.31	0.10	0.77	0.11

Table 16: Conversion Factors based on MAIS, France

	Car occupant		Motorcyclist		Pedal cyclist		Pedestrian		Other		All	
MAIS	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
1 or 2[1]	1.32	2.38	1.35	3.13	4.69	10.39	1.01	1.90	1.52	2.67	1.43	2.69
3	0.35	0.03	0.69	0.11	1.64	0.27	0.43	0.08	0.69	0.05	0.52	0.05
4	0.12	0.00	0.10	0.01	0.26	0.00	0.10	0.01	0.30	0.01	0.12	0.01
5	0.05	0.00	0.05	0.00	0.07	0.00	0.03	0.00	0.07	0.00	0.05	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
All[1]	1.84	2.41	2.18	3.25	6.67	10.66	1.58	2.00	2.58	2.73	2.11	2.75
≥3	0.51	0.03	0.83	0.12	1.97	0.27	0.57	0.10	1.06	0.06	0.68	0.06

[1] *factors are likely to be underestimated*

Greek Study

The Greek study was carried out with data for 1996–2003 from the Island of Corfu. The island has a population of approximately 110,000 people and is located in the North Ionian Sea. The study was based on data from the Greek Emergency Department Injury Surveillance System (EDISS) which collects data from the Regional Hospital of Corfu (the only public hospital on the island) and the Road Traffic Police database.

Table 17: The linkage results, by road user type, Greece

		Police		Not	Grand
	Road user	Seriously injured	Slightly injured	police	Total
Hospital	Bicyclist	2	4	98	104
	Driver	141	615	5,949	6,705
	Passenger	43	191	1,375	1,609
	Pedestrian	30	94	624	748
	Unknown	24	118	1,959	2,101
	Subtotal	240	1,022	10,005	11,267
Not hospital	Driver	54	315		369
	Passenger	26	153		179
	Pedestrian	22	78		100
	Subtotal	102	546		648
Grand Total		342	1,568	10,005	11,915

Table 18: Conversion Factors based on Length of Stay, Greece

Length of	Driver		Passenger		Pedestrian		Unknown		All	
Stay	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
Overnight	3.83	4.46	2.38	2.97	0.37	2.76	15.65	13.33	3.91	4.63
1-3 days	1.35	1.22	0.88	0.78	0.85	0.70	0.45	0.94	1.15	1.04
>3 days	0.55	0.33	0.36	0.20	0.56	0.35	0.10	0.23	0.49	0.30
All	5.73	6.01	3.62	3.95	1.78	3.81	16.20	14.51	5.55	5.97
>=1 day	1.90	1.55	1.24	0.98	1.41	1.05	0.56	1.17	1.64	1.34

Table 19: Conversion Factors based on MAIS, Greece

	Driver		Passenger		Bicyclist		Pedestrian		Unknown		All	
MAIS	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight	Serious	Slight
1 or 2[1]	4.62	6.40	3.19	4.42	8.33	20.58	2.49	3.91	11.82	15.17	4.54	6.39
3	0.47	0.11	0.27	0.08	0.00	1.00	0.31	0.13	0.53	0.07	0.40	0.11
4	0.06	0.01	0.00	0.00	0.17	0.17	0.13	0.00	0.17	0.02	0.06	0.01
5	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
All[1]	5.14	6.53	3.46	4.50	8.50	21.75	2.93	4.04	12.51	15.26	5.00	6.51
>=3	0.53	0.13	0.27	0.09	0.17	1.17	0.45	0.14	0.69	0.09	0.46	0.12

[1] factors are likely to be underestimated

Hungarian study

The study in Hungary was carried out with data from the Károlyi Sándor Hospital in Budapest, one of the 4 Regional Trauma Centres in the city. The accidents occurred between 1 August 2004 and 31 January 2006.

Table 20: The linkage results, by road user type, Hungary

		Police			Not police	Grand Total
	Road user	Fatally injured	Seriously injured	Slightly injured		
Hospital	Car Occupant	2	67	202	273	544
	Motorcyclist	0	52	44	182	278
	Pedal Cyclist	0	10	7	219	236
	Pedestrian [1]	1	45	62	89	197
	Other	0	0	8	31	39
	Subtotal	3	174	323	794	1,294
Not hospital	Vehicle Occupant	54	374	1,359		1,787
	Pedestrian	29	111	238		378
	Subtotal	83	485	1,597		2,165
Grand Total		86	659	1,920	794	3,459

[1] includes 8 casualties recorded by police as vehicle occupants

Table 21: Conversion Factors based on Length of Stay, Hungary

Length of Stay	Vehicle occupant		Pedestrian		All	
	Serious	Slight	Serious	Slight	Serious	Slight
Overnight	0.12	0.34	0.04	0.17	0.10	0.31
1-3 days	0.10	0.13	0.08	0.16	0.10	0.13
>3 days	0.35	0.032	0.37	0.069	0.35	0.038
Total	0.57	0.50	0.48	0.40	0.55	0.49
>=1 day	0.45	0.16	0.44	0.22	0.45	0.17

Table 22: Conversion Factors based on MAIS, Hungary

	Vehicle Occupant		Pedestrian		All	
MAIS	Serious	Slight	Serious	Slight	Serious	Slight
1+2[1]	0.83	1.28	0.86	1.16	0.84	1.27
3	0.43	0.04	0.22	0.02	0.38	0.04
4	0.06	0.00	0.08	0.00	0.06	0.00
5	0.02	0.00	0.05	0.00	0.03	0.00
6	0.00	0.00	0.01	0.00	0.00	0.00
All[1]	1.35	1.33	1.21	1.19	1.32	1.31
>=3	0.52	0.04	0.35	0.03	0.48	0.04

[1]factors are likely to be underestimated

Note that, as with Table 8, certain of these results will need to be adjusted for reasons that are explained in Section 4.3.

Dutch Study

The Dutch study was carried out with data for 1997-2003 from the whole of the Netherlands. Rather than serious casualties, the data refer to hospitalised (Hosp) casualties.

Table 23: The linkage results, by road user type, the Netherlands

		Police		Not police	Grand
	Road user	Hosp.	Slight		Total
Hospital	Car/van occupant	21,176	4,590	12,570	38,336
	Motorcyclist	3,847	831	3,266	7,944
	Moped	9,230	2,477	11,146	22,853
	Pedal cyclist	10,323	2,732	32,006	45,061
	Pedestrian	3,620	781	3,693	8,094
	Other	539	86	672	1,297
	Subtotal	48,735	11,497	63,353	123,585
Not hospital	Car/van occupant	1,479		1,309	2,788
	Motorcyclist	222		198	420
	Moped	599		530	1,129
	Pedal cyclist	644		570	1,214
	Pedestrian	213		188	401
	Other	48		31	79
	Subtotal	3,205		2,826	6,031
Grand Total		51,940	11,497	66,179	129,616

Table 24: Conversion Factors based on Length of Stay, the Netherlands

Length of	Car occupant		Motorcyclist		Moped rider		Pedal cyclist		Pedestrian		Other		All	
Stay	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight
Overnight	0.13	0.017	0.10	0.022	0.14	0.017	0.22	0.029	0.12	0.022	0.08	0.009	0.15	0.021
1-3 days	0.40	0.036	0.46	0.067	0.49	0.045	0.89	0.083	0.49	0.061	0.39	0.029	0.52	0.051
>3 days	0.39	0.014	0.68	0.049	0.69	0.030	1.22	0.055	0.71	0.046	0.39	0.014	0.65	0.029
All	0.92	0.067	1.25	0.138	1.31	0.092	2.33	0.167	1.31	0.129	0.86	0.052	1.32	0.101
>=1 day	0.78	0.050	1.14	0.116	1.17	0.075	2.11	0.138	1.20	0.107	0.78	0.043	1.17	0.080

Table 25: Conversion Factors based on MAIS, the Netherlands

	Car occupant		Motorcyclist		Moped rider		Pedal cyclist		Pedestrian		Other		All	
MAIS	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight	Hosp.	Slight
1[1]	0.19	0.019	0.12	0.021	0.14	0.014	0.23	0.022	0.14	0.020	0.16	0.014	0.18	0.019
2	0.41	0.029	0.71	0.088	0.72	0.057	1.27	0.099	0.74	0.080	0.42	0.026	0.68	0.056
3	0.20	0.006	0.31	0.020	0.35	0.014	0.70	0.031	0.33	0.019	0.19	0.007	0.34	0.015
4	0.02	0.001	0.03	0.001	0.03	0.001	0.05	0.002	0.04	0.002	0.02	0.001	0.03	0.001
5 or 6	0.02	0.000	0.03	0.001	0.02	0.001	0.02	0.000	0.02	0.000	0.01	0.000	0.02	0.000
All[1]	0.83	0.055	1.20	0.131	1.26	0.086	2.27	0.154	1.26	0.122	0.80	0.048	1.25	0.090
>=3	0.23	0.007	0.37	0.022	0.40	0.015	0.77	0.033	0.38	0.021	0.23	0.008	0.39	0.016

[1]*factors are likely to be underestimated*

Spanish study

Two linkage studies have been carried out in Spain, one in the city of Barcelona and the other in the rural region of Castilla y Leon. It was concluded that the results from Barcelona were not representative, so they will not be presented. Only overall results could be prepared in either study, i.e. not by road user type. The reason is illustrated by Table 26: the road user type is unknown in the medical records. The data come from July-December 2005.

Table 26: The linkage results, by road user type, Castilla y Leon

		Police				Not police	Grand
		Fatal	Serious	Slight	Not coded		Total
Hospital	Car occupant	6	176	104			286
	Pedestrian	4	47	13			64
	Pedal cyclist	0	9	4			13
	Motor cyclist	2	56	10			68
	Other	1	38	23			62
	Unknown	0	0	0		1,143	1,143
	Subtotal	13	326	154		1,143	1,636
Not hospital	Car occupant	155	757	3,460			4,372
	Pedestrian	17	105	316			438
	Pedal cyclist	5	28	79			112
	Motor cyclist	20	183	432			635
	Other	37	210	644			891
	Unknown	0	0	2	27		29
	Subtotal	234	1,283	4,933	27		6,477
Grand Total		247	1,609	5,087	27	1,143	8,113

Table 27: Conversion Factors based on Length of Stay, Castilla y Leon

Length of Stay	Seriously injured	Slightly injured
Overnight	0.01	0.007
1-3 days	0.19	0.051
>3 days	0.46	0.050
All	0.67	0.107
>=1 day	0.66	0.101

Table 28: Conversion Factors based on MAIS, Castilla y Leon

	Serious	Slight
1 or 2[1]	1.22	1.06
3	0.16	0.01
4	0.08	0.00
5	0.03	0.00
6	0.00	0.00
All[1]	1.48	1.07
>=3	0.26	0.02

[1] *factors are likely to be underestimated*

4. Synthesis

This Section brings together the results of the national studies presented in the previous section. First, however, the studies themselves will be reviewed briefly, drawing upon the detailed report of each study in the Appendix.

Predictably, all studies used accident data from national accident databases that had been compiled from police accident reports. Most studies used files of medical data compiled by national or regional authorities from hospital records. The medical files available in the Czech Republic and Hungary were not sufficient to carry out the study, however, so the medical data had to be assembled from hospital records specifically for these studies. This is clearly a more expensive process and has restricted the scale of these studies. The national medical file available in Austria only included a subset of the injury codes, which hampered that study considerably.

There were problems of obtaining access to medical records in most studies, although these were overcome successfully. It is worth recalling, however, that it was originally envisaged that a study would be carried out in Belgium, but this had to be aborted when the Belgian partner found it impossible to negotiate access to the necessary data.

Access to anonymised medical records for research purposes is often problematical. Modern linkage techniques such as those used in this project, however, make these data increasingly valuable. Ways need to be found to persuade the custodians of these data to allow them to be used for purposes that support the broader aims and welfare of society. As the extent to which the number and severity of road accidents recorded in national databases represents reality comes under greater scrutiny, the need for this type of study will increase.

While the concept of linkage as described in Section 2.2 is straightforward, a variety of approaches was adopted by the partners. It was originally envisaged that there would be an exploratory phase of the project where these could be compared and harmonised if appropriate. Resource and time constraints meant that this was not possible, but it would certainly be important to include this in any future research of this type.

Nevertheless, the various linkage approaches were applied rigorously and all work in similar ways using the same variables to identify potential matches, so there is no reason to suppose that the results would have differed significantly if a common, optimised technique had been applied in all studies.

As indicated by Table 1, the extent of the eight studies varies widely in time and space: from the whole of the Netherlands from 1997-2003 to the Czech town of Kromeriz in 2003–05. Similarly, the size of the combined datasets varies widely, from 1.6 thousand records from the Czech study to 201 thousand records from Scotland. Summary details of the studies are shown in Table 29.

Table 29: Summary of linking results.

	Linked police and hospital records, by police severity				Hospital not police records	Police not hospital records				Neither police nor hospital records	Total
	Fatal	Serious	Slight	Unkn		Fatal	Serious	Slight	Unkn		
AU	104	2,882	1,689	1,382	12,010	854	5,325	39,800	5,187		69,233
CZ		41	217		575		129	687			1,649
FR		3,647	17,616		56,479		1,215	11,500			90,457
GR		240	1,022		10,005		102	546			11,915
HU	3	174	323		794	83	485	1,597			3,459
NL		48,735	11,497		63,353		3,205			2,826	129,616
ES	13	326	154		1,143	234	1,283	4,933	27		8,113
UK	427	17,434	8,764		20,672	2,544	12,831	138,334			201,006

It is inevitable that the strength of the results achieved by various studies differs, certainly on statistical grounds and also potentially on the degree to which results from a part of a country may be nationally representative. Overall, however, the results achieved represent an important step forward in comparing the numbers of road accidents and casualties across a range of countries.

4.1 MAIS and LoS compared

One of the main aims of Task 1.5 is to recommend a definition of ‘serious casualty’ for use in international comparisons of casualty data from the CARE database. When the common methodology was being defined, it was seen that the broad choice lay between a definition based on the Length of Stay of road accident casualties in hospital, and a definition based on their MAIS scores.

Medical authorities tend to be critical of the use of Length of Stay as an indicator of injury severity, e.g. Brasel et al (2007). To the medical layman, it certainly appears that Length of Stay is likely to be influenced far more by clinical practices and the availability and organisation of hospital services than by the level of road safety. It appears that results based on MAIS are more likely to monitor casualty and severity trends reliably than results based on Length of Stay.

Results from the Scottish linkage study provide information that is highly relevant to this choice of basis. The trends in the linked data between 1980 and 2005 show how MAIS and Length of Stay for road accident casualties have developed over a quarter of a century. The operational procedures for SHIPS were unchanged between 1997 and 2005, and indeed for many years previously, so any changes in the annual data cannot result from changes in the SHIPS data collection procedure. They must be caused by changes in the number and nature of casualties, or in the criteria used to admit, treat and discharge hospital in-patients.

Figure 4 compares the distribution of Length of Stay in the linked casualty data with three ranges: 0 days (admitted and left hospital on the same day), 1-3 days and over 3 days. Note that the y-axis scales differ. This uses the definition of Length of Stay that was applied in the earlier study (Keigan et al, 1999) for consistency, so the data from the current study have been recalculated using these categories. There are clear overall trends, with a shift towards shorter stays in hospital.

Figure 4: Distribution of Length of Stay in Scottish linked casualty data

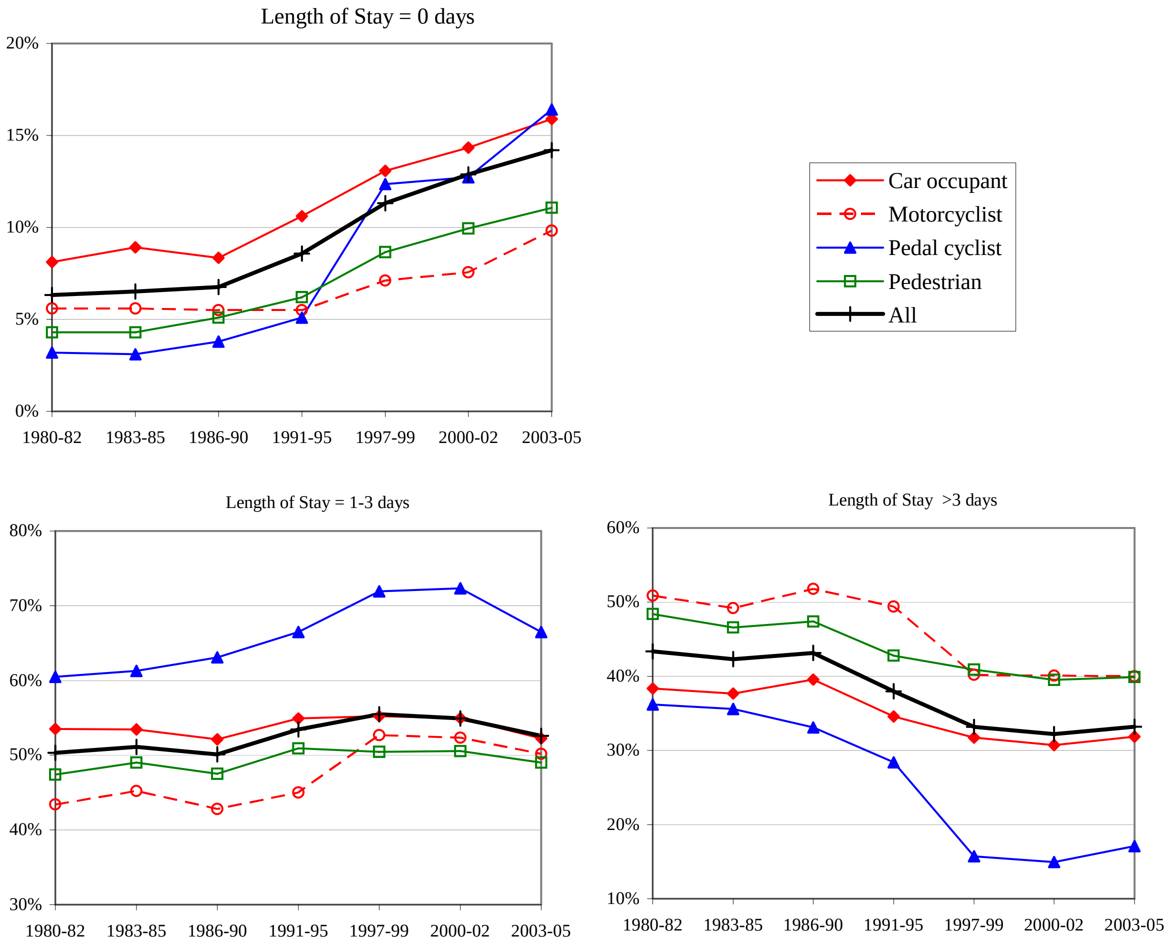
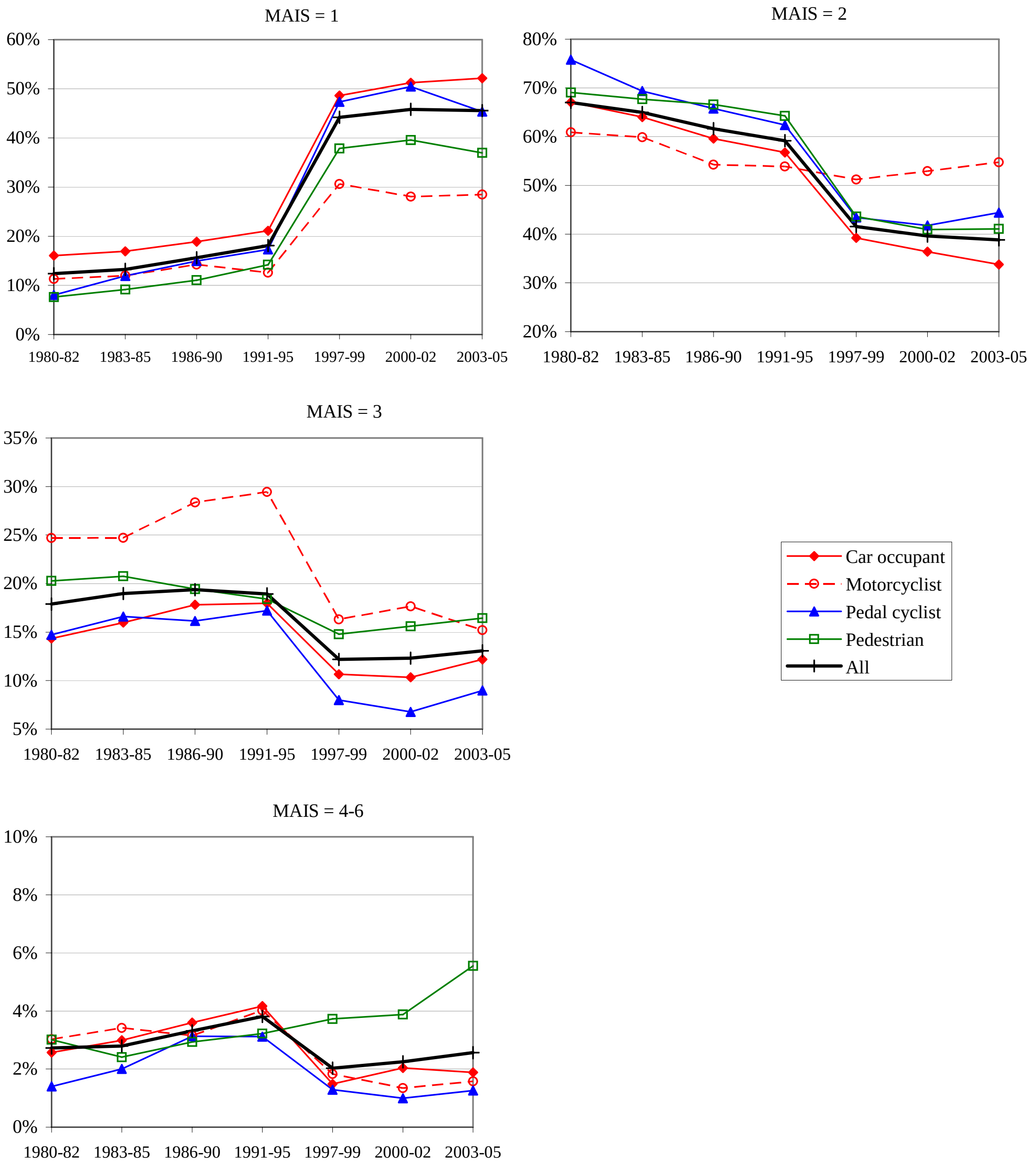


Figure 5 presents the corresponding comparison for MAIS. This comparison is affected by the switch from ICD9 to ICD10: SHIPS used the ICD9 system until 1996, while the ICD10 system was introduced in 1997 and a new mapping from ICD to MAIS had to be adopted. The Figure shows major increases in the proportion of casualties with MAIS 1 between 1991-95 and 1997-99, and corresponding reductions with higher MAIS. It is clear that the combination of the ICD10 codes and the new mapping has tended to yield lower MAIS scores.

Figure 5: Distribution of MAIS in Scottish linked casualty data



The ICD10 system is likely to provide less accurate estimates of injury severity than ICD9, for the reasons discussed in Section 2.3. Nevertheless, provided that the bias in the MAIS values calculated from the ICD10 codes is consistent between countries and years, it should still provide a valuable benchmark for international comparisons of road accident data.

The Figure does indicate that results based on mapping ICD9 codes to MAIS are not comparable with results based on mapping ICD10 codes. Before and after the switchover, however, the trends show a consistent pattern which is likely to reflect changes in road safety rather than in external influences. Unfortunately, Table 1 shows that both systems have been used in the national studies. Section 4.4 adjusts the ICD10-based conversion factors to ICD9 in an attempt to achieve consistency.

4.2 Definition of ‘hospitalised’ person

The title of the final stage of Task 1.5 in the SafetyNet programme of work is “Adoption of the common definition of hospitalised persons”. The underlying aim of the Task is to achieve a definition of a serious casualty that could be applied in all countries. At the time the programme was being developed, it seemed that the most likely outcome would be a definition such as “spent more than 3 days in hospital”: hence the use of the term hospitalised.

Overall, however, the evidence from the Scottish linkage study supports other arguments against basing the new definition on Length of Stay, and in favour of adopting a definition based on MAIS. It would be perverse, then, to adopt a definition based on Length of Stay simply because of the terminology used in 2003 when preparing the programme.

The only remaining decision concerns the MAIS range to choose for the definition of serious casualty. In principle, the threshold could be taken as 2, since AIS 2 describes a moderate injury and indeed there are appreciable numbers of cases of casualties who die with MAIS=2. The development of the estimation procedure in Section 3, however, explains that it is not possible to estimate MAIS 1 and 2 separately with the data available in some countries, so the minimum feasible value for the threshold is 3. Coincidentally, the AIS documentation refers to 3 as a serious injury. The conversion factors from the national studies show that adopting a higher threshold would yield rather small numbers of serious casualties.

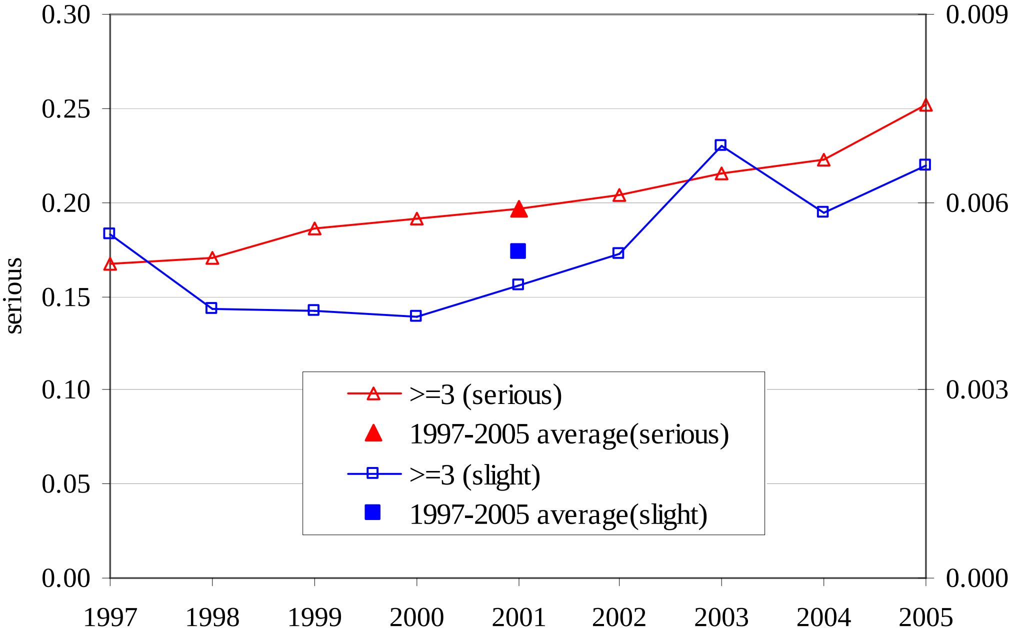
Recommendation

Accordingly, it is concluded that the optimal definition of serious casualty for use with the CARE database should be a non-fatal casualty with MAIS between 3 and 6 (inclusive).

4.3 Trends in the Conversion Factor

The results of Figure 5 from the Scottish linkage between 1980 and 2005 show clear trends in the MAIS distribution over time, and the detailed results from the matching process in Section 7.8 show that the relationship between the casualty data recorded by the police and by the hospitals is also dynamic. Thus, it is likely that Conversion Factors calculated annually would differ from the averages for 1997-2005 that were presented in Section 3.1. This is confirmed by Figure 6, which presents the annual factors used to estimate the number of casualties with MAIS>=3 from the number of serious and slight casualties in the STATS19 data. The factors for converting slight casualties are less than for serious, so use the right-hand scale.

Figure 6: Annual MAIS-based Conversion Factors in Scottish linked casualty data



Although the annual results are more susceptible to random variation than the results for nine years of data, it is clear that there are trends in the factors. Ideally, the conversion factors should be recalculated regularly to take account of potential variations in the casualty data and their relationship with the medical data.

4.4 Adjustment of ICD10-based results

It was seen in Section 4.1 that MAIS scores based on ICD10 data are not comparable with MAIS scores based on ICD9 data, and that ICD9-based scores are likely to be more reliable. Table 1 shows that the Czech and UK national studies have used ICD10 data while three used ICD9 (the direct coding system of the French and Hungarian studies is comparable with ICD9). Hence, it is preferable to adopt ICD9 for the comparison, and a method is needed to adjust the Czech and UK results to be comparable with ICD9-based results.

As yet there has been no study to compare the effect on severity trends of changing from ICD9 to ICD10, so the only feasible approach is to make use of the trends in the SHIPS/STATS19 data. Figure 7 extends Figure 5 to show the proportion of linked casualties with MAIS>=3. A simple statistical model has been used to extrapolate the 1980-95 data and estimate the proportions that would have been found in 1997-99 if ICD9 had been used. For each road user group, the estimated proportion would have been greater than the actual proportion and this would have continued through the following years, so the conversion factors in Table 8 for MAIS>=3 would have been greater if ICD9 had been used. Table 30 shows the adjustments necessary, and applies these to the factors from the last row of Table 8.

Project co-financed by the European Commission, Directorate-General Transport and Energy

Figure 7: Proportion of Scottish linked casualties with MAIS>=3

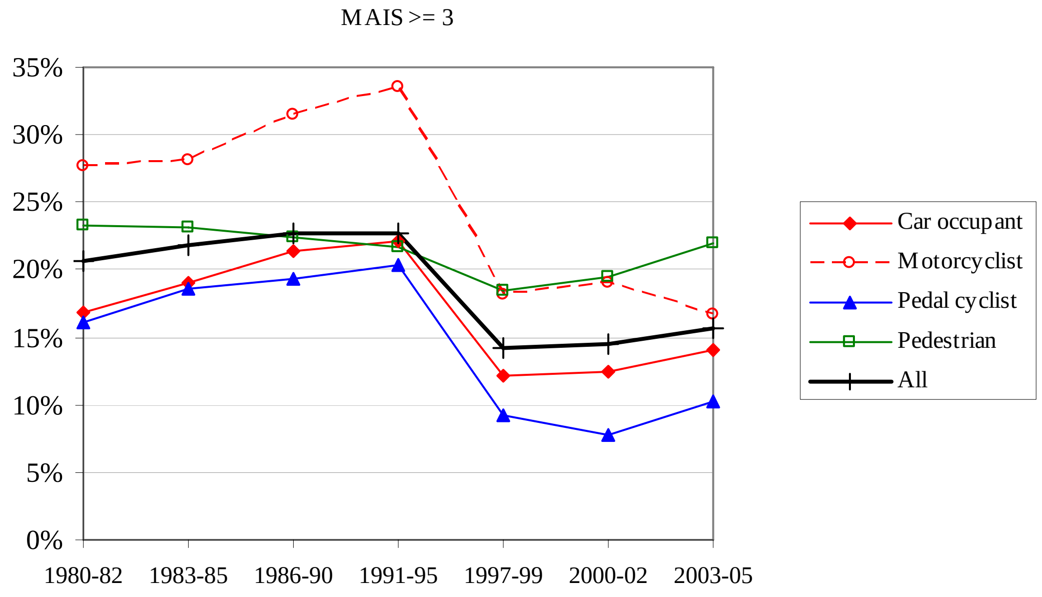


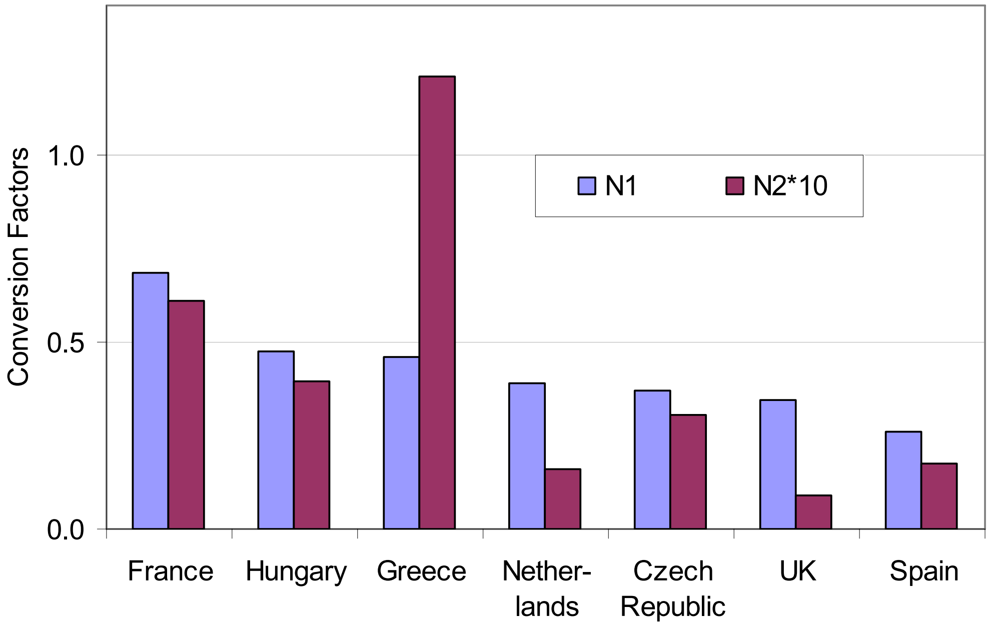
Table 30: Proportion of casualties in 1997-99 with MAIS>=3

	Actual (ICD10-based)	Estimated (ICD9-based)	Adjustment to ICD9	Conversion Factors Serious	Slight
Car occupant	12.1%	24.7%	2.03	0.30	0.008
Motorcyclist	18.1%	35.4%	1.95	0.52	0.022
Pedal cyclist	9.3%	22.1%	2.38	0.69	0.021
Pedestrian	18.5%	21.2%	1.14	0.26	0.009
Other				0.30	0.007
All	14.1%	24.5%	1.74	0.34	0.009

Applying the overall adjustment to the Czech conversion factors (Table 13) gives 0.37 for serious casualties and 0.031 for slight.

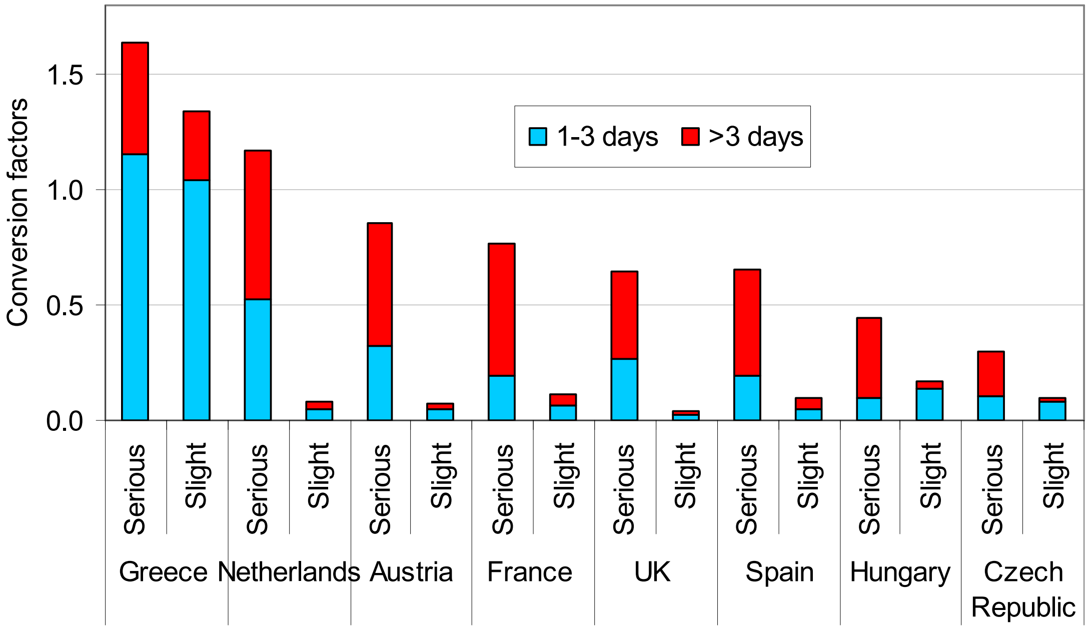
The overall results from the national studies are brought together in the following two Figures. First, Figure 8 presents the conversion factors for MAIS>=3, with the factor for slight casualties multiplied by 10 (the Czech and UK factors have been adjusted to ICD9). For example, from Table 16, for every serious casualty in the French accident data there are 0.68 casualties with MAIS>=3, while for every slight casualty there are 0.06.

Figure 8: Conversion Factors for MAIS>=3, all road users



Next, Figure 9 compares the conversion factors for Length of Stay>=1 day, distinguishing between 1-3 and >3 days. Even the ranking of countries in the two Figures is different.

Figure 9: Conversion Factors for Length of Stay>=1 day, all road users



The clear differences shown in these Figures between countries confirm that casualty reporting practices differ markedly, and that it would be misleading to compare national casualty data without adjustment. The differences demonstrate the need for the use of conversion factors such as those presented here.

4.5 An application

The conversion factors presented in Section 3 (adjusted to ICD9 basis as necessary) will now be applied using actual CARE data, to illustrate the ways in which the scope for international comparisons is expanded by the results of this research. Naturally, the strength of the results depends upon the assumption that the factors are nationally representative.

Table 31 presents the 2003-05 average annual casualty totals from CARE. At the time the data were downloaded, no post-2003 data were available for the Netherlands, only data from 2005 were available for Hungary and no data were available for the Czech Republic. The definition of serious and slight casualty in France changed in 2005, so the conversion factors only apply up to 2004.

Table 31: 2003-05 annual average casualty totals

	Killed	Serious	Slight
Austria	859	13,956	41,367
France (2002-04)	6,414	19,898	100,587
Greece	1,644	2,338	18,650
Hungary (2005)	1,278	8,320	19,185
Netherlands (2001-03)	1,003	10,881	29,608
Spain	4,861	23,323	117,286
UK	3,454	32,445	254,253

Source: CARE database, June 2007

To distinguish the number of serious casualties estimated according to the definition proposed above, i.e. with MAIS between 3 and 6, these will be referred to as serious* casualties. The conversion factors illustrated in Figure 8 are applied in Table 32 to estimate the number of serious* casualties, showing the components N1 and N2. The relation between the number of serious and serious* casualties varies widely. In Greece, the estimated number of serious* casualties considerably exceeds the number of serious casualties, while in the Netherlands, Spain and the United Kingdom it is less than one half.

Table 32: Estimation of the number of Serious* casualties

	Serious casualties			Slight casualties				
	CARE total	factor 1	N1	CARE total	factor 2	N2	Serious* =N1+N2	Serious* / Serious
France	19,898	0.68	13,612	100,587	0.061	6,157	19,768	0.99
Greece	2,338	0.46	1,081	18,650	0.121	2,259	3,339	1.43
Hungary	8,320	0.48	3,962	19,185	0.040	761	4,723	0.57
Netherlands	10,881	0.39	4,254	29,608	0.016	474	4,728	0.43
Spain	23,323	0.26	6,084	117,286	0.018	2,059	8,143	0.35
UK	32,445	0.34	11,130	254,253	0.009	2,298	13,428	0.41

In order to allow for the size of country, fatality rates per million population are often compared, also fatality rates per million motor vehicles. Table 33 compares these fatality rates with the rates of serious* casualties. IRTAD counts of population and motor vehicles in 2005 have provided the denominators for these calculations.

Table 33: Casualty rates per million population and per million motor vehicles, 2003-05

	Rate per million population		Rate per million vehicles	
	Killed	Serious*	Killed	Serious*
France (2002-04)	106	326	173	532
Greece	148	301	248	503
Hungary (2005)	127	468	379	1,402
Netherlands (2001-03)	61	290	116	548
Spain	112	187	176	294
UK	57	223	103	399

These results need to be qualified in several ways. The main qualification is that some conversion factors may not be nationally representative, principally because of the choice of study area. In addition, the data for the Netherlands come from an earlier period than for the other countries as the post-2003 Dutch fatality data were not available in CARE in June 2007. It is known that the national casualty totals have fallen since 2003 so the actual rates for the Netherlands for 2003-05 are lower than shown in the table.

Also, no account has been taken of trends in the conversion factors. In view of the results in the previous section, it might seem better to choose a period for each country that was centred on the period of the national study. This would lead to greater inconsistency among the periods being compared, and introduce a different form of bias. The approach adopted above seems preferable, but it will be important to take account explicitly of trends in the conversion factors in any future study of this general type.

5. Conclusions and Recommendations

At present, international comparisons of the level of road safety rely almost exclusively on the analysis of data relating to fatal accidents and fatalities. This is unsatisfactory, for by most criteria non-fatal accidents and casualties impose a burden on society that is at least as great. It is potentially misleading to analyse the level of road safety in a country only in terms of fatal accidents and fatalities.

The CARE database contains details of non-fatal accidents and casualties, so it is essential to allow these data to contribute to international comparisons of road safety. The problems with doing this have been well known for many years, and scientifically acceptable methods are needed to overcome these problems. The studies presented in this report represent an important step in this direction. They do not offer a complete solution since they only deal with the more seriously injured casualties, namely those that attend hospital for treatment. Nevertheless, this is the most important section of the spectrum of non-fatal casualties, and the underlying principles may be developed in future to extend the coverage. The results that have been achieved demonstrate that the problems mentioned above are not simply theoretical but are real and acute.

The original objective for this project included those casualties who were slightly injured, but the resources available have meant that no practical progress has been made in this direction. At present, police reports are the only systematic source of information about those with less serious injuries, although subject to the limitations that have been examined in this report. An independent source would be required in order to assess the completeness of these reports. Such data collection is expensive, and has not been attempted within this project.

Nevertheless, approaches exist that could be considered for future research. One possibility would be to include questions relating to road accident in the EC‘s SARTRE survey. The relatively low incidence of road accidents in much of Europe may rule this out on statistical grounds, but it would be possible to include questions in the regular national surveys carried out by many Governments, such as the General Household Survey in Great Britain. There is relevant experience in the Netherlands (AVV, 2002).

Various technical problems have been encountered and surmounted in the course of this project. It is probably too early to say that the coefficients estimated from these linkage studies allow the national casualty data from the participating countries to be compared in an unbiased fashion. Nevertheless, they provide important new information and demonstrate clearly how fully reliable and nationally representative coefficients may be prepared.

While the national studies have been carried out for a specific purpose within the SafetyNet project, the linked data sets have a research potential that extends far beyond this purpose and that has not been considered in this report.

Recommendations

1. The definition of a “serious” casualty for use with the CARE database should be a non-fatal casualty with MAIS between 3 and 6 (inclusive). At present this can only be implemented in the countries reported here (other than Austria). It is recommended that a new study be carried out so that this definition can be implemented in more countries. Naturally, this study would take into account the lessons learned in the study reported here.

2. This approach applied in this project represents an important step forward in comparing the numbers of road accidents and casualties across a range of countries. In view of the variation of the Conversion Factors among the countries studied, however, it would not be wise to generalise results to other countries. A new study is needed that includes more Member States. Also, where the current study included only part of a country, the coverage within that country should be extended where possible.

3. The CODES system of US Federal Government’s National Highway Traffic Safety Administration offers an example of how this might be achieved. The NHTSA routinely supports the state-wide linkage of accident and medical records in about 30 States. The EC could support regular national linkage studies in the Member States. The results would have many benefits in addition to the preparation of conversion factors for use with CARE.

4. Police accident reporting practices in any country will evolve over time, so the relationship between the police casualty statistics and the actual number of casualties is also likely to change. Clear evidence of this has been seen in both studies that examined developments over time (the Netherlands and Scotland). It will be necessary to repeat the linkage studies regularly in order to update the conversion factors.

5. The methods used to link police and medical records in the national studies differ in detail, as described in the reports in the Appendix. Some differences were inevitable since the levels of detail in the datasets being linked differed from country to country. Other differences, however, reflected differences in background and experience among those carrying out the linkage. Ideally, each of the various methods would have been applied to each dataset to see whether the linkages achieved depended significantly upon the method used, and if so to identify the optimal method. Unfortunately the resources available for this study did not allow for this comparative phase to be carried out, but it would undoubtedly be valuable for any future study of this type to incorporate such a comparison of the methods available among the partners.

6. The methods used to estimate MAIS from ICD injury codes need better validation. It is not necessary or desirable to confine this to samples of road accident casualties, a broad application to all injuries however caused would also be very helpful.

7. It would be valuable to estimate the number of casualties missed by both the police and the hospital reporting, in order that the conversion factors can fully account for under-reporting. The number of these casualties can be estimated by the capture-recapture method, which requires certain assumptions to be met. In particular, the fact that the police under-reporting rate is associated with characteristics such as injury severity and mode of transport must be taken into account. This can be done by stratification, in other words estimating the number of missed casualties (and hence the total number of casualties) in each strata defined by these characteristics. The conversion factors should not only be estimated according to injury severity and mode of transport (as done here) but also according to other relevant characteristics such as number of vehicles involved and type of road (e.g. urban/rural). Moreover, if we estimate conversion factors using regional data and wish to apply them nationally, the need to adjust for urban/rural characteristics is greater.

8. This study has linked police and hospital records, so inevitably can say nothing about those road accident casualties with lesser injuries that do not attend hospital for treatment. While individually these people are less severely injured than those who attend hospital, they are likely to be more numerous. It will be important to develop techniques that can prepare conversion factors for this group of casualties. The new IDB (Injury Database) of DG-Sanco will contain detailed information about traffic accidents in the near future from samples of accidents across the EU. This new source of information could potentially be used in linkage studies.

6. References

Association for the Advancement of Automotive Medicine (1990). *The Abbreviated Injury Scale (1990 revision)*, Des Plaines, Illinois.

AVV, Ministerie van Verkeer en Waterstaat, (2002). *Verkeersongevallen in Nederland 2001*, Heerlen, The Netherlands. http://www.rws-avv.nl/pls/portal30/docs/3736.PDF

Brasel K J, Lim H J, Nirula R and Weigelt J A (2007). *Length of Stay: an appropriate quality measure?* Archives of surgery, vol 142, pp 461-466.

Crash Outcome Data Evaluation System (CODES). http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/codes.html,, visited 25 June 2007

Department of Health and Human Services. Health Resources and Services Administration. *From E to VWXY Cause of Injury Codes.* ftp://ftp.hrsa.gov/mchb/vwxy.pdf

European Centre for Injury Prevention, University of Navarra (2006). *Algorithm to transform ICD-10 codes AIS and ISS, version 1 for SPSS.* Pamplona, Spain

Expert Group on Injury Severity Measurement (EGISM) (2004). *Discussion document on injury severity measurement in administrative datasets.* http://www.cdc.gov/nchs/data/injury/DicussionDocu.pdf.

Lopez D G, Rosman D L, Jelinek G A, Wilkes G J and Sprivulis P C (1999). *Complementing police road-crash records with trauma registry data – an initial evaluation*. Accident Analysis and Prevention, Vol. 32, pp. 771-777.

MacKenzie, E. J., Sacco, W et al. (1997). *ICDMAP-90: A users guide.* Baltimore, The Johns Hopkins University School of Public Health and Tri-Analytics, Inc.

National Center for Health Statistics (2007). *International Classification of Diseases 10th Revision (ICD-10).* http://www.cdc.gov/nchs/about/major/dvs/icd10des.htm.

Osler, T., Rutledge, R., et al. (1996). *ICISS: an international classification of disease-9 based injury severity score.* J. Trauma 41 (3), 380–388.

Stephenson S, Langley J and Cryer C (2005). *Effects of service delivery versus changes in incidence on trends in injury: a demonstration using hospitalised traumatic brain injury.* Accident Analysis and Prevention 2005; 37(5):825-832.

World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.* Geneva

7. Appendix A

This Appendix includes eight reports that present details of the national studies whose results appeared in the main report. Each report was prepared by the person responsible for carrying out that national study. The reports have a shared structure, but are formatted independently.

7.1 Study carried out in Austria

Report prepared by Stefan Hoeglinger (KfV)

7.1.1 Introduction

The Austrian national study combined the police database of accidents throughout Austria in the year 2001 with the Austrian hospital discharge database.

7.1.2 Description of data sources

The police database

The police database of Austria comprises data of road accidents on public roads with at least one person injured. This statistics should cover 100% of all relevant cases in Austria. The police collect the data by filling in a standardised form. All these papers are collected by the Federal Statistics body (Statistik Austria) and transformed into an electronic format. Every month, the Federal Statistics body is sending out the monthly data of road accidents. Persons committing suicide in a road traffic accident are not included in this database. If the probability is very high that the person died due to a serious health problem (e.g. heart attack) the respective record will also be removed from the database.

For the linking procedure and the preparation of matrixes 1 and 2 accident data of the year 2001 are used. In this year 57.223 persons were either injured or killed in a road accident. In contradiction to other European countries the Austrian police use the injury type “injury unknown” when the police agent on the scene does not know if the person is slightly or seriously injured. Within the CARE system these persons are generally treated as seriously injured. Like most European countries, Austria is using the 30 day definition for fatalities: If a person dies within 30 days after the accident he or she will be counted as a road accident fatality. Uninjured persons involved in the accidents are excluded from this linking process.

The hospital discharge database

In the hospital discharge database administrative and medical data of all in-patients of 270 Austrian hospitals are collected. The hospital discharge records are designed to fulfil the needs of financial compensation for medical services.

This statistics should cover 100% of all relevant cases in Austria. Out-patients are not recorded in this database. Data is collected by the hospitals and transferred on an electronically basis to the Ministry for Health and Women. The ministry passes the data to the Federal Statistics body (Statistik Austria) which prepares the so called “Spitalsentlassungsstatistik” (hospital discharge statistics).

For our needs we analysed all in-patients injured in road accidents, corresponding to ICD 10 coding (S00-T99) and the external cause of injury “not at work” or “at work”. In this database only the main diagnosis at the moment of discharging from the hospital is recorded. Before 2001, the hospitals reported the main diagnosis using the ICD 9 coding system.

The hospital discharge database is case-orientated instead of person-orientated. If a person is transferred from one hospital to another he or she is recorded twice. It is thus not possible to recognise that he or she is just a single person. As the hospitals collect data about the patient’s age when leaving the hospital, the age at the moment when admitted to the hospital has to be recalculated by using the date of birth and the date of admission to the hospital.

The hospital discharge data is published on an annual basis. Due to some late transmission of data by hospitals, there is a delay of 1½ years between the year of discharge and the year of data publication. Every year, selected indicators from this database are calculated and sent to EUROSTAT, WHO-HFA (Health for All) and also to the OECD database (OECD – Health Data).

In order to prepare data for the linking procedure, data from 2001 and 2002 were used in order to prepare a database with recorded persons who were admitted to a hospital in 2001. Only records in the hospital discharge database with an indication of a road traffic accident (“not at work” or “at work”) where used for the linking procedure. (IX14_Verletzungsg; Cause of Injury ISIS S30) Without this limitation around 250,000 records of injuries and toxications are in the database each year. When applying this limitation only 18,067 records are remaining.

7.1.3 Description of the linking process

Before the linking procedure is started, the data files have to be properly prepared. All this was done by using the standard software SPSS and MS Access.

Variables used for the linking process

Table 34 shows the key variables used in the linking process.

Table 34: Key variables in the police and hospital databases

Common name	Hospital discharge database		Police database on road accidents	
	Explanation	Variable	Explanation	Variable
Date	Date when admission to the hospital (no unknown in the database)	H_Date	Date when the accident happens (no unknown in the database)	P_Date
-	-	-	Hour of the day when the accident happens -0h-21h and 22h-24h (no unknown in the database)	P_Hour
Sex	Gender (no unknown in the database)	H_Sex	Gender (no unknown in the database)	P_Sex
Austrian or Foreigner	Nationality of person (derived from the passport not from the country of living) (no unknown in the database)	H_Foreigner	Nationality of person (derived from the passport not from the driving license or license plate; no unknown in the database)	P_Foreigner
Age	Age (no unknown in the database)	H_Age	Age (24 records with unknown in the database)	P_Age
Federal state	Federal state where the hospital is located. (no unknown in the database)	H_Fed	Federal state where the accident happened. (no unknown in the database)	P_Fed
Living federal state	Federal state where the person lives. (no unknown in the database)	H_Fed	-	-

As mentioned above, in the police database only records of injured or killed persons are documented and from the hospital discharge database only records of persons involved in a road accident are used. This pre-selection is necessary because the variables for the linking process are very limited. Otherwise the chance of having a random (and therefore more likely wrong) match between records of both databases would be relatively high.

Technical preparation

All the variables used (Table 34) are prepared to use common values and variables. For the linking process the following additional variables are added to the databases.

Table 35: Additional variables used for the linking process between the police and hospital databases

Police database	Hospital database
P_Nr	H_Nr
P_Lsi	H_Lsi
P_Pointer1	H_Pointer1
P_Dist1	H_Dist1
P_Pointer2	H_Pointer2
P_Dist2	H_Dist2
P_Matchnr	H_Matchnr
P_Selectivity	H_Selectivity
P_Distfinal	H_Distfinal

First, the databases were arranged in ascending order of the variable H_Date and P_Date. After that, the records were consecutively numbered using the variable H_Nr and P_Nr. Later on these numbers are refered to as the rank numbers of the record.

P_Lsi and H_Lsi indicate the status of the record. The values for this variable are:

-2	died on the scene
-1	undecided
0	matched
1	not matchable
2	matchable but not matched due to more than one perfect fits. “Double zeros”

The variable P_Dist1 and H_Dist1 is used to calculate the distance to the best neighbour of a record. The variable P_Dist2 and H_Dist2 contain the distance to the next best neighbour of the relevant record. The variable P_Pointer1 and H_Pointer1 refer to the rank number of the best neighbour. The variable P_Pointer2 and H_Pointer2 refer to the rank number of the next best neighbour. P_Selectivity and H_Selectivity show the calculated final selectivity between matched pairs. P_Distfinal and H_Distfinal show the final distance between matched records.

Preparation of the record set of the hospital admission database

For the analysis it is necessary to derive a database which contains records of those persons who were admitted to the hospital in the year 2001. As the 2001 hospital discharge database contains only records of persons who left the hospital in the year 2001, all records from the hospital discharge databases of the years 2001 and 2002 with an admission date in the year 2001 were selected. The result was that 233 persons were admitted to the hospital in 2001 and left the hospital in 2002. It can be assumed that there are just a few cases of people who were admitted to the hospital in 2001 and left the hospital after the year 2002. This was thus not analysed.

We presume that a very high percentage of all relevant road accident victims get admitted to a hospital within 4 days after the accident. Therefore the 136 hospital admission database records until 4 January 2002 where included as well.

The variable "Age" of the hospital discharge database contains the age of the person when leaving the hospital. Therefore the variable, H_Age (age at the moment of hospital admission) is calculated by using this variable "Age", the date of admission to the hospital and the date of birth.

The Linking procedure

The attempt to match the Austrian police's road traffic accident database with the hospital discharge (admission) database used a procedure very similar to the one proposed by SWOV in 2001.

The police database for 2001 contains 57.223 road traffic accidents, whereas the hospital database contains 18.067 hospitalisations. By simply joining every record of the police database with every record of the hospital database about 1 billion matches would have to be assessed. By using a slightly modified procedure as the one proposed by SWOV the assessment is limited to a reasonable amount. The quality of the linking process has been improved steadily by running iterative series of tests.

The similarity of the matches between records of both databases is calculated by the distance function which is described in "The distance function". The quality or uniqueness of the match is calculated by the selectivity function. This is explained in "The selectivity function". In "The linking procedure" the linking procedure itself is described.

The distance function

The personal ID-number, which could serve as a primary key, is not recorded in the Austrian police and hospital reports, so a set of other characteristics has to be used to match the respective records. These key variables (see Table 34) are used in the distance function.

When comparing the values of the key variables they are sometimes not correctly registered or even missing. Furthermore, the time span between the occurrence of the accident and the time of admission to the hospital depends on some parameters. It is uncertain that a seriously injured person is brought to a hospital with a delay of two days. But it is more probable that a slightly injured person might have some problems two days after the accident and then go to the hospital. To quantify the similarity between two records of the police and hospital databases a generalised distance function has been defined by SWOV. A very low distance close to zero indicates a very high probability that the person in the police database is the same as the one recorded in the hospital database. If the distance is higher (because the key variables are different in some variables) the probability that this matched pair refers to the same victim is smaller.

As mentioned by SWOV (2001) the distance function can be described as follows:

> “Let the hospital database contain N_1 records and the police database contain N_2 records, and let c_{ik} denote the category of record i on key variable k (k = 1,…., m), then the distance between record i (i=1, …, N_1) and record j (j = 1,…, N_2) is defined as:
>
> $$d_{ij} = \sum_{k=1}^{m} \delta(c_{ik}, c_{jk}). \qquad (1)$$
>
> Very generally, the term $\delta(c_{ik}, c_{jk})$
>
> $$\delta(c_{ik}, c_{jk}) = \begin{Bmatrix} 0 \;\; if\; c_{ik} = c_{jk} \\ a_k \;\; if\; c_{ik} \neq c_{jk} \\ b_k \;\; if\; c_{ik}\; and/or\; c_{jk} \;\; \text{missing} \end{Bmatrix} \qquad (2)$$
>
> Although the values of a_k and b_k in (2) are defined for each key variable, they all have in common that they increase the distance between two records when the records contain unequal categories and/or missing information on a key variable. “

The determination of a_k and b_k for the following key variables was ruled by the assumption that a distance of 100 corresponds to a probability of about 50% that two records refer to the same victim (see Figure 10). The determination of the distances is a weakness of the used methodology. By carrying out several trials the distances were chosen manually.

Figure 10: Determination of distances: Probability – Distance function

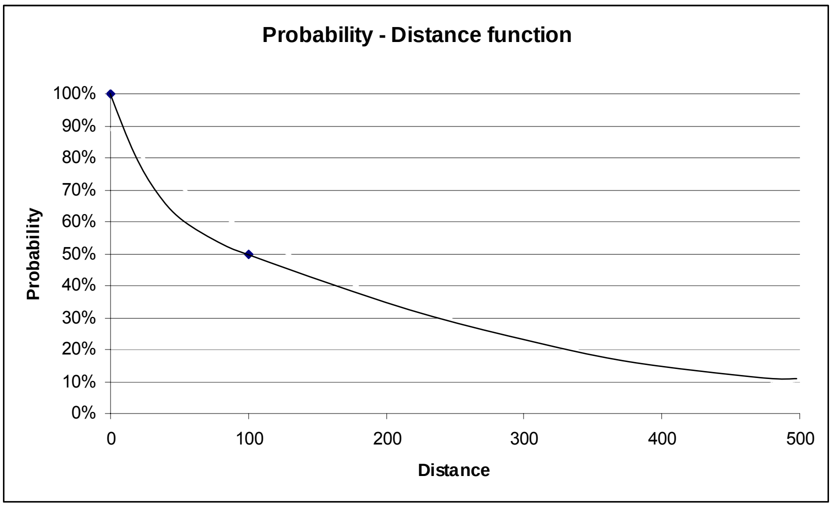


Table 36: Values for a_K for the key variable "Date"

P_Inj_Name	P_Hour_Name	Date_Diff	ak
Slightly injured	before 21	0	100
Slightly injured	after 21	0	100
Slightly injured	before 21	1	120
Slightly injured	after 21	1	110
Slightly injured	before 21	2	150
Slightly injured	after 21	2	150
Slightly injured	before 21	3	200
Slightly injured	after 21	3	200
Slightly injured	before 21	4	300
Slightly injured	after 21	4	300
Injury unknown	before 21	0	0
Injury unknown	after 21	0	0
Injury unknown	before 21	1	30
Injury unknown	after 21	1	10
Injury unknown	before 21	2	50
Injury unknown	after 21	2	50
Injury unknown	before 21	3	100
Injury unknown	after 21	3	100
Injury unknown	before 21	4	200
Injury unknown	after 21	4	200
Seriously injured	before 21	0	0
Seriously injured	after 21	0	0
Seriously injured	before 21	1	50
Seriously injured	after 21	1	20
Seriously injured	before 21	2	1000
Seriously injured	after 21	2	1000
Seriously injured	before 21	3	1000
Seriously injured	after 21	3	1000
Seriously injured	before 21	4	1000
Seriously injured	after 21	4	1000
Killed within 24 hours	before 21	0	0
Killed within 24 hours	after 21	0	0
Killed within 24 hours	before 21	1	50
Killed within 24 hours	after 21	1	20
Killed within 24 hours	before 21	2	1000
Killed within 24 hours	after 21	2	1000
Killed within 24 hours	before 21	3	1000
Killed within 24 hours	after 21	3	1000
Killed within 24 hours	before 21	4	1000
Killed within 24 hours	after 21	4	1000
Killed between 24 and 48 hours	before 21	0	0
Killed between 24 and 48 hours	after 21	0	0
Killed between 24 and 48 hours	before 21	1	50
Killed between 24 and 48 hours	after 21	1	20
Killed between 24 and 48 hours	before 21	2	1000
Killed between 24 and 48 hours	after 21	2	1000
Killed between 24 and 48 hours	before 21	3	1000
Killed between 24 and 48 hours	after 21	3	1000
Killed between 24 and 48 hours	before 21	4	1000
Killed between 24 and 48 hours	after 21	4	1000
Killed between 48 and 72 hours	before 21	0	0
Killed between 48 and 72 hours	after 21	0	0
Killed between 48 and 72 hours	before 21	1	50
Killed between 48 and 72 hours	after 21	1	20
Killed between 48 and 72 hours	before 21	2	1000
Killed between 48 and 72 hours	after 21	2	1000
Killed between 48 and 72 hours	before 21	3	1000
Killed between 48 and 72 hours	after 21	3	1000
Killed between 48 and 72 hours	before 21	4	1000
Killed between 48 and 72 hours	after 21	4	1000
Killed between 72 and 30 days	before 21	0	0
Killed between 72 and 30 days	after 21	0	0
Killed between 72 and 30 days	before 21	1	50
Killed between 72 and 30 days	after 21	1	20
Killed between 72 and 30 days	before 21	2	1000
Killed between 72 and 30 days	after 21	2	1000
Killed between 72 and 30 days	before 21	3	1000
Killed between 72 and 30 days	after 21	3	1000
Killed between 72 and 30 days	before 21	4	1000
Killed between 72 and 30 days	after 21	4	1000

In Table 36, P_Inj_Name is the severity of injury reported in the police database. P_Hour_Name indicates if the accident happens between 0h and 21h or between 22h and 24h. Date_diff is the number of days between the date when the accident happened and the date where the patient was admitted to the hospital.

b_k: n/a as no date is missing in one of the databases

Table 37: Values for a_k for the key variable "Sex"

P_Sex_Name	H_Sex_Name	ak
male	male	0
male	female	2000
female	male	2000
female	female	0

Where P_Sex_Name is the gender information in the police database and H_Sex_Name is the gender information in the hospital file.

b_k: n/a as no gender information is missing in one of the databases

Key variable: Austrian or Foreigner

a_k: $a_k = 0$ if the value of this variable is the same in both databases

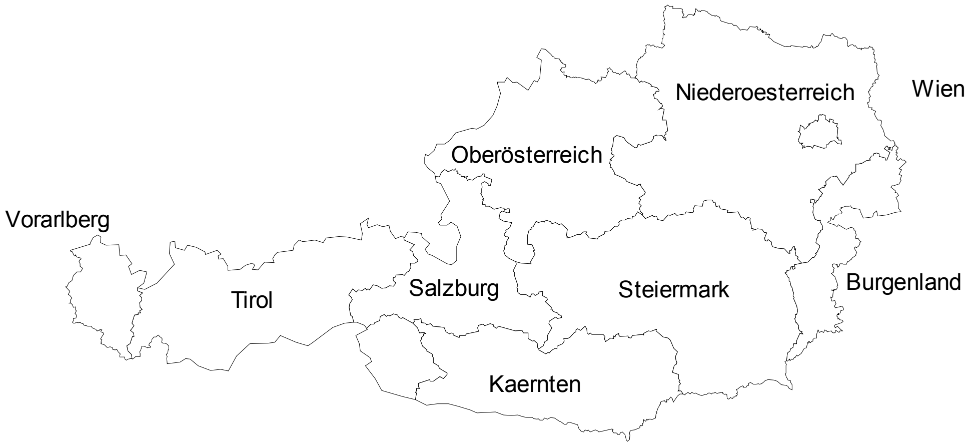
b_k: n/a as no information is missing in one of the databases

If the value of this variable differs in the databases, the match is treated as not linkable at all.

Key variables about the location in Austria

The following pages describe key variables related to federal states in Austria. Figure 11 shows a map of Austria including all federal states to get an impression about the chosen a_k.

Figure 11: Map of Austria



Key variable: Federal state

a_k:

Table 38: Values for a_k for the key variable "Federal state"

P_Fed_Name	H_Fed_Name	ak
Burgenland	Burgenland	0
Burgenland	Kaernten	250
Burgenland	Niederoesterreich	125
Burgenland	Oberoesterreich	250
Burgenland	Salzburg	500
Burgenland	Steiermark	125
Burgenland	Tirol	35
Burgenland	Vorarlberg	500
Burgenland	Wien	35
Kaernten	Burgenland	500
Kaernten	Kaernten	0
Kaernten	Niederoesterreich	250
Kaernten	Oberoesterreich	350
Kaernten	Salzburg	125
Kaernten	Steiermark	125
Kaernten	Tirol	35
Kaernten	Vorarlberg	500
Kaernten	Wien	35
Niederoesterreich	Burgenland	125
Niederoesterreich	Kaernten	500
Niederoesterreich	Niederoesterreich	0
Niederoesterreich	Oberoesterreich	250
Niederoesterreich	Salzburg	500
Niederoesterreich	Steiermark	125
Niederoesterreich	Tirol	35
Niederoesterreich	Vorarlberg	500
Niederoesterreich	Wien	35
Oberoesterreich	Burgenland	500
Oberoesterreich	Kaernten	500
Oberoesterreich	Niederoesterreich	75
Oberoesterreich	Oberoesterreich	0
Oberoesterreich	Salzburg	250
Oberoesterreich	Steiermark	125
Oberoesterreich	Tirol	35
Oberoesterreich	Vorarlberg	500
Oberoesterreich	Wien	35
Salzburg	Burgenland	500
Salzburg	Kaernten	125
Salzburg	Niederoesterreich	350
Salzburg	Oberoesterreich	75
Salzburg	Salzburg	0
Salzburg	Steiermark	250
Salzburg	Tirol	35
Salzburg	Vorarlberg	500
Salzburg	Wien	35
Steiermark	Burgenland	50
Steiermark	Kaernten	125
Steiermark	Niederoesterreich	125
Steiermark	Oberoesterreich	125
Steiermark	Salzburg	125
Steiermark	Steiermark	0
Steiermark	Tirol	35
Steiermark	Vorarlberg	500
Steiermark	Wien	35
Tirol	Burgenland	500
Tirol	Kaernten	125
Tirol	Niederoesterreich	500
Tirol	Oberoesterreich	500
Tirol	Salzburg	75
Tirol	Steiermark	500
Tirol	Tirol	0
Tirol	Vorarlberg	250
Tirol	Wien	35
Vorarlberg	Burgenland	500
Vorarlberg	Kaernten	500
Vorarlberg	Niederoesterreich	500
Vorarlberg	Oberoesterreich	500
Vorarlberg	Salzburg	500
Vorarlberg	Steiermark	500
Vorarlberg	Tirol	35
Vorarlberg	Vorarlberg	0
Vorarlberg	Wien	35
Wien	Burgenland	500
Wien	Kaernten	500
Wien	Niederoesterreich	500
Wien	Oberoesterreich	500
Wien	Salzburg	500
Wien	Steiermark	500
Wien	Tirol	35
Wien	Vorarlberg	500
Wien	Wien	0

P_Fed_Name indicates the federal state where the accident happened (mentioned in the police database) and H_Fed_Name indicates in which federal state the hospital is located.

b_k: n/a as no information is missing in one of the databases

Key variable: Federal state of living

a_k:

Table 39: Values for a_k for the key variable “Living”

P_Fed_Name	H_Liv_Name	ak
Burgenland	Burgenland	0
Kaernten	Burgenland	150
Niederoesterreich	Burgenland	70
Oberoesterreich	Burgenland	150
Salzburg	Burgenland	150
Steiermark	Burgenland	70
Tirol	Burgenland	150
Vorarlberg	Burgenland	150
Wien	Burgenland	40
Burgenland	Kaernten	150
Kaernten	Kaernten	0
Niederoesterreich	Kaernten	150
Oberoesterreich	Kaernten	150
Salzburg	Kaernten	70
Steiermark	Kaernten	70
Tirol	Kaernten	70
Vorarlberg	Kaernten	150
Wien	Kaernten	150
Burgenland	Niederoesterreich	70
Kaernten	Niederoesterreich	70
Niederoesterreich	Niederoesterreich	0
Oberoesterreich	Niederoesterreich	70
Salzburg	Niederoesterreich	150
Steiermark	Niederoesterreich	70
Tirol	Niederoesterreich	150
Vorarlberg	Niederoesterreich	150
Wien	Niederoesterreich	40
Burgenland	Oberoesterreich	150
Kaernten	Oberoesterreich	150
Niederoesterreich	Oberoesterreich	70
Oberoesterreich	Oberoesterreich	0
Salzburg	Oberoesterreich	70
Steiermark	Oberoesterreich	70
Tirol	Oberoesterreich	150
Vorarlberg	Oberoesterreich	150
Wien	Oberoesterreich	150
Burgenland	Salzburg	150
Kaernten	Salzburg	70
Niederoesterreich	Salzburg	150
Oberoesterreich	Salzburg	70
Salzburg	Salzburg	0
Steiermark	Salzburg	70
Tirol	Salzburg	70
Vorarlberg	Salzburg	150
Wien	Salzburg	150
Burgenland	Steiermark	70
Kaernten	Steiermark	70
Niederoesterreich	Steiermark	70
Oberoesterreich	Steiermark	70
Salzburg	Steiermark	150
Steiermark	Steiermark	0
Tirol	Steiermark	150
Vorarlberg	Steiermark	150
Wien	Steiermark	150
Burgenland	Tirol	150
Kaernten	Tirol	70
Niederoesterreich	Tirol	150
Oberoesterreich	Tirol	150
Salzburg	Tirol	70
Steiermark	Tirol	150
Tirol	Tirol	0
Vorarlberg	Tirol	70
Wien	Tirol	150
Burgenland	Vorarlberg	150
Kaernten	Vorarlberg	150
Niederoesterreich	Vorarlberg	150
Oberoesterreich	Vorarlberg	150
Salzburg	Vorarlberg	150
Steiermark	Vorarlberg	150
Tirol	Vorarlberg	70
Vorarlberg	Vorarlberg	0
Wien	Vorarlberg	150
Burgenland	Wien	40
Kaernten	Wien	70
Niederoesterreich	Wien	40
Oberoesterreich	Wien	70
Salzburg	Wien	70
Steiermark	Wien	70
Tirol	Wien	150
Vorarlberg	Wien	150
All federal states	Not Austria	0
All federal states	Unknown	0
All federal states	Austria but unknown	0

P_Fed_Name indicates the federal state where the accident happened (mentioned in the police database) and H_Liv_Name indicates the federal state where the hospitalised person lives.

b_k: n/a as no information is missing in one of the databases

Project co-financed by the European Commission, Directorate-General Transport and Energy

Key variable: Age

a_k: $a_k = 0$ if the value of this variable is the same in both databases

b_k: n/a as no information is missing in one of the databases

If the value of this variable differs in the databases, the match is treated as not linkable at all.

The selectivity function

The similarity of the records can be quantified by calculating the distance between two records of the two databases. If a record of the police database finds a very similar record (small distance) in the hospital database it is important to control if there are also other hospital records existing with a small distance to the initial record in the police database. In such a case, the uniqueness of the initial pair can be criticised.

The selectivity of a matched pair is the minimum of the differences of the distances to its best neighbour and its next best neighbour of each record. Whereas the best neighbour is the record in the other database with the lowest distance and the next best neighbour is the record in the other database with the second lowest distance to the selected current record.

If the selectivity is high, the uniqueness is high too. If the selectivity is low, it is debatable whether the best neighbour pair or the next best neighbour refers to the same person.

The linking procedure

The linking procedure is implemented by using the standard software MS Access 2003. In the following section the method of how to find matches between the hospital and the police database is described. The matches should have a high probability of referring to the same person in each database. The section “first pass” is more or less similar to the paper of SWOV (2001). In the section “second pass” an additional function is implemented to cover all possible cases of such a linking procedure.

First initiation

The first initiation starts after both databases were properly prepared.

In this initiation the following variables are filled with constant values:

P_Lsi = -2 or -1	H_Lsi = -1
P_Dist1 = 100000	H_Dist1 = 100000
P_Dist2 = 100000	H_Dist2 = 100000
P_Pointer1 = -1	H_Pointer1 = -1

P_Pointer2 = -1 H_Pointer2 = -1

If the severity of injury reported in the police database indicates that the person died on the scene the variable concerning the status of the record P_Lsi is set to -2 (died on the scene). In all other cases the value of P_Lsi and H_Lsi is set to -1 (undecided)

At the beginning of the linking procedure very high distances are applied to the relevant variables. During the linking procedure these variables are recalculated and in case of a lower distance the variables are replaced with new values.

An initial value of “-1” is selected for all pointers.

First pass

In the first pass the best neighbour and the next best neighbour is determined by using the distance function.

If a person died on the scene (P_Lsi = -2) the related record will be excluded from further examination as it can be assumed that this person will not be transferred to a hospital.

As mentioned above, only records where the admittance to the hospital occurs within 4 days after the accident were considered. The first record of the police database is compared with all records in the hospital database satisfying this time span restriction. When two records are compared it is checked, if any of the already stored distances in the variable .._Dist1 is higher as the new calculated one. If so, the old distance is updated by the new distance and stored in both records. If the new calculated is higher than the distance stored in the variable .._Dist1, but lower than the distance stored in the variable .._Dist2, the variable .._Dist2 is updated with the new calculated distance and stored in both records of the databases.

Also the corresponding pointers H_Pointer1, H_Pointer2, P_Pointer1, P_Pointer2 are updated with the rank number of the relevant records of the other database. This procedure is done for both records which are compared.

After that, the second record of the police database is compared to all records in the hospital database also satisfying the mentioned timespan restriction. The same systematic treatment is applied. This procedure continues until the last record of the police database is processed.

At the end of this first pass all records which found a similar record in the other database contain a pointer which refers to the rank number of its best neighbour in the other database and also a pointer which refers to the rank number of its next best neighbour in the other database. Each of these neighbours is combined with the distance between these records in both databases.

Second initiation

In this initiation the following variables are filled with these values:

P_Matchnr = 0	H_Matchnr = 0
P_Selectivity = 0	H_Selectivity = 0
P_Distfinal = 100000	H_Distfinal = 100000
P_Lsi = 1 or -1	H_Lsi = 1 or -1

P_Matchnr and H_Matchnr indicate the rank number of the record in the other database if the pair is defined as a match. Zero is choosen as initial value.

P_Selectivity and H_Selectivity indicate the selectivity between the best neighbour and the next best neighbour of a record. As initial value zero is chosen.

P_Distfinal and H_Distfinal store the final distance of the matched pair. (It is not necessarily the distance beween the best neighbours). A very high distance is used as an initial value.

If the pointer to the best neighbour indicates that no best neighbour was found in the first pass (P_Pointer1 = -1 or H_Pointer1 = -1) these records will be excluded from further analysis. (P_Lsi = 1; H_Lsi = 1); If at least a best neighbour was found the status for these records will remain as “undecided” (P_Lsi = -1; H_Lsi = -1)

Determination of “Double zero” matches

If a record has a distance of zero to its best neighbour and to its next best neighbour, it is uncertain which record has to be taken for the match. Both neighbours fit perfect.

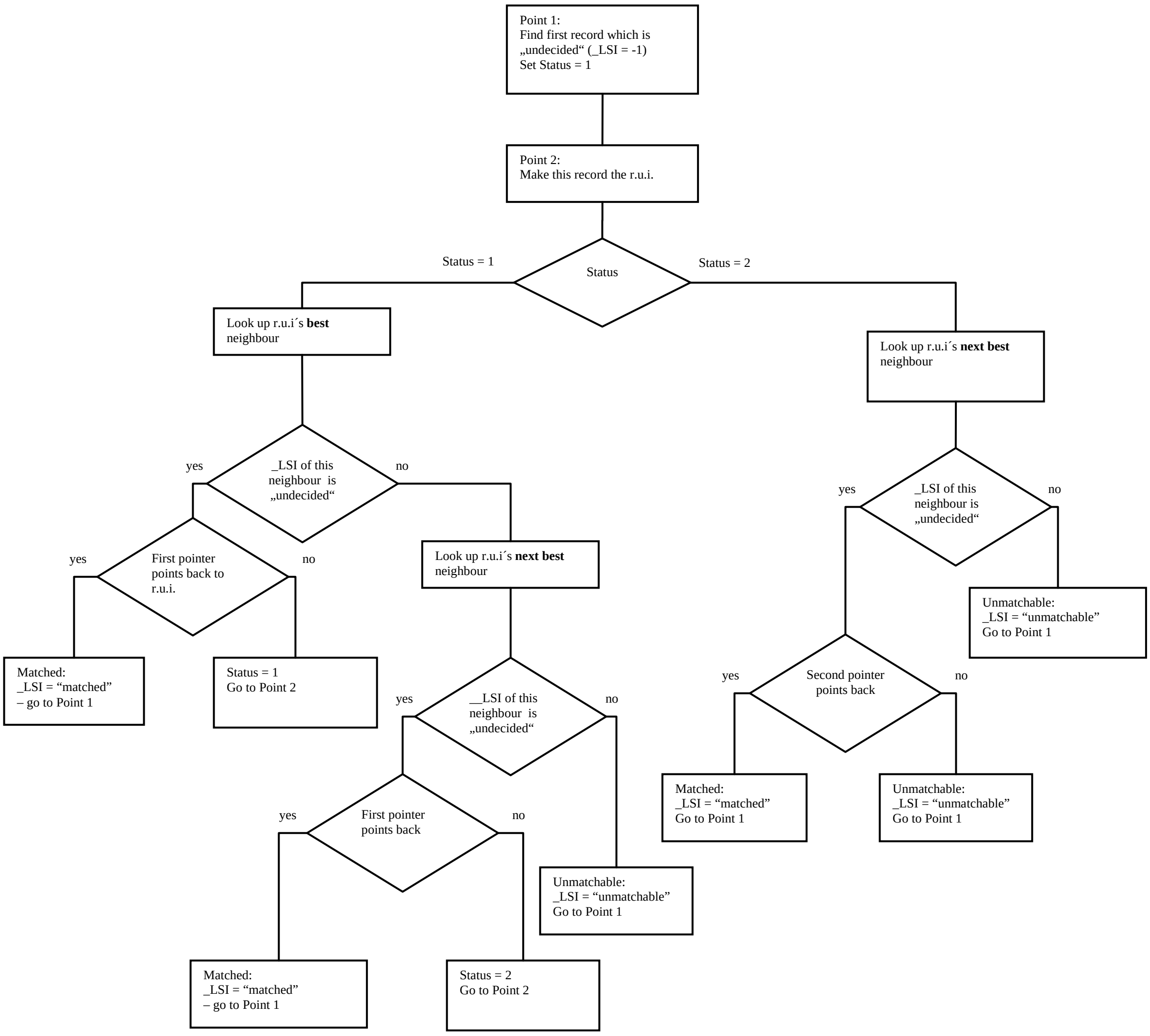
For this reason every record with such a double zero situation has to be marked as “double zeros” (.._LSI = 2). If the best neighbour of a record refers to a double zero match, this record also has to be marked with ..LSI = 2.
This record points to a record in the other database pretending that this is a perfect fit, but the other record is highly uncertain. To put it dramatically, a double zero record is poisoning every record surrounding it.

Records with status.._LSI = 2 have to be excluded from further processing as it is already known at this stage of the linking process that the situation will not change when running the second pass. Double zero records are a problem when linking the Austrian police database and the hospital database. The problem occurs due to lack of comparable information in both databases.

Second pass

In the second pass the determination of the records which should be linked is performed. This recursive procedure is almost the same as described in the paper of SWOV but enhanced with one function to find also matches between two records mentioned as next best neighbours in both databases. Figure 12 shows a flow chart which describes this recursive algorithm.

Figure 12: Description of the procedure used for the second pass


Point 1:
Find first record which is „undecided“ (_LSI = -1)
Set Status = 1
Point 2:
Make this record the r.u.i.
Status = 1
Status
Status = 2
Look up r.u.i´s **best** neighbour
Look up r.u.i´s **next best** neighbour
yes
_LSI of this neighbour is „undecided“
no
yes
_LSI of this neighbour is „undecided“
no
yes
First pointer points back to r.u.i.
no
Look up r.u.i´s **next best** neighbour
Unmatchable:
_LSI = “unmatchable”
Go to Point 1
Matched:
_LSI = “matched”
– go to Point 1
Status = 1
Go to Point 2
yes
__LSI of this neighbour is „undecided“
no
yes
Second pointer points back
no
yes
First pointer points back
no
Matched:
_LSI = “matched”
Go to Point 1
Unmatchable:
_LSI = “unmatchable”
Go to Point 1
Unmatchable:
_LSI = “unmatchable”
Go to Point 1
Matched:
_LSI = “matched”
– go to Point 1
Status = 2
Go to Point 2

Description of names used:

r.u.i	Record under investigation
Status	status to detect if a pair who is pointing at each other as next best neighbour could be matched.
First pointer	H_Pointer1 or P_Pointer1
Second pointer	H_Pointer2 or P_Pointer2

This procedure helps to find the following matches of records:

1. if the record in the first database points to its best neighbour in the other database and this record also points back to the record in the first database as its best neighbour.

2. if the record in the first database points to its best neighbour in the other database and this record also points back to the record in the first database as its next best neighbour.

3. if the record in the first database points to its next best neighbour in the other database and this record also points back to the record in the first database as its best neighbour.

4. if the record in the first database points to its next best neighbour in the other database and this record also points back to the record in the first database as its next best neighbour.

Calculation of selectivity and final distance

In this part of the linking procedure, the uniqueness of a match quantified by selectivity and final distance is calculated. First the difference between the distances of a record to its best neighbour and to its next best neighbour is calculated. This is done for records of both databases. The lowest difference of a matched pair is then stored in both records of the matched pair as their selectivity. To determine final distance the distance to the neighbour which is taken as matched neighbour is stored in each database.

Taking into account selectivity and final distance, matched pairs can be divided in those which have a high probability to refer to the same person (Matchtype: “usable matched”) and those where it is doubtful if the match is ok (Matchtype: “not usable matched (Dist, Selectivity)”). A matched pair is “usable matched if the following conditions are fulfilled:

- Final distance <= 100 and
- Selectivity >= 100

These values have been chosen carefully after processing several tests.

7.1.4 Results

As no set of records with proofed quality of linking is available, the linking procedure could not be evaluated perfectly. To see if the linking procedure is calculating reasonable results, two alternative ways of checking the plausibility of the matched records is chosen.

Reliability test I – Differences in affected body regions

For this test we assumed that injuries of motorcyclists including moped riders and car occupants differ significantly when looking at the affected body region.

Table 40 and Figure 13 show the mode of transport by affected body region for matched pairs with a matchtype “usable matched”. The standardised residuals show for example that knee and lower leg injuries of motorcyclists and moped riders differ significantly from the average due to the standardised residual of 12,1. (standardised residuals +/- 1,94 would lead to a significance level of 5%). In all body regions rows the chi-square assumption is fulfilled. (Predicted value > 5)

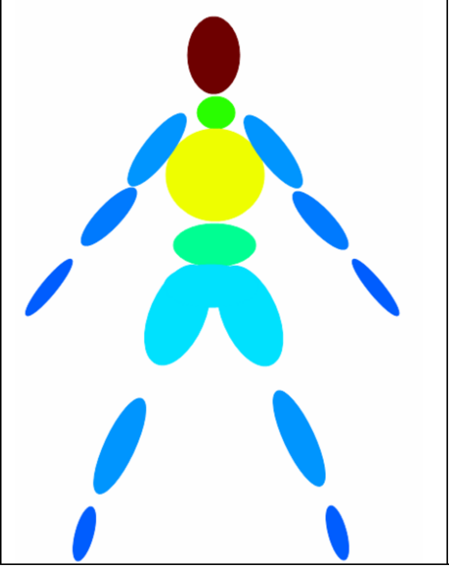
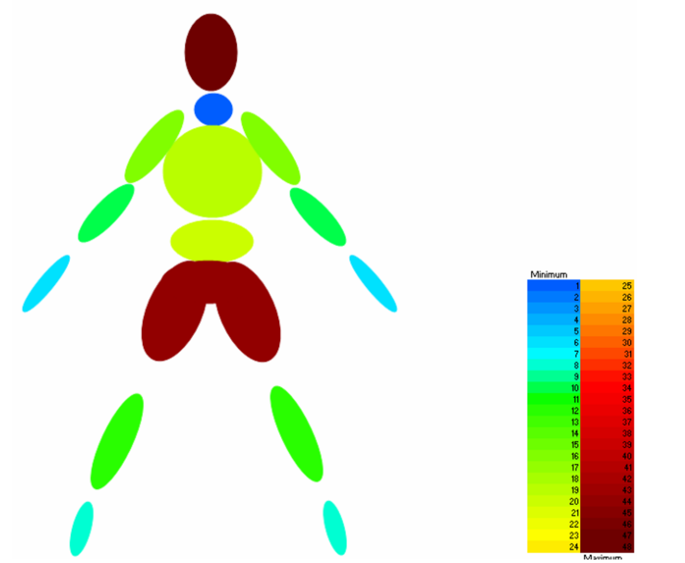
This leads to the assumption that the linking procedure does not match just records randomly. If so, there would be no significant difference by mode of transport.

Table 40: Affected body regions of car passengers and motorcyclists (only usable matched persons: 3.624 records)

Bodyregion * road user type Crosstabulation

			road user type		
			Car occupant	Motorcyclist incl. Moped rider	Total
Bodyregion	Abdomen injuries	Count	239	91	330
		Std. Residual	-,9	1,5	
	Ankel injuries	Count	48	40	88
		Std. Residual	-2,4	4,3	
	Elbow injuries	Count	69	48	117
		Std. Residual	-2,2	3,9	
	Hand injuries	Count	51	32	83
		Std. Residual	-1,6	2,8	
	Head injuries	Count	1167	210	1377
		Std. Residual	3,5	-6,3	
	Hip and thigh injuries	Count	100	59	159
		Std. Residual	-2,0	3,6	
	Knee and lower leg injuries	Count	164	196	360
		Std. Residual	-6,7	12,1	
	Neck injuries	Count	299	12	311
		Std. Residual	3,9	-7,1	
	Shoulder injuries	Count	94	74	168
		Std. Residual	-3,1	5,5	
	Thorax injuries	Count	543	88	631
		Std. Residual	2,7	-4,9	
Total		Count	2774	850	3624

Figure 13: Affected body regions of car passengers and motorcyclists incl. Moped riders (only usable matched persons: 3.624 records)

Car occupants	Motorcyclists incl. Moped riders
	

Reliability test II – Linking different years

For this test hospital data from the year 2001 were linked to police data from the year 2003. In the “date” - variable the year was changed from 2003 to 2001. Theoretically there should be no “useable matched” records after running the linking procedure.

Table 41 shows that 1.567 records were found to be useable. This is 30,5% of the “useable matched” records when linking both databases with data from the year 2001. Therefore it can be assumed that 30,5% of these matched pairs do not refer to the same person in each database. This high share reveals that not enough information is available in both databases for getting reliable results.

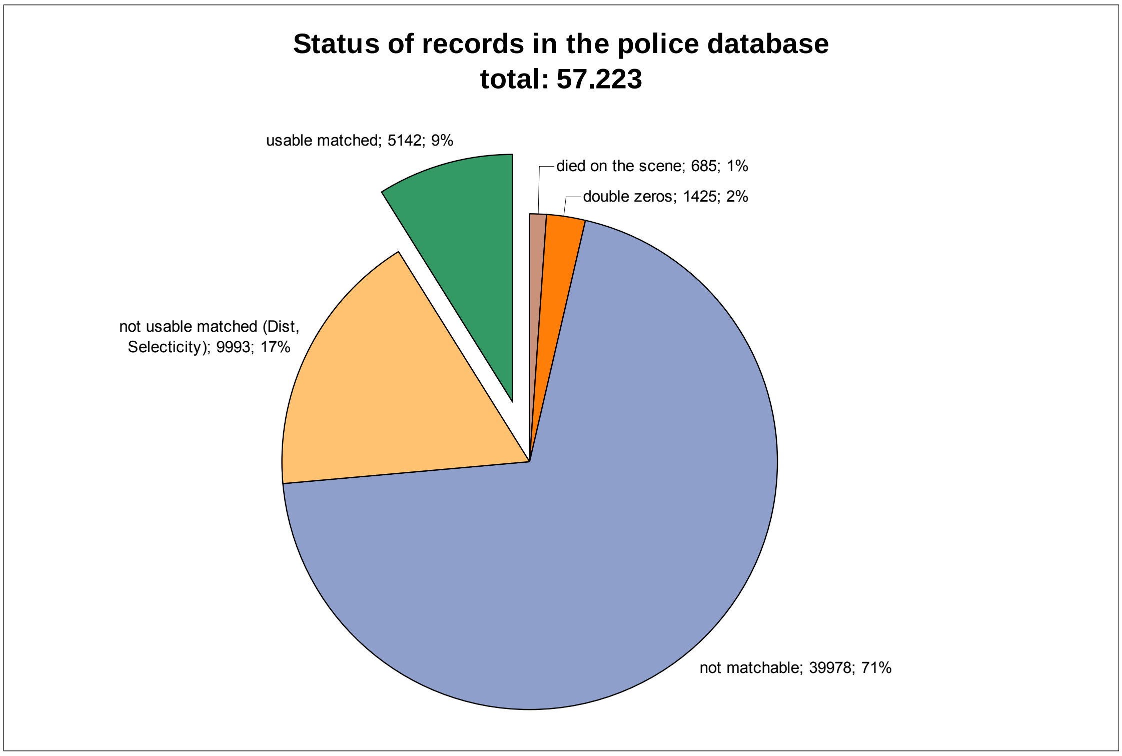
Table 41: Overview on linking procedure when linking police data from 2003 with hospital data from 2001

Source	Matchtype	Ergebnis
Hospital	double zeros	177
	not matchable	1258
	undecided	1789
Hospital Ergebnis		3224
Police	Died on the scene	614
	double zeros	290
	not matchable	42065
Police Ergebnis		42969
Police + Hospital	not usable matched (Dist, Selecticity)	13276
	usable matched	1567
Police + Hospital Ergebnis		14843
Gesamtergebnis		61036

Overview and discussion

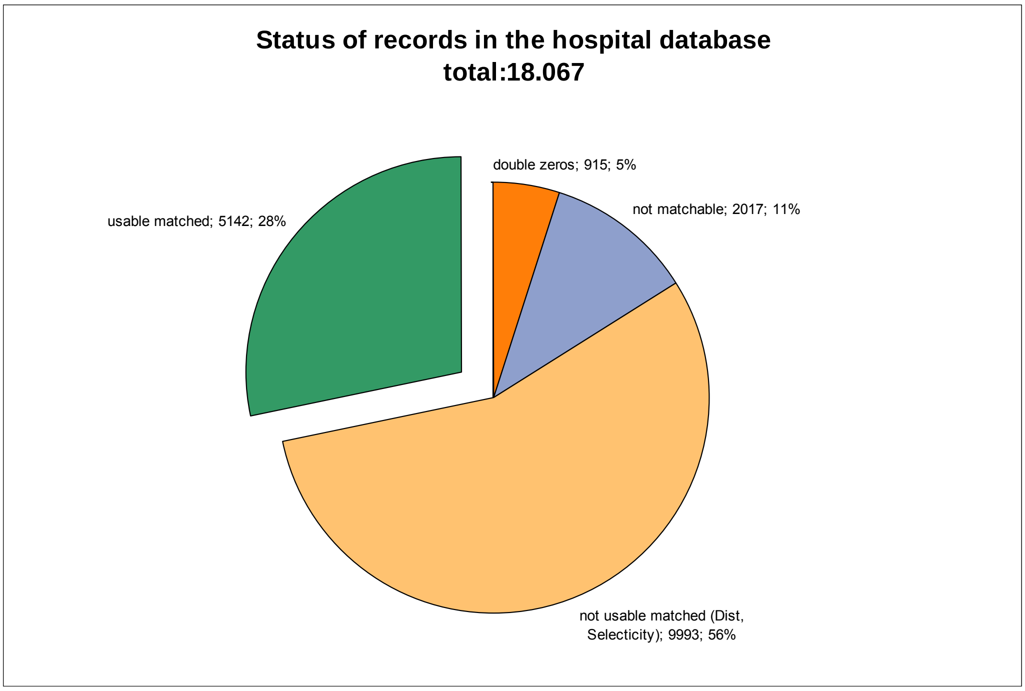
Figure 14 and Figure 15 show the status of all records in the hospital and police database after the linking process.

Figure 14: Status of records in the police database



Project co-financed by the European Commission, Directorate-General Transport and Energy

Figure 15: Status of records in the police database



In the police database 71% of all records cannot be matched. As the hospital database contains only in-patients and no outpatients, this amount can be explained by the police coded “slightly injured” persons, who – in case they were coded properly – do not go to the hospital (73% of all records).

It seems that we do have to less comparable information in both databases. This is indicated by the high percentage at 5% of all hospital records were the status is “double zero”. These records have exactly the same key variables in both databases. E.g. in Austria only the age, but not the date of birth is recorded in the police database. The linking procedure would be much more efficient if the police database would also contain this information.

When looking at the total numbers, as presented in Figure 14 and Figure 15, it could be assumed that 915 records of double zeros from the hospital database refer to the same persons as 915 double zeros from the police database but, who fits with whom can not be determined. The remaining double zeros from the police database (1.425-915 = 510) can not treated as matchable.

Preparation of the matrixes

In order to prepare the matrixes a major problem is the lack of information about the mode of transport in the hospital database. E.g., it is known how many records in the police database refer to persons in a car. But there is no information on how many records in the hospital database refer to an accident where the person was sitting in a car. Therefore it is not possible to calculate how many people in a car were not coded by the police. Additionally, it is impossible to calculate the grey marked cells in the following table on the level of the mode of transport.

Table 42: Possible combinations of presence or absence of hospitalised road traffic victims in police and hospital databases. (SWOV 2001)

	In hospital database	*Not in hospital database*	*Not a hospitalised road traffic victim*
In police database	**In both databases**	**Only in police database**	**Road traffic victim but not hospitalised**
Not in police database	**Only in hospital database**	**In neither database**	
Not a hospitalised road traffic victim	**Hospitalisation not caused by a road traffic accident**		

In order to provide a full picture of the linking procedure every record of both databases is counted in the matrixes. If two records are matched “usable matched” or “not usable matched” all needed cells of the matrixes can be calculated. In all other cases there is a lack of information and cells, which could not be calculated, are marked with “n/a”

Additional information to the variables proposed by TRL concerning the design of matrix 1 and matrix 2 is provided. These are:

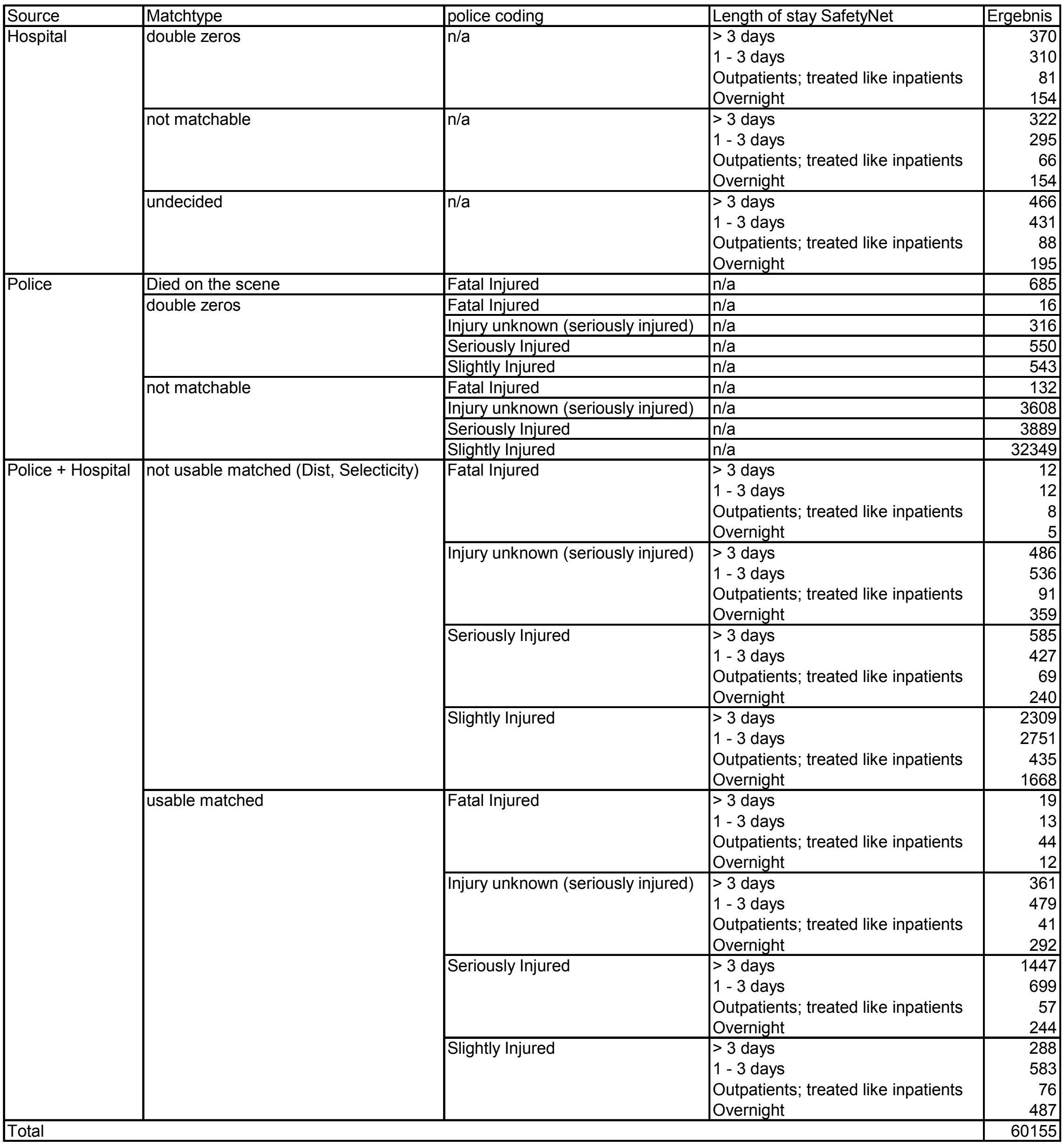
- Variable Source: This indicates wether this is a matched pair (Police + Hospital) or a record could not find another record (Police; Hospital)
- Variable Matchtype:
  - Died on the scene
  - Double zeros
  - Not matchable
  - Undecided
  - Not usable matched (Dist, Selectivity)
  - Usable matched

This information is needed to analyse why data are not matched properly.

Matrix 1 for the matched pairs

Table 43 shows an overview on matrix1 where police coding, length of stay in hospital and road user type can be crossed.

Table 43: Matrix 1 overview



Source	Matchtype	police coding	Length of stay SafetyNet	Ergebnis
Hospital	double zeros	n/a	> 3 days	370
			1 - 3 days	310
			Outpatients; treated like inpatients	81
			Overnight	154
	not matchable	n/a	> 3 days	322
			1 - 3 days	295
			Outpatients; treated like inpatients	66
			Overnight	154
	undecided	n/a	> 3 days	466
			1 - 3 days	431
			Outpatients; treated like inpatients	88
			Overnight	195
Police	Died on the scene	Fatal Injured	n/a	685
	double zeros	Fatal Injured	n/a	16
		Injury unknown (seriously injured)	n/a	316
		Seriously Injured	n/a	550
		Slightly Injured	n/a	543
	not matchable	Fatal Injured	n/a	132
		Injury unknown (seriously injured)	n/a	3608
		Seriously Injured	n/a	3889
		Slightly Injured	n/a	32349
Police + Hospital	not usable matched (Dist, Selecticity)	Fatal Injured	> 3 days	12
			1 - 3 days	12
			Outpatients; treated like inpatients	8
			Overnight	5
		Injury unknown (seriously injured)	> 3 days	486
			1 - 3 days	536
			Outpatients; treated like inpatients	91
			Overnight	359
		Seriously Injured	> 3 days	585
			1 - 3 days	427
			Outpatients; treated like inpatients	69
			Overnight	240
		Slightly Injured	> 3 days	2309
			1 - 3 days	2751
			Outpatients; treated like inpatients	435
			Overnight	1668
	usable matched	Fatal Injured	> 3 days	19
			1 - 3 days	13
			Outpatients; treated like inpatients	44
			Overnight	12
		Injury unknown (seriously injured)	> 3 days	361
			1 - 3 days	479
			Outpatients; treated like inpatients	41
			Overnight	292
		Seriously Injured	> 3 days	1447
			1 - 3 days	699
			Outpatients; treated like inpatients	57
			Overnight	244
		Slightly Injured	> 3 days	288
			1 - 3 days	583
			Outpatients; treated like inpatients	76
			Overnight	487
Total				60155

If a person is released from the hospital within the same day, one day is stored in the variable “XDauer” of the Austrian hospital database. If a person is overnight in the hospital two days are stored in this variable. The conversion from “Xdauer” to the definition of “Length of stay SafetyNet” used in this report is as follows:

Table 44: Transformation of "Xdauer" to "Length of stay SafetyNet" used in this report

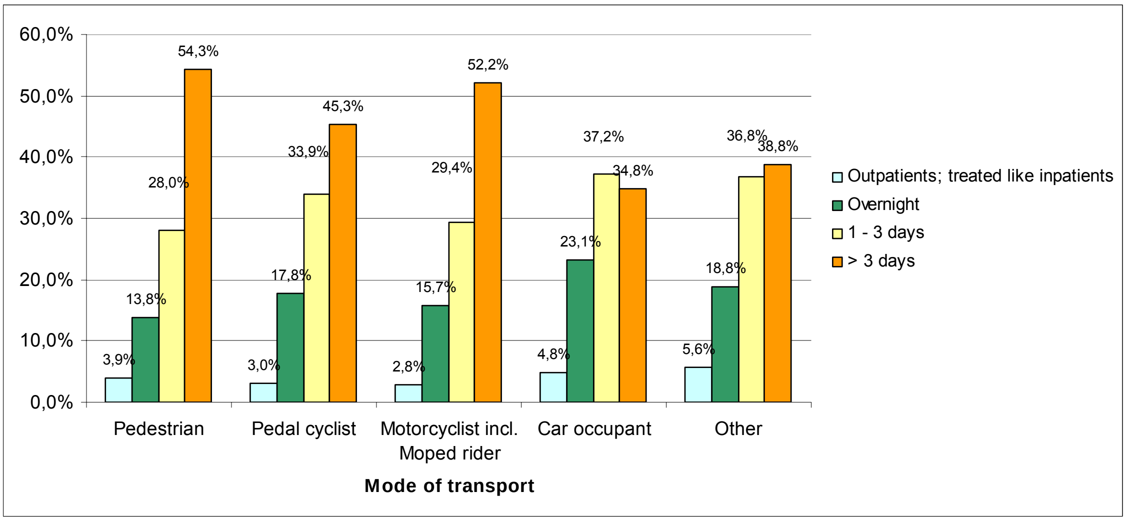
Xdauer	Length of stay SafetyNet
1	Outpatients; treated like inpatients
2	Overnight
3	1 - 3 days
4	1 - 3 days
5	1 - 3 days
6 and more	> 3 days

1.434 persons are coded "slightly injured" in the police database but these persons went to the hospital after the accident. This shows that the police underestimated the severity of injuries.

In the police database 132 persons coded "fatal injured" are not matchable at all. This could be a topic for further investigations as the fatalities stored in the police database are often cross-checked with the Austrian mortality statistic. As this statistic is used for the population statistic, it can be assumed that the quality of this database is very high. Therefore, the high share of 132 missing persons in the hospital database is doubtful. E.g. further investigations on a single record revealed that this person was transfered to hospital A after the accident but transfered to hospital B after some days. In contradiction to the description of the Austrian hospital discharge database there was no record about the stay of the person in hospital A - only about hospital B.

Figure 16 shows the distribution of length of stay in hospital by mode of transport. A higher proportion of of a long stay in hospitals can be shown for pedestrians and motorcyclists including moped riders.

Figure 16: Distribution of length of stay in hospital by mode of transport (only usable matched persons: 5.142 records)



Matrix 2 for the matched pairs

Conversion of ICD-10 Codes to MAIS

In Austria the international statistical classification ICD-10 is used in the hospital discharge database. Before 2001 hospitals reported the main diagnosis by using the classification ICD-9. Therefore, a transformation from ICD-10 to MAIS is carried out.

Within the EC project Apollo a conversion from ICD-10 diagnosis to ISS has been developed at the University of Navarra. Based on this code a conversion from ICD-10 to MAIS is used for the Austrian hospital discharge diagnosis. Not for all ICD-10 codes MAIS can be calculated. In such cases the value of MAIS is "n/a"

Problems occur when adapting the Apollo code to this data. The Apollo code was developed to use up to three diagnoses per record. In the Austrian hospital discharge database only the main diagnosis is stored. It could be assumed that this is the reason why the conversion does not work very properly.

To show the problem, records of the hospital database are chosen, where there is an indication in the hospital database that this person left the hospital due to death. In the database 171 records are marked with this value. Out of these records 126 records show a diagnosis which can be transformed to MAIS. The following table shows the number of cases and the calculated MAIS values. For only one record, a MAIS score of 6 can be calculated. As the source for this table is only the hospital database all records should have a MAIS value of 6 as these persons left the hospital dead.

Table 45: MAIS scores of persons who died in the hospital

MAIS Score	Number of records
1	10
2	40
3	45
4	12
5	18
6	1
Total	126

Therefore, the conversion of the ICD-10 codes to MAIS using only one diagnosis, as in the Austrian hospital database, could be criticised.

Calculated Matrix 2

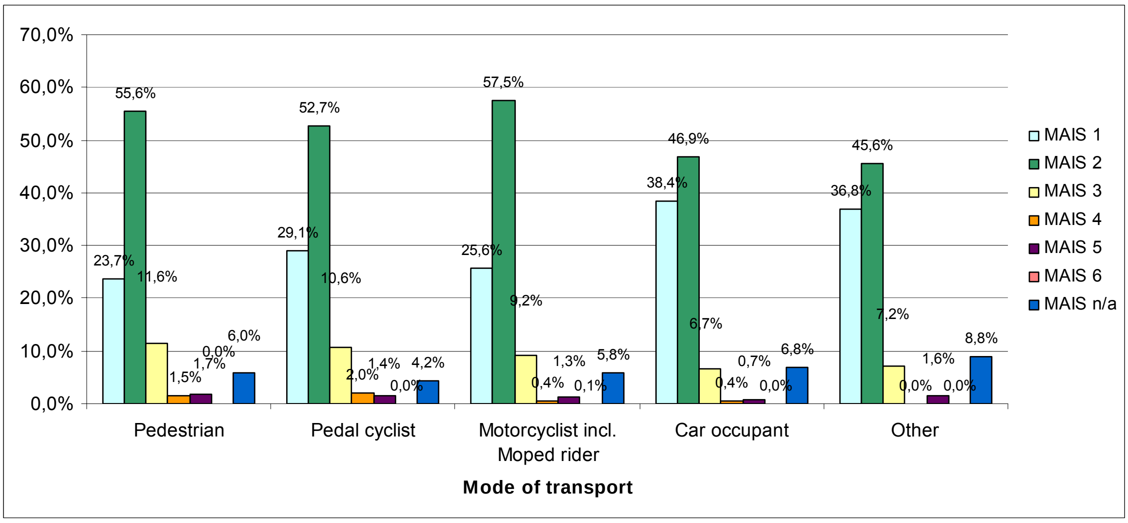
Table 46 shows a matrix where mode of transport, police coding and MAIS is crossed. Only "useable matched" pairs are used in this table.

Table 46: Overview Matrix 2 (only useable matched records)

Sum of Counts		MAIS							
road user type	police coding	1	2	3	4	5	6	n/a	Total
Car occupant	Fatal Injured	5	17	11	4	2		14	53
	Injury unknown	357	305	34	1	1		58	756
	Seriously Injured	273	700	135	6	16		57	1187
	Slightly Injured	497	361	19	2	3		73	955
Motorcyclist incl. Moped rider	Fatal Injured	2	3	1			1		7
	Injury unknown	70	95	10		1		8	184
	Seriously Injured	89	377	66	4	11		31	578
	Slightly Injured	71	46	6				14	137
Other	Fatal Injured					1		1	2
	Injury unknown	19	27	3		1		7	57
	Seriously Injured	36	72	14		2		9	133
	Slightly Injured	37	15	1				5	58
Pedal cyclist	Fatal Injured		1	5	2				8
	Injury unknown	27	52	6	2	4		2	93
	Seriously Injured	60	140	34	5	2		6	247
	Slightly Injured	58	70	8	1	1		13	151
Pedestrian	Fatal Injured	1	6	5	2	2		2	18
	Injury unknown	28	46	4	1	2		2	83
	Seriously Injured	45	185	50	5	4		13	302
	Slightly Injured	53	61	3		1		15	133
Total		1728	2579	415	35	54	1	330	5142

Figure 17 presents the distribution of MAIS by different modes of transport. Although the transformation of ICD-10 to MAIS is doubtful, differences can be observed. Car occupants have a higher share of MAIS 1, especially compared to pedestrians.

Figure 17: Distribution of MAIS by mode of transport (only usable matched persons: 5.142 records)



7.1.5 Conclusions

The computer assisted linking process of the Austrian hospital and the police database described in this report tried to calculate two matrixes which should enable further calculation of underreporting rates (conversion factors).

The design of the calculated matrixes fits the requirements of SafetyNet WP1 Task 5. Nevertheless, the calculated output has to be treated with care. “Useable matched” calculated records are likely not always to refer to the same persons. It can be assumed that 30,5% of these matches are wrong. The reason for this is not the linking procedure itself, but the lack of comparable information on both databases. This lack explains the high share of “not useable matched” records due to selectivity, distance (56%) and the “double zero” records (5%).

The calculated MAIS is based on a transformation of ICD-10 Codes to MAIS. This transformation produces doubtful results. Therefore, no conversion factors based on MAIS can be derived from this linking process.

To calculate underreporting rates for different modes of transport, information about the mode of transport is needed in both databases. The Austrian hospital discharge database does not contain this information. No conversion factors for different modes of transport can be calculated with the result of this report. As the Austrian hospital discharge database contains only in-patients, information about slightly injured persons (who do not go to the hospital) are not treated.

Future steps & recommendations:

- Results of the linking procedure could be highly improved if e.g. the date of birth is also mentioned in the police database. A project called “UDM – Unfalldatenmanagement” with the aim of changing the paper based accident investigation into a computer based investigation could change this situation.
- More analysis is required to detect why there are so many “unmatchable” records.
- More investigation about the recording procedure of persons, who are admitted to a hospital, but than transferred to another hospital later, is needed.
- The new IDB (Injury Database) of DG-Sanco will also contain information about the mode of transport in traffic accidents in near future. Unlike the Austrian hospital discharge database, the IDB is based on a survey carried out in different member states of the EU. More information about the accident will be available with this new database which could be used to relaunch of the calculation of underreporting rates.

7.1.6 References

Apollo Project: http://www.unav.es/preventiva/traffic_accidents/pagina_5.html University of Navarra – *Apollo Project*, visited on 5th of January 2007.

Clark, DE. (2004). *Practical introduction to record linkage for injury research*. Injury Prevention 2004, Vol. 10, pp186-191.

Crash Outcome Data Evaluation System (CODES): http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/codes.html, NHTSA, visited on 20th of June 2007

Howe GR. *Use of computerized record linkage in cohort studies*. Epidemiol Rev. 1998, 20(1):112-21.

Jaro MA. *Probabilistic linkage of large public health data files.* Stat Med. Mar 15-Apr 15 1995, 14(5-7):491-8.

SWOV (2001). *A new linking procedure for the determination of the total number of hospitalised road traffic victims by comparing police and hospital reports*. Netherlands

7.2 Study carried out in Czech Republic

Report prepared by Jan Tecl (CDV)

7.2.1 Introduction

The Czech national study comparison and linking the records of road traffic victims from the police database and the records of treated or hospitalised persons from the hospital database was carried out in the framework of SafetyNet project as probably the first study of this kind in the Czech Republic in 2007.

The two datasets with details of injured persons were for the Czech district of Kromeriz in the years 2003 – 2005. Records from the central police database and the district hospital were used.

7.2.2 Description of data sources

The police database

The police database of the Czech Republic is widely believed to be a quite reliable road accident data source, and has been developed over more than 30 years. It is organised in three levels: district, regional and central. Data collected at the district level are verified and transmitted to the regional level, again verified and transmitted to the central level, again verified and then stored to the central database on the State Police Directory. The data collecting process is carried out by means of the road accident form and this process is now highly computerised. The accident report form contains 74 variables related to the 4 main groups (accident, vehicle, driver / passenger, pedestrian).

By law, there is a legal obligation to report to the police all accidents on public roads, not only those with any person injured but also those with only material damage (with damage over defined financial limit - at present 50 000 CKR, 20 000 CKR from 2001 until 2006 and only 1000 CKR before 2001.

The hospital discharge database

The situation for the hospital data for road accident victims is not so clear in the Czech Republic. There is no central hospital injury database as in other countries. Nevertheless, the statistics of hospitalised and killed persons from all hospitals in the country are collected by the Institute of Health Information and Statistics, that is the administrator of the National Health Information System. The following reports are, among others, processed by this Institute:

Statistics of deceased (national demographic statistics data of deceased persons, amended by the cause of death by diagnosis in ICD-10 system) and

National register of hospitalised (individual hospitals data about hospitalised persons, basic diagnosis and possibly also other diagnoses, possibly operation and death cause diagnoses).

Project co-financed by the European Commission, Directorate-General Transport and Energy

The problem is that only data are available from this Institute (summary statistical publications), but disaggregate data records are available - even without personal identifiers. This strict personal data protection is defined by the law.

From the local point of view, the statistics of separate hospitals vary considerably in their structure and reliability, so it is very difficult to obtain from here some reasonable and utilizable data. Finally, one hospital from Kromeriz in central Moravia (about 70 km from Brno) was chosen, as the example of the hospital with relatively reliable and available data. This hospital operates for the town with 30 000 inhabitants (and the neighbouring area), which is about 0,3 % of total inhabitants of the Czech Republic.

There are two main problems. Firstly, it is uncertain how closely the operational zones of the police district and the hospital match. Secondly, the hospital statistics are predominantly oriented to the medical elements of the case (diagnosis of the injury), while the elements related to the accident circumstances are not the focus of attention of the medical personnel, even in the hospitals with better level of data collection.

7.2.3 Description of the linking process

Variables used for the linking process

The best linking variable would be surely the ID-number of injured person. This possibility is, of course, excluded by the law for personal data protection. Consequently, alternate variables must be used.

Although most of the variables in the police and hospital databases are disjoint, some corresponding variables could be found. The following variables have been chosen for the linking process:

- date of accident
- year of birth,
- sex,
- type of road user.

Skateboard and roller skate users are registered in the hospital database as a specific type of road user, but they are considered as pedestrians for the linking procedure.

Some tolerance is allowed for variables in the linking process:

day of accident: +1 day in the hospital database,
year of birth: ± 5 years
type of road user: some difference in the hospital database is allowed

Other important variables which are necessary for the final evaluation and comparison, are the injury severity (from the police database) and the length of stay in the hospital and MAIS (Maximum Abbreviated Injury Scale). MAIS is derived from the ICD-10 system code. The codes used are:

Road user type:
1=pedestrian
2=pedal cyclist
3=motorcyclist
4=car occupant
9=other

Police severity:
1=fatality
3=seriously injured
4= slightly injured
-1= not matched

MAIS:
1=MAIS 1
2=MAIS 2
3=MAIS 3
4=MAIS 4
5=MAIS 5
6=MAIS 6
9=unknown MAIS'
-1=not matched

Length of hospital stay:
1='outpatient' *(0 night)*
2='overnight' *(1 night)*
3='hospitalised 1-3 days' *(1-3 nights)*
4='hospitalised more than 3 days' *(>4 nights)*
8='hospitalised but unknown length'
-1='not matched'

The third party parameter has not been taken into consideration because there is no corresponding value in the hospital database.

The linking procedure

The procedure used for the data linking is a probabilistic method accomplished by the semi-automatic way with a manual checking of linked records. The data of both groups (police and hospital) are ordered by the date of the accident. Then the records with the same or near linking parameters are gradually offered for the linking in two passes. The definitive linkage can be accepted or rejected.

7.2.4 Results

Table 47: Total number of linked records

	police data	no police data	total
hospital data	266	575	841
no hospital data	845	-	
total	1111		1686

matched records

In total, 266 from 1111 (23,9%) police records were matched among the hospital records. 266 from 841 (31,6%) hospital records were matched among the police records. The total registration rate is 65,9% (1111/1686).

Table 48: Number of linked records for type of user

type of user	police and hospital data	only police data	only hospital data	police total	hospital total	total
pedestrians	28	68	62	96	90	158
pedal cyclists	60	135	422	195	482	617
motorcyclists	19	74	15	93	34	108
car occupants	148	523	76	671	224	747
other	11	45	0	56	11	56
total	**266**	**845**	**575**	**1111**	**841**	**1686**

The registration rate is 60,8% for pedestrians, 31,6% for pedal cyclists, 86,1% for motorcyclists, 89,8% for car occupants and practically 100% for other road users.

Table 49: Number of linked records for age and sex

age / sex	police and hospital data	only police data	only hospital data	police total	hospital total	total
men						
0-14	10	37	62	47	72	109
15-24	44	173	90	217	134	307
25-49	71	258	140	329	211	469
50-69	30	76	63	106	93	169
70+	12	28	13	40	25	53
women						
0-14	7	21	37	28	44	65
15-24	26	67	43	93	69	136
25-49	34	105	76	139	110	215
50-69	27	57	38	84	65	122
70+	5	23	13	28	18	41
total	**266**	**845**	**575**	**1111**	**841**	**1686**

The registration rate is 43,1% for men aged 0-14, 70,7% for men aged 15-24, 70,1% for men aged 25-49, 62,7% for men aged 50-69 and 75,5% for men at least 70 years old (66,8% for all men). The registration rate is 43,1 % for women aged 0-14, 68,4% for women aged 15-24, 64,7% for women aged 25-49, 68,9% for women aged 50-69 and 68,3% for women at least 70 years old (64,2% for all women).

Table 50: Matrix for MAIS and police severity linking

	police severity			
MAIS	1: fatal	3: serious	4: slight	-1: not matched
1: pedestrians				
1	0	1	9	46
2	0	2	8	9
3	0	4	1	6
4	0	0	0	0
5	1	1	0	0
6	1	0	0	1
9	0	0	0	0
-1	4	20	44	0
2: pedal cyclists				
1	0	0	35	358
2	0	3	12	52
3	0	3	2	10
4	0	3	0	2
5	0	1	0	0
6	1	0	0	0
9	0	0	0	0
-1	5	20	110	0
3: motorcyclists				
1	0	2	8	13
2	0	0	5	2
3	0	1	1	0
4	0	0	0	0
5	0	1	0	0
6	1	0	0	0
9	0	0	0	0
-1	2	18	54	0
4: car occupants				
1	0	4	84	61
2	0	7	37	12
3	0	5	4	2
4	0	1	1	0
5	2	1	0	1
6	2	0	0	0
9	0	0	0	0
-1	18	64	441	0
9: other				
1	0	0	6	0
2	0	0	4	0
3	0	1	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
9	0	0	0	0
-1	0	7	38	0

Table 51: Matrix for length of stay and police severity linking

	police severity			
length of stay	1: fatal	3: serious	4: slight	-1: not matched
1: pedestrians				
1: outpatient	1	1	11	52
2: overnight	0	0	1	1
3: 1 - 3 days	0	1	5	6
4: >3 days	1	6	1	3
8: unknown	0	0	0	0
-1: not matched	4	20	44	0
2: pedal cyclists				
1: outpatient	1	2	40	386
2: overnight	0	0	0	2
3: 1 - 3 days	0	3	8	22
4: >3 days	0	5	1	12
8: unknown	0	0	0	0
-1: not matched	5	20	110	0
3: motorcyclists				
1: outpatient	1	2	11	14
2: overnight	0	0	0	0
3: 1 - 3 days	0	0	2	1
4: >3 days	0	2	1	0
8: unknown	0	0	0	0
-1: not matched	2	18	54	0
4: car occupants				
1: outpatient	2	5	93	67
2: overnight	0	0	0	0
3: 1 - 3 days	0	6	27	6
4: >3 days	2	7	6	3
8: unknown	0	0	0	0
-1: not matched	18	64	441	0
9: other				
1: outpatient	0	0	9	0
2: overnight	0	0	0	0
3: 1 - 3 days	0	1	1	0
4: >3 days	0	0	0	0
8: unknown	0	0	0	0
-1: not matched	0	7	38	0

7.2.5 Conclusions

The linking process has been carried out for the first time in the Czech Republic. It seems however, that police data are significantly more reliable than hospital data about accidents victims because the hospital data are not collected so carefully. It would be necessary, for more accurate results, to improve and unify the system of collecting hospital accident statistics. Further, the linking procedure may have been affected by the possibility that the not catchment area may not correspond fully.

It can be seen from the results that the lowest registration rate is for bicyclists (probably most of them were injured in accidents involving a single bicycle) and then for pedestrians (but sometimes it not clear from the hospital database if the pedestrian was really injured in a traffic accident that involved another person).

7.3 Study carried out in France

Report prepared by Emmanuelle Amoros (INRETS)

7.3.1 Introduction

WP1 Task 1.5 aims at estimating factors to correct for under-reporting, to be applied on the CARE data. As in the other countries, the CARE data for France are made of the national police data. The medical data necessary for comparison with the police data are provided in France by a road trauma registry. This registry covers all victims of road crashes that occurred in the Rhône county and who seek medical attention in health facilities of the county or close surroundings. The Rhône county is a large county of 1.6 million inhabitants. It consists of a large city Lyon, its suburbs and a rural area in the north part.
The record-linkage of the police data and the registry data is being conducted every year, as a routine. It is described below. Results of the linkage are also provided.

Only the French specificities of the data and of the record-linkage are described here. The common methodology used to estimate the under-reporting correction factors is described elsewhere.

7.3.2 Description of data sources

Police data

The French police are required by law to write a crash report for every road crash causing at least one casualty. A road crash is officially defined as a crash involving at least one vehicle and occurring on the network open to public traffic. Skateboard or roller skate users are considered as pedestrians by the police, and, as such, are only classified as road casualties if hit by a vehicle. There is no restriction about motorised vehicles, in other words there is no exclusion criteria on bicycles.

The police crash report should report all the people involved in the crash: killed, injured and non-injured ones. Injured are classified into slightly or seriously injured : casualties requiring a hospital stay of 6 days or more are categorised as ‘seriously injured’, whereas casualties requiring less than 6 days of hospital stay (including outpatients) are categorised as ‘slightly injured’.
The police crash report contains detailed information on the crash, the crash environment and conditions, the vehicles involved, but it contains limited information on the people involved.

These police reports are paper reports; most of the information they contain is recorded into electronic files, according to a standardised format.

The police dataset used here is the one restricted to the Rhône county: only crashes that occur in the Rhône county are selected.

Hospital data (road trauma registry)

The registry covers all casualties from road crashes in the Rhône county who seek medical attention in health facilities. Inclusion criteria are broader than the police ones: off-road crashes are not excluded; roller skate, skate-board or scooter users are not considered as pedestrians but as road users using a mean of transport and are hence included, whether hit by another vehicle or not. The registry is based on the participation of all health care facilities in the county (and its close surroundings) that may receive victims of a traffic crash: it includes some 150 health care facilities: from emergency departments, intensive care units, surgery units... to rehabilitation departments, as well as pre-hospital emergency care. The registry includes both inpatients and outpatients, i.e. all casualties, whether hospitalised or not.

Information collected for each casualty consists of a few crash characteristics and of the following casualty characteristics: gender, date of birth, place of residence, hospital stay, hospital transfer if relevant, and accurate injury assessment. Indeed, for each subject, injury assessment is based on the whole set of diagnoses provided by the different health services the subject may have gone through. Plain text diagnoses are coded by the registry physicians according to the Abbreviated Injury Scale (AIS), 1990 revision. Each injury is assigned a severity score, ranging from AIS 1 (minor) to AIS 6 (beyond treatment). To measure the overall severity for casualties with multiple injuries, the MAIS is used (maximum AIS severity score).

Length of hospital stay is estimated by the number of nights spent at the hospital, with the number of nights being obtained from date of discharge and date of admission.

The categories used are the following:
- outpatient (0 night)
- overnight (1 night)
- 1-3 days (2-4 nights)
- >3 days (>4 nights)
- inpatient, but unknown length of stay

Preparation of the data

The data cover the 1996-2003 time period. The (French) police definition of road casualties are applied, as these are the definition of the data in the CARE database. It implies that the following road users have been excluded:

- roller, scooter and skate boards users if not hit by a vehicle,
- uninjured road users.

Fatalities have also been excluded since they have high rates of reporting.

7.3.3 Description of the linking process

Methodology

We have first implemented a record-linkage methodology (Clark, 2004) to be applied retrospectively on the police and medical datasets, once they are available. The 1996-2001 data have been linked this way. The method used is both probabilistic and manual.

It is manual because a major linking variable (place of accident: city/village and details such as street name) cannot be standardised into numerical codes. This variable is unformatted free text which cannot be standardised and coded without important loss of information.
The method is manual in the sense that the person in charge of the linking process goes through any single record (of the police dataset), trying to match it with one from the medical registry dataset; this process is performed on the computer screen, assisted by a user-friendly specific application.

It is a probabilistic method as we allow for some possible error in the linking variables. Typically, we allow for the date of accident to differ by 1 or 2 days, if the other linking variables agree… No matching weights based on probabilities are computed (since it was not possible on one of the linking variables). The decision of linking two records is made based on how many linking variables agree, which ones and on which values.

The linking variables are :

- date of crash,
- time of crash,
- location of crash (town/district/village and details such as road(s) number or street(s) name),
- date of birth (only year and month are available) of the casualties,
- gender of the casualties,
- road user type of the casualties

the most important ones being date of crash, location of crash, year and month of birth of casualties

From 2002 onwards :
In order to improve the exhaustiveness of the registry and the completeness of its information (i.e. reduce the number of missing values), the use of the police data is now part of the registry recording procedure. That is to say, every time a casualty is about to be recorded in the registry (from a notification form), it is first checked whether this casualty can be found in the police dataset. If so, the registry record is created by specifying the link with the record found in the police data and by copying police information about the crash (location, type of crash). If the casualty is not found in the police data, the registry record is created, using information from the notification form.

Linking software

A specific software was developed in Visual Basic ; it works in the Microsoft Access environment.

The software is basically a user-friendly way of comparing the two datasets. It allows for different sorting on the linking variables, and pre-selecting of the records that match on date of accident and date of birth (year and month). It displays casualties records grouped within accident. Values of all different linking variables are displayed. One goes through every police casualty record, tries to find the corresponding record in the medical registry dataset. Two records are linked with a “press-button”, and hence be selected out of the records to be linked.

From 2002 onwards :

A specific software was developed in Visual Basic ; it works in the Microsoft Access environment. It works very much in the same way.

Problems encountered and solutions

Crash location is missing in 16% of the medical registry casualties. Since crash location is a major linking variable, a missing value often means that this casualty will not be linked. One must however keep in mind that at best, since the police file is not even half the size of the registry file, one could match about one registry casualty out of two. In other words, 16% of missing crash location (in the registry casualties) does not mean 16% of missed links. This was confirmed (see section 5 – reliability of the linkage)

There is nothing much that can be done about this missing crash location. The registry staff already sends letters to the casualties when some data is missing, and obtains some but not all.

7.3.4 Results

The results of the linkage are provided in the following tables:

Table 52: Number of linked records

	Number of records from police data	Number of records from hospital data
Linked	21 310 (62.7%)	21 310 (27.0%)
Non-linked	12 668 (37.3%)	56 479 (73.0%)
Total	33 978 (100 %)	77 789 (100 %)

In the police data (restricted to injured), a little less than two thirds were linked with the hospital data. Conversely, less than one third of hospital data were linked with hospital data.

Table 53: Police data and link status according to police severity

Police reported severity	Number of records	Proportion linked
Killed (at 6 days)	excluded	-
Seriously injured	4 862	75.4%
Slightly injured	29 116	60.6%
Non-injured	excluded	-
Total	33 978	62.7%

The proportion of police records linked to the hospital data increase with police severity classification.

Table 54: Hospital data and link status according to MAIS

Length of stay	Number of records	Proportion linked
MAIS 1	56 043	22.9
MAIS 2	15 535	34.9
MAIS 3	4 110	53.1
MAIS 4	752	64.6
MAIS 5 and 6	248	71.4
Unknown MAIS	1 100	16.8
total	77 789	27.0

The proportion of hospital records linked to police records increases with the MAIS, from 23% at MAIS 1 to 71% at MAIS 5-6.

Table 55: Hospital data and link status according to length of hospital stay

Length of stay	Number of records	Proportion linked
Killed (at 6 days)	excluded	-
outpatients	64 112	23.6
overnight	5 062	36.9
hospitalised 1-3 days	2 799	43.0
hospitalised > 3 days	4 187	59.8
Hospitalised, unknown length of stay	1 627	33.9
total	77 789	27.0

Similarly, the proportion of hospital records linked with the police records increases with length of hospital stay.

Casualties characteristics according to registration

We provide distributions of a number of casualties characteristics, according to source registration. This is defined in 3 groups: police-only data, hospital-only data and intersection between police and hospital data. Being based on independent data the 3 groups can be compared. The comparison of casualties characteristics between hospital-only and ‘police intersection hospital’ provides a description of the police “filter” on casualties and hence enables the identification of bias factors for police under-reporting.

Table 56: Road user type according to registration source

Road user type	Police only	Police ∩ hospital	Hospital only
pedestrians	13.3	14.5	7.7 %
cyclists	3.0	3.5	17.7 %
motorised-two wheelers	15.9	17.8	20.7 %
car occupants	63.4	60.6	49.9 %
others	4.4	3.5	4.0 %
	100.0 % (n=12 668)	100.0 % (n= 21 310)	100.0 % (n= 56 479)

There is hardly any difference in road user type distribution between police-only casualties and police∩hospital casualties. This indicates an absence of bias on road user type in hospital reporting.

On the contrary, the distributions of road user type are different between hospital-only and police∩hospital : there are far fewer cyclist casualties in the police intersection hospital data than in the hospital only data. This indicates a bias towards less reporting of cyclists compared to other road user types in the police data.

Table 57: Presence of third party, according to registration source

Third party	Police only	Police ∩ hospital	Hospital only
yes	84.2	84.1	55.7 %
no	15.8	15.9	44.3 %
total	100.0 % (n=12 668)	100.0 % (n= 21 310)	100.0 % (n= 56 479)

What we mean by ‘third party’ is whether a (human) opponent no matter is he/she was injured or not, and no matter if he/she was a pedestrian, or someone in a vehicle, whether motorised or not, and whatever the type (bicycle, car, van , bus, train, tram..).

There is no difference in third party distribution between police-only casualties and police ∩ hospital casualties, indicating no bias on the existence of third party in the crash in hospital reporting. On the contrary, these third party distributions are highly different between the police∩hospital data and hospital-only data: a much higher proportion of casualties with no third party involved in the hospital-only data compared to the police data. This indicates a bias towards lower reporting in the police data of casualties involved in crashes with no third party than casualties involved in a crash with a third party.

Table 58: Crash location: road type, according to registration source

Road type	Police only	Police ∩ hospital	Hospital only	Hospital only
motorways	12.2	11.1	6.7	9.8
state and county roads	30.0	31.9	7.6	11.2
local roads	54.6	54.2	49.7	72.8
off-road, other	3.2	2.7	4.2	6.2
unknown	0.0	0.0	31.7	Not accounted for
	100.0 % (n=12 668)	100.0 % (n= 21 310)	100.0 % (n= 56 479)	100.0% (n= 38 554)

There is hardly any difference in road type distribution between police-only casualties and police ∩ hospital casualties, indicating no bias on road type (where the crash occurred) in hospital reporting.

The further comparison of road type distribution is hindered by a large proportion of unknown in the hospital-only data. However it seems that the distribution of road type is different between hospital-only data and police intersection hospital data: there would a smaller proportion of casualties from crashes on local roads and off-road in the police data than in the hospital-only data. This indicates that the police under-reporting is worse for those crashes. In other words, there is some bias on road type in police reporting.

Table 59: Crash location: "greater Lyon: inside or outside" according to registration source

Urban/rural	Police only	Police ∩ hospital	Hospital only	Hospital only
Lyon	34.8	32.6	19.5	24.7
Lyon suburbs	39.1	47.3	35.6	45.1
outside 'greater Lyon'	25.9	19.9	23.8	30.1
unknown	0.2	0.2	21.1	Not accounted for
	100.0 % (n=12 668)	100.0 % (n= 21 310)	100.0 % (n= 56 479)	100.0% (n= 44 557)

There is some difference in the distribution of where the crash occurred between police-only casualties and police ∩ hospital casualties: casualties involved in crashes that occurred far from Lyon are less often found in hospital data: it is probable that slight casualties who crashed far from a hospital are less likely to go to the hospital than similarly slight casualties who crashed within 'greater Lyon'. In other words there is a bias on urban/rural area in hospital reporting.

Again, there is a large proportion of unknown in the hospital-only data. However the distributions of casualties according to "greater Lyon" (crash location) seems different between hospital-only and Police ∩ hospital: the proportion of casualties who crashed outside the "greater Lyon" is smaller in the police ∩ hospital than in the hospital only. This indicates a worse police under-reporting for these casualties. More generally, it means a bias on this characteristic in police-reporting too.

Table 60: Crash location: police type area, according to registration source

Police type area	Police only	Police ∩ hospital	Hospital only	Hospital only
urban police area	61.0	61.4	36.1	49.8
rural police area	22.5	21.5	30.3	41.9
urban motorway police area	16.5	17.0	6.0	8.2
unknown	0.0	0.0	27.6	Not accounted for
	100.0 % (n=12 668)	100.0 % (n= 21 310)	100.0 % (n= 56 479)	100.0% (n=40 915)

There is no difference in police type distribution between the police-only casualties and the police intersection hospital casualties: this indicates that there is no bias on police type area in hospital reporting.

Again, there is a large proportion of missing data in hospital-only data. However the distributions of police type seem different. There is a higher proportion of casualties from rural police area in the registry than in the police intersection hospital casualties: this indicates a bias on police type in the reporting of casualties in the police data. More precisely, the under-reporting is worse in rural police area. This corroborates with the previous finding of worse under-reporting outside the “greater Lyon” area.

Table 61: Age of casualty, according to registration source

Age	Police only	Police ∩ hospital	Hospital only
0-14	6.5	7.5	13.4%
15-24	29.6	30.7	30.3%
25-49	43.8	43.7	38.2%
50-69	13.4	12.6	8.8%
70 and over	5.6	5.1	3.0%
unknown	1.1	0.4	2.8%
	100.0 % (n=12 668)	100.0 % (n= 21 310)	100.0 % (n= 56 479)

There is hardly any difference in age distribution between police-only casualties and police ∩ hospital casualties. This indicates no bias on age in hospital reporting. There is some difference in age distribution between hospital-only casualties and police ∩ hospital casualties: younger casualties in the hospital-only data. This indicates a slight bias on age in police reporting.

Table 62: Gender of casualty according to registration source

Gender	Police only	Police ∩ hospital	Hospital only
male	60.5	60.3	62.6%
female	39.5	39.7	37.4%
total	100.0 % (n=12 668)	100.0 % (n= 21 310)	100.0 % (n= 56 479)

There is no difference on gender distribution between police-only casualties and police hospital casualties, indicating no bias on gender in hospital reporting. There is a very slight difference in gender distribution between hospital-only casualties and police hospital casualties. This indicates a possible slight bias on gender in police reporting.

Police under-reporting of road casualties and its associated bias factors have been studied using a multivariate analysis (Amoros et al., 2006). It was mainly shown that : 1) police under-reporting is inversely and strongly associated with injury severity, 2) police under-reporting is strongly related to both road user type and involvement of a third party. Casualties in crashes involving a third party (pedestrian or another vehicle) are more police –reported than those without; cyclists are far less police-reported than other road users types. 3) police under-reporting is strongly associated with the combination of road type, crash environment (“greater Lyon”: inside vs. outside) and police force area.

Reliability of the linkage

The reliability of the linkage was assessed on the 2001 data (Amoros et al., 2007). Once the previously described record-linkage was performed, a survey on casualties only identified in the police data was conducted. It consisted in going back to the police paper reports and retrieving additional information. In particular, names were collected. Since names are available in the road trauma registry (since 2000), it was then possible to conduct an additional linkage using names on casualties not previously linked.

First names and surnames (married name if applicable for women) were used. Names were compared using the Soundex code to allow for typing, spelling or transliteration mistakes. Pairs were pre-selected on matching first name and surnames (Soundex coded) ands then linked if also matching on year and month of birth, date of crash and crash location. For the 2001 data, the standard record-linkage yields 2813 linked casualties, 1322 police-only casualties and 7823 registry-only casualties.

The additional record-linkage was able to find 148 additional linked casualties, in other words 3.6% of the police casualties, or 1.4% of the registry casualties. We checked why these additional linked casualties were not found by our “standard” record-linkage: it was either because there were 2 or 3 errors in the major linking variables or because the crash location was missing in the registry.

A second assessment of the linkage reliability was provided by estimating the number of false positives and false negatives (Amoros et al., 2007). False positives are pairs linked whereas corresponding to two distinct casualties; false negatives are non-linked pairs whereas corresponding to the same casualty (where casualty is defined by an individual and a crash, since an individual can be involved in more than one crash). The estimation method was largely inspired by two papers : Brenner, 1994 and Brenner and Schmidtmann, 1996. It is based on probability computations and approximations of these: probability of disagreement in the linking variables (because of errors) for any pair of records from the same casualty, and probability of agreement (by chance) for any pair of records from distinct casualties.

On the 2001 data, it was estimated that there were 97 false positives and 396 false negatives. This corresponds to 2.3% and 9.6% respectively of the police records, and 0.9% and 3.7% of the trauma registry records. It further means that the number of linked pairs should be increased by 299 (396-97) and the number of non-linked pairs decreased by the same amount.

In conclusion, the number of matches missed by the record-linkage is reasonably small. Compared to the total number of casualties of the aggregate dataset (police U registry), it is of course smaller.

7.3.5 Conclusions

As regards to the generalisation to the whole country of correction factors estimated at the Rhône county level: the underlying assumption is that the police practices (of road casualties reporting and severity classification) are homogenous throughout the country. This is supported in France by the centralised structure of each police type. As an example it means that the degree of police under-reporting of cyclists is assumed to be the same whatever the region. By estimating a correction factor for each road user type, the variation of the distribution of road user type throughout France can be taken into account. The same is true for injury severity; injury severity varies across France since it varies between urban and rural areas. Road user type and Injury severity are the two under-reporting bias factors used by the common methodology.

There are other important characteristics that display both varying degree of police-reporting and varying distributions throughout France : road type, urban/rural or police type, and possibly third party involvement. We have shown that they are biasing factors in police reporting of casualties. These characteristics display different distributions across France: road type distribution does vary since some counties do not have motorways for instance. Urban/rural distribution or its correlated police type distribution does of course vary. Third party involvement distribution may vary since it is correlated with traffic density and hence with urban/rural distribution.

The fact that in this study we take account of only two under-reporting factors (those requested for all countries in the project: injury severity) does therefore reduce the quality of the estimations. An estimation of the totality of road casualties with under-reporting correction factors estimated according to 5 major bias factors is being conducted; it also includes the estimation of the number of non-observed (non-reported) casualties, through the capture-recapture approach. It will be published separately.

7.3.6 References

Amoros E, Martin J L, Laumon B, 2006. *Under-reporting of road crash casualties in France*. Accident Analysis and Prevention, 38(4), 627-535

Amoros E, Martin J L, Laumon B, 2007. *Estimating non-fatal road casualties in a large French county, using the capture-recapture method*. Accident Analysis and Prevention, May;39(3):483-490

Brenner, H., 1994. *Application of capture-recapture methods for disease monitoring: potential effects of imperfect record linkage*. Methods of Information in Medicine 33 (5), 502-506.

Brenner, H. and Schmidtmann, I., 1996*. Determinants of homonym and synonym rates of record linkage in disease registration*. Methods of Information in Medicine 35 (1), 19-24.

Clark, D. E., 2004. *Practical introduction to record linkage for injury research*. Injury Prevention 10 (3), 186-191.

7.4 Study carried out in Greece

Report prepared by George Yannis, Petros Evgenikos, Antonis Chaziris (NTUA) Eleni Petridou, Nikos Dessypris (CEREPRI)

7.4.1 Introduction

This report describes the Greek national study on the identification of the road accident underreporting level within the framework of Task 1.5 of SafetyNet WP1, aiming to estimate the real numbers of road accident casualties, by addressing the issue of Police under-reporting. The results of this report will allow for the development of appropriate correction coefficients to be applied on the Greek road accident data for the estimation of the real number of road accident casualties.

Moreover the development of under-reporting coefficients will allow for comparisons between the statistics provided by several EU countries on non-fatal injuries. Currently, the only comparable measurement units available in CARE, the European Union road accident database with disaggregate data, are the numbers of fatal accidents and of people killed, where the degree of underreporting is acceptably small in most EU Member States. The same is not true for non-fatal accidents and for people suffering non-fatal injuries, therefore the numbers of non-fatal accidents and of people seriously and slightly injured cannot be compared between the several EU countries.

Given that the present study aimed to identify possible links between accident data derived from the Accident and Emergency Departments of the hospitals and the Police road accident data files, it was anticipated that data incompatibility problems would be encountered as the data collection systems used by the hospital Emergency Departments and the Police differ not only on their variable sets but also on the common variables definitions. In order to develop the data matrices (as described in the common methodology) to allow for the development of the under-reporting coefficients, a methodological tool comprising the following steps was adopted.

The primary target in order to achieve compatibility between the hospital and the Police data was to define an appropriate study area to ensure that no accident casualties reported by the Police were transferred to any hospital other than the one under investigation. This target was achieved by selecting the General Regional hospital of Corfu as the source of medical road accident casualty data, thus the criterion of a precisely defined "catchment" area was met, as all casualties within the island of Corfu are primarily transferred to the emergency department of this hospital, even though some more severe injuries can be subsequently transferred to better equipped hospitals, mainly in the city of Athens.

In order to perform record linkage between the hospital and the Police data files one should primarily identify a set of common appropriate variables to be used for the data linking, and ensure that the same definitions concerning these variables were adopted in both databases (e.g. fatal injuries are considered those in which death has occurred within 30 days following an accident). This objective was fulfilled by adopting transformations and/or aggregations in specific variables of the hospital database in order to be comparable with the respective Police database variables.

The final step concerned the data matching procedure itself. By using the selected variables and values the file with common records in both databases (matched cases) was extracted. These records could be grouped by any of the selected variables and values allowing their further processing, whereas non-matching cases were grouped in a separate file, which would also be used in order to produce the data matrices needed for the elaboration of the correction coefficients.

7.4.2 Description of data sources

The study was based on data from two different sources, the Greek Emergency Department Injury Surveillance System (EDISS) and the Road Traffic Police database.

Hospital data files

The Greek Emergency Department Injury Surveillance System (EDISS) was the first surveillance system covering all type of injuries and is operated in the Emergency Departments of four hospitals in Greece, two of which are located in the Greater Athens area (Asclipeion Trauma Hospital and A. Kyriakou Children's Hospital), the third being in the county of Magnesia, in Greek mainland and the fourth on the island of Corfu. The first hospital is dedicated mostly to adult trauma patients, while the second hospital is one of the two major Children's hospitals in the Greater Athens area and covers on alternate days three quarters of the childhood (<14 years old) population of this area. The last two hospitals are regional (district) hospitals, covering the population of the respective two administrative regions of Greece. The two regions are: the partly rural and partly industrial region of Magnesia in the Greek mainland and the tourist island of Corfu in the Ionian sea.

In order to successfully link road casualty records between the hospital and the Police databases, the selected Police road accident data set should refer to the "catchment" areas of the examined hospitals. The EDISS database includes data from four hospitals, two of which are located in the capital of Greece. The "catchment" area of these two hospitals cannot be precisely delimited because of the large number of hospitals situated in the greater area of Athens. Data from the Regional hospital of Volos are also difficult to be linked with the respective Police data files, as a significant number of admissions to this hospital concern casualties from the adjacent interurban road network but also from the "Sporades" group of islands located outside the bay of Magnesia, therefore a specific "catchment" area cannot be precisely defined. The only hospital collecting the necessary medical data and meeting all the requirements about a predefined "catchment" area is the Regional hospital of the island of Corfu, as all road accident casualties on the island requiring medical treatment are primarily carried to this hospital (even cases which are subsequently transferred to other hospitals). The common methodology for the identification of the underreporting level developed within the framework of SafetyNet Task 1.5 was applied to medical data from this hospital as well as police accident data for the prefecture of Corfu, for the period between 1996 and 2003.

Data collection in EDISS was carried out by specially trained medical staff that interviewed patients suffering from any type of injury (or their guardians). A pre-coded questionnaire was used, complemented with a short free text describing the injury event. The questionnaire is a modified version of the basic form used by all European Union member states, which participate in the European Home and Leisure Accident Surveillance System. The Greek version of this questionnaire also includes additional variables on traffic and occupational injuries, thus the whole spectrum of injuries, by nature, external cause and intent is covered by EDISS. The questionnaire covers socio-demographic variables, injury characteristics and treatment of the injured individual. For those who are eventually hospitalised additional data are also collected. The recorded information is entered in a computerised database with continuous data quality control system.

More specifically, the variables recorded in the Greek EDISS are:

Accident mechanism
Activity at the time of the accident, Sports
Location of the accident
Products involved
Accident description
Patient: Age, sex, nationality
Treatment: Follow-up treatment
Date and time of attendance
Length of stay
Injury severity score (ISS)
Diagnosis: Type of injury (2 possible injuries)
Injured body part (2 possible parts)
Administrative information: Country code, Hospital identification number, Patient identification number
First aid
Place of residence
Vehicle type (for road accidents)
Occupation

Data derived from the EDISS database concern the years 1996-2003, as after 2003 the data collection was suspended.

Police data files

Data collection on road accidents involving casualties (injury or death) in Greece is carried out by the Police at national scale since 1963. Whenever an injury or death occurs as a result of a road accident, a special force of the Police (Department of Road Accidents) carries out an investigation (not only on-site). Initially the Police fills-in on the spot an autopsy report.

Data of the autopsy report are not computerised, but they are processed and analysed only at a general level (total numbers, etc). Furthermore, the main characteristics of an accident (cause, type, time, day, place etc.) are registered in the incident book of the Police department in charge of the accident.

Data collected by the Police in the appropriate collection form, which is further processed by the National Statistical Service of Greece (NSSG), are computerised and available for further analyses. The NSSG competent service receives by the departmental police offices all road accident collection forms filled-in by the police officers in charge of the accident. Two copies are filled-in, one of which is submitted to the NSSG, while the other is kept in the local Police department.

The road accident variables recorded by the Police are the following:

Accident related

Date (year, month, day of the month, day of the week, hour)
Location
Area type
Road type
Number of casualties
Number of vehicles involved
Pavement type
Weather, lighting and pavement conditions
Accident type
Vehicle manoeuvre

Vehicle related

Vehicle type
Vehicle nationality
Vehicle make and model
First registration year
Vehicle age
Number of drivers and passengers
Alcotest data (hour, place, results)
Driving license data (nationality, category, age)

Person related

Road user type
Gender
Age
Nationality
Use of safety equipment
Casualty severity
Position in vehicle
Trip purpose

All data derived from the Police database concerned the prefecture of Corfu and the years 1996-2003, referring to the same area and time period as the EDISS road accident casualty records.

Data processing

In order to link road accident casualty data, a common subset of variables was used to identify the cases to be linked in the two databases. These variables were: road user type (driver, passenger, pedestrian, bicyclist); time of occurrence (year, month, day); age of the road user (in single years); gender of the road user; nationality of the road user and mode of transport. The time of the accident was included in the dataset but was not used for the data linking. This variable was not considered reliable in the EDISS database, as it was recorded while interviewing the casualty/patient (or his/her guardian) at the emergency department of the hospital, in some cases many hours after the accident occurred. Consequently, time of the accident was used as supplementary information for linking cases for which other information was missing or the information provided by the other the variables was insufficient.

The definition of the "mode of transport" variable differs between the EDISS and police collection systems, therefore necessary transformations were implemented. More specifically, the value "pedestrian" is included in the EDISS "mode of transport variable", while it is not included in respective variable of the police data file (for pedestrian casualties, the motor vehicle which collided with the pedestrian is recorded, and the value "pedestrian" is recorded as value only in the "person class" variable). In order to obtain a compatible value set, the value "pedestrian" in EDISS was replaced by the vehicle which collided with the pedestrian (which was obtained by a third variable included in EDISS). Therefore the values included in the EDISS variable "mode of transport" are passenger car, motorcycle, moped, bicycle, small truck, large truck and bus.

As the data files used for the purposes of this study included road accident casualties on the island of Corfu, any particularities that arise from the characteristics of the selected study area (demographical, traffic related etc) were taken into account. More specifically, during vacation periods and especially during the summer months, a large number of foreign tourists are visiting the island of Corfu, explaining the relatively high proportion of road accidents involving foreign people recorded during these periods. This fact was exploited to increase the efficiency of the data linking procedure by including the variable concerning the nationality of the road user in the set of variables used for the data linkage.

Apart from the common variables and values, some additional variables included in the hospital dataset were also included in the data file with the linked records. These variables concerned the Length of Stay (LoS) in the hospital, and the Abbreviated Injury Scale (AIS score) for each casualty recorded. Concerning the AIS scores, these were extracted from the ICD9 scores for each casualty recorded in the hospital database. These additional variables were used to produce the matrices described in the common methodology, allowing for the further processing and the calculation of the under-reporting coefficients.

The "Length of stay" variable is widely used among hospital databases in Europe, although it is not defined in the same way throughout the countries. The Greek definition of length of stay is based on whether the date has changed while the patient was being hospitalised. For example, if the patient was hospitalised while the date has not changed (e.g. from 05:00 to 19:00 at the same day), or if the patient was hospitalised while the date changed only once (e.g. from 18:00 to 09:00 in the next morning), then the length of stay would be

1. If the patient stayed hospitalised while the date changed twice, the length of stay is 2, etc.

Therefore, the Greek definition of the LoS is not necessarily related to the nights spent, but with the change of date while the patient was being hospitalised. It is also evident that the length of stay will be 1, even for some cases where the patient did not spend the night in the hospital.

Any ethical issues that arise from the use of person identification information related to road accident casualties were also taken into account. Variables concerning personal information (casualty names, personal addresses, date of birth etc) were not included in the data files used for the record linkage. Any information that could be used for the identification of a casualty at personal level is considered confidential and could not be used for research purposes.

The data linking was carried out in two distinct phases, each one concerning fatalities and injuries respectively. Concerning non-fatal injuries, the severity of the accident is defined differently in the hospital and the police databases. More specifically, the police defines injury severity using personal judgement of the police officer filling-in the accident form, while the hospitals define injury severity on the basis of the days of hospitalisation. Thus, many injuries defined as serious in the hospital database could be identified as slight by the police and vice versa. Consequently, linking the police serious injuries with the hospital serious injuries and the police slight injuries with the hospital slight injuries could result an important loss of potentially linked cases which were recorded according to a different injury severity definition in the two databases. On that purpose, the data linking was carried out using the total number of injuries (serious and slight together) in the two databases.

After completing the data linkage for all injuries, an additional task was performed in order to identify the proportion of casualties for which the accident severity was differently reported (serious in the one database and slight in the other) in the hospital and the police data files. This proportion was relatively high (only 47,8% of the linked records' injury severity agreed between the two databases) verifying the choice to link aggregated injury type data files.

7.4.3 Description of the linking process

The Regional Hospital of Corfu was the only one out of the four hospitals collaborating in EDISS meeting the clearly defined "catchment" area requirements. Indeed, all road traffic casualties occurring on the island and requiring medical treatment are initially assessed in the Emergency Hospital Department, even the most severe cases which are subsequently transferred in other hospitals for specialised care. The linking procedure was applied on data concerning the Emergency Medical Department of the Hospital and the Road Traffic Police data for the prefecture of Corfu between 1996 and 2003.

The present study aimed to link hospital and police road accident casualty data files. The two data files included a different series of variables and values, therefore all transformations / aggregations required to obtain the data grouped by the same variables (as already described in "Data description") were adopted. After all necessary adjustments have been made, all data grouped by the linking variables were included in a single file which was the input for the linking process.

In order to identify the common records between the two datasets specialised statistics software was used. The input file included both hospital and police records for the whole series of years grouped by the variables used for the linking. The utility "*identify duplicate cases*" was used to identify the matched pairs of records in the data file, based on the given set of linking variables.

The result of this procedure grouped the data file in the following way:

- A case from one file matched with one case from the other file (perfect matches).

- A case from one file matched with more than one cases from the other file. In that case the additional variables (such as the time of the accident) were used to identify the most likely among the various matches. The selection was done manually.

- Two (or more) cases from the same file were matched. These cases concerned accidents with identical characteristics in the same database, therefore they were not taken into account in the data linkage.

- More than one cases from one file were matched with more than one cases from the other file. Once again, additional variables were used to identify the most likely pairs among the various possible matches.

- Non-matched cases.

During the linkage process, however, it was revealed that some values of the common set of variables used for the cross-checking did not match for several records in both datasets, although they seemed to refer to the same casualty. Inconsistencies between the values could be attributed to incorrect reporting from the part of the patient in the case of the hospital data or to misjudgement on the part of the person completing the collection form, in the case of the police data or simply due to an error while processing the data. In order to include these cases to the linked data file one of the following two approaches were adopted:

1) Repeating the linking process using each time a different subset of less common variables as the criterion for linkage in order to examine a broader range of possible matches. For example, when the nationality of the road user was unknown in the police database and known in the EDISS database, excluding the nationality variable from the matching subset of variables

would allow inclusion of the case in the possible matching groups, based on the matching results of the remaining variables

2) Adopting less strict rules while linking the data, thus, allowing a tolerance interval for the values of certain variables. Consequently, some cases could be matched although slight differences on the values of one linking variable, was observed, while the rest of the variables provided sufficient evidence for linking the cases. The variables for which such tolerance intervals were allowed were the following:

   Date of the accident (a tolerance interval of one or two days before and/or after the actual date of the accident could be allowed).

   Nationality of the road user - only in cases where the nationality was recorded as "Albanian" in the police data file and as "from North Epirus" in the EDISS data file. North Epirus is a region officially within the borders of Albania with a minority of Greek population, which is recorded using a separate value in the EDISS database. A large number of people from North Epirus are living on the island of Corfu.

   Age of the road user (a tolerance interval of 5 years was adopted in some cases, in order to include possible rounding of the age in the databases).

   Mode of transport variable (some cases where "moped" was recorded in the one data file and "motorcycle" in the other were linked, when sufficient evidence from the rest of the variables was provided. Such links were established also for a few cases where small truck was recorded in the one data file and passenger car in the other).

Each time the record linkage was taking place, a set of linked hospital-police records was being copied into a separate file containing the linked cases. All other records were entering the next iteration.

After performing several times the same procedure, using each time different variable sets all possible matches were identified. The iterations were stopped when the set of variables used for the data linking was judged to be small enough therefore unable to provide sufficient data for a pair of records to be linked.

Obviously, when less strict rules for the record linking were used, the manual checking for each pair of records was more important in order to avoid linking irrelevant records. The records left in the end were the non-matched cases. These data were transferred into separate files for the hospital and the police data. This file contained road accident casualty records, recorded by the police and not recorded by the hospital and road accident casualties recorded by the hospital and not recorded by the police.

7.4.4 Results

Summary results of the overall linking results are provided in this section. Moreover, some additional results concerning the presentation of the data linkage in disaggregate form are presented in the following Tables. More details concerning the main results of the Greek national study, the data matrices used for the extraction of the under-reporting coefficients, as well as data concerning the length of stay or the AIS scores for the casualty records, are provided in the main report.

The main results of the record linkage between the hospital road accident casualty database records of the General Regional hospital of Corfu and the police records referring to the respective "catchment" area are summarised in the following Table 63.

Table 63: Summary of linking results

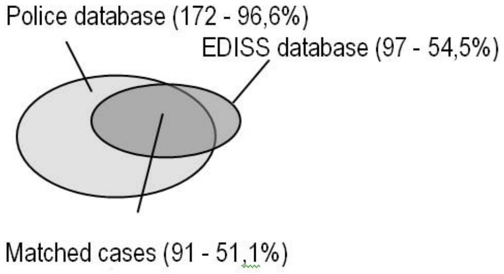
	Fatalities	Injuries
Records in Hospital database	97	11.267
Records in Police database	172	1.910
Linked cases	91	1.262
Not linked hospital records	6	10.005
Not linked police records	81	648
Total records (real number)	**178**	**11.915**

During the eight year study period (1996-2003) 11.364 road accident casualties contacted the Corfu Hospital Emergency Medical Department and out of them 97 (0,9%) died on arrival to the hospital or during hospitalization. For the same period 2.082 road accident casualties were reported by the police and out of them 172 (8,3%) were declared as deaths.

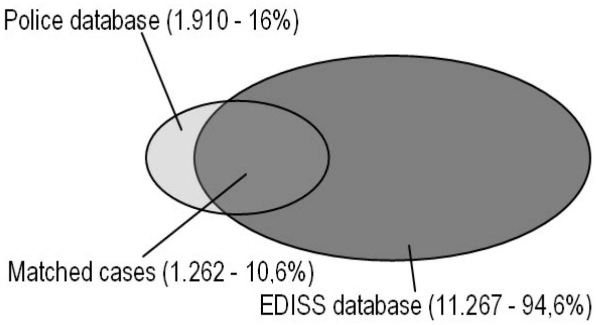
Only 6 out of the total 178 cases (3,3%) concerning fatalities were not reported in the police database, while the respective percentage for injuries is significantly higher (10.005 out of 11.915, corresponding to a proportion of 84%). Additionally, a high percentage of road accident fatalities are not reported in the hospital database, most likely due to a large number of deaths at the accident site, never contacting the emergency department of the hospital. The number of non fatal injuries reported by the Police and not reported in the EDISS database is significantly lower (648 out of 11.915, 5,4%).

The results of the linking procedure are presented below in schematic form.

Project co-financed by the European Commission, Directorate-General Transport and Energy


Police database (172 - 96,6%)
EDISS database (97 - 54,5%)
Matched cases (91 - 51,1%)

Results of data linkage for fatal injuries


Police database (1.910 - 16%)
Matched cases (1.262 - 10,6%)
EDISS database (11.267 - 94,6%)

Results of data linkage for non-fatal injuries

The results of the linking procedure were further processed in order to provide the figures concerning the hospital and the police casualty reporting in disaggregate form. More specifically, the following Tables present the linking results by the casualty's age and gender, mode of transport (for pedestrian casualties the linked vehicle was recorded) and nationality.

Table 64: Fatalities by age groups and gender

	Linked			EDISS only			Police only		
Age groups	Male	Female	Total	Male	Female	Total	Male	Female	Total
<15	1	0	1	1	0	1	3	1	4
16-18	4	1	5	0	0	0	4	0	4
19-24	11	6	17	0	1	1	14	2	16
25-40	25	5	30	0	1	1	15	3	18
41-64	15	4	19	2	1	3	13	5	18
65+	7	5	12	0	0	0	15	6	21
Unknown	4	3	7	0	0	0	0	0	0
Total	67	24	91	3	3	6	64	17	81

Table 64 presents the overall linking results for fatal injuries in a disaggregate form, by age and gender of the road user. The small number of non-reported cases though (only 6 cases not reported by the police) does not allow for further conclusions concerning the police underreporting by age and gender.

Table 65: Injuries by age groups and gender

Age	Linked				EDISS only				Police only			
groups	Male	Female	Unknown	Total	Male	Female	Unknown	Total	Male	Female	Unknown	Total
<15	43	24	0	67	408	222	0	630	19	18	0	37
16-18	119	35	0	154	910	398	1	1.309	22	18	0	40
19-24	222	83	0	305	1.527	926	1	2.454	99	48	0	147
25-40	266	134	0	400	2.042	1.148	6	3.196	158	63	0	221
41-64	161	70	1	232	1.181	545	2	1.728	81	50	1	132
65+	75	22	0	97	442	194	2	638	34	21	1	56
unknown	6	1	0	7	36	14	0	50	5	5	5	15
Total	892	369	1	1.262	6.546	3447	12	10.005	418	223	7	648

The linking results concerning non-fatal injuries are presented in disaggregate form (by age and gender) in Table 65. As it can be derived by the above Table, the police underreporting level of people younger than 15 years old is larger than the respective level in any other age group, as 90,4% of the respective hospitalised casualties are not included in the police database. On the other part, people aged more than 65 years old are more recorded by the police than any other age group, as the relevant underreporting level is 86,8%. With reference to the gender, females tend to be less recorded by the police (9,7% of the women included in the EDISS database were matched in both database, comparing to 12% of the men). Moreover, from the above Table 3, it can be concluded that males younger than 15 years old from the casualties included in the EDISS database, are less recorded by the police, comparing to the males of any other age group (90,5% of the males aged less than 15 years old are not recorded). On the other part, females aged 16-18 years old are less recorded by the police, comparing to any other female age group (91,9% of the females aged 16-18 years old are not recorded).

Table 66: Fatalities and injuries by vehicle type

	Fatalities			Injuries		
Vehicle type	Linked	EDISS only	Police only	Linked	EDISS only	Police only
Bicycle	0	0	0	8	120	5
Tricycle	0	0	0	5	5	3
Moped	13	0	19	224	3	108
Motorcycle	33	0	29	498	6.303	211
Car	39	2	28	452	1.545	286
Small truck	2	0	2	51	35	26
Truck	1	1	0	5	17	2
Bus	0	0	2	9	16	3
Tractor	0	0	1	7	1	1
Unknown	3	3	0	3	1.941	3
Other	0	0	0	0	19	0
Total	91	6	81	1.262	10.005	648

Table 66 shows the linking results of both fatalities and non-fatal injuries by the casualty's mode of transport (for pedestrian casualties the related vehicle is recorded). The number of casualties related to motorcycles is bigger than the respective number for any other mode of transport, both fatal and non-fatal accidents. More specifically, almost 63% of the EDISS only non-fatal injuries refer to motorcycle related accidents, while the respective figure for passenger cars is approximately 15%.

Table 67: Fatalities and injuries by nationality

	Fatalities			Injuries		
Nationality	Linked	EDISS only	Police only	Linked	EDISS only	Police only
Greek	67	6	57	958	7.532	342
Albanian	5	0	5	72	535	46
Italian	3	0	3	26	321	30
British	6	0	8	96	392	97
German	4	0	5	40	388	39
Other	6	0	3	70	837	94
Total	91	6	81	1.262	10.005	648

Table 67 presents the linking results concerning both fatalities and non-fatal injuries by nationality. These results could prove useful for identifying differences in the underreporting levels between natives and foreigners. Excluding Greek casualties, the highest proportion both in fatal and non-fatal injuries is observed for the British people, followed by the Albanians.

In summary, the results presented in these Tables provide a more detailed view of the road accident casualty data linkage between the EDISS and the police databases for the prefecture of Corfu. It is evident though that disaggregate data concerning fatalities refer to a small number of records and their further division in subgroups (age, gender etc) cannot be used for further calculations due to the small size of the sample. Nevertheless, the non-fatal injury figures present a great potential for further analyses as the sample size is considerably larger.

7.4.5 Conclusions

The methodology used to link road accident casualty data between the hospital (EDISS) database and the respective police data files is based on using a statistics utility for identifying duplicate cases within a single data file. The procedure used for the data linking as described in Section 7.4.3 is not fully computerised and its correct application strongly depends on the manual checking of the linked records. Although it provides reliable results (every linked record pair is manually checked and it is very unlikely that a matched pair will not be found) it would be very difficult to implement such a procedure on large data files (containing a long time series and a bigger area of study), as the manual checks of the record matches would be extremely time consuming. In the present study the sample consisted of 12.000 records approximately, most of which could not be linked to another record (only 1.262 matched cases).

The input file used within this study consisted of both hospital and police records for the whole series of years (1996 - 2003) grouped by the six variables used for the linking (road user type, date of occurrence, age of the road user, gender and nationality of the road user and mode of transport) and other variables which would serve either as supportive information on whether a record pair should be linked or not (for cases where more than two records were linked using only the six variables) or as information to be included in the data matrices required for the extraction of the correction coefficients (as described in the common methodology). The inclusion of several variables in the common data file made possible to extract the linked and not linked data files in a disaggregate form, using each time a different variable (or set of variables) to describe the data, therefore it can allow for the development of more detailed correction coefficients.

A study concerning road accident casualty under-reporting coefficients at a national level should always be based on appropriate data, meaning that the sample should not only have a statistically significant size (e.g. for a series of years) but also being representative for the selected country. Therefore, in order to evaluate the results and the conclusions from this study, any restrictions that may arise from the collected data should be taken into account. The data used for the present study concerned the island of Corfu, representing only a proportion of the overall road accident casualties in Greece, being also an ideal "catchment" area for matching hospital and police data. The high presence of tourists during the summer period does not necessarily affects the matching of the two files, however further investigation could be proved useful.

This study revealed not only the great potential for linking hospital and police data but also the need for further investigation of such links in order to come up with larger samples and thus more complete results.

7.4.6 References

Dessypris N., Petridou E., Skalkidis Y., Moustaki M., Koutselinis A. and Trichopoulos D. (2002). *Countrywide estimation of the burden of injuries in Greece: a limited resources approach*. J Canc Epidem & Prev 2002; 7: 123-129.

Petridou E, Gatsoulis N, Dessypris N, Skalkidis Y, Voros D, Papadimitriou Y and Trichopoulos D (2000). *Imbalance of demand and supply for regionalized injury services: a case study in Greece*. International Journal for Quality in Health Care; Vol 12; Issue 2; p.115-113.

7.5 Study carried out in Hungary

Report prepared by Dr. Gábor Merényi, Olivér Zsigmond, Árpád Tóth, Dr. Péter Holló (KTI)

7.5.1. Introduction

In Hungary, all hospitals report on their discharged patients to Gyógyinfok, the firm engaged in data management at the National Health Insurance Fund. The report includes the discharged patients' date of birth, the time of their hospitalisation and discharge, their gender as well as the social insurance identification number (TAJ). On the basis of this identification, the case may be followed even if the patient is transported to another hospital. In Hungary, too, disease (injury) recording is implemented in accordance with the BNO-10 (ICD-10) code system. In case of accident, the accident cause has also to be registered in the BNO-10 code system. Traffic accidents, and only this type of accidents, all begin with „V", therefore they can be easily selected from the database.

From the aspect of the task, unfortunately the Gyógyinfok Database has several disadvantages:

- Because of data protection aspects, the TAJ is not included into the database of the police, therefore the person cannot be identified as simply as that
- Because of different financing, the attended outpatients are recorded in another database, and until 2006 it was not compulsory to code the cause of the accident, therefore, traffic accidents cannot be retrieved from this database.
- The BNO (ICD) –10 codes are not suitable for specifying the AIS (MAIS) severity scale demanded by the task with the software available with the database.

Because of the above reasons and for the sake of more precise data collection, the whole national database has not been used, but all data from one selected trauma centre were analysed and compared with the police data.

7.5.2. Description of data sources

This centre is the Károlyi Sándor Hospital in Budapest. This hospital is one of the 4 Regional Trauma Centres in Budapest. All the injured adults of the traffic accidents occurring in five northern districts of the Capital (III.,IV.,XIII.,XIV.,XV.), and in the approximately 25-30 settlements of the agglomeration, as well as on the northern main roads (2,2A,2B,3,M3,10,11) leading to the Capital are brought here on five days of the week (Tuesday-Wednesday-Friday-Saturday-Sunday) throughout 24 hours.

From the hospital's computerised database, the data of all the patients were processed who had been injured in traffic accidents and hospitalised in the

Project co-financed by the European Commission, Directorate-General Transport and Energy

hospital's department in the period between 1 August 2004 and 31 January 2006.

The following data have been selected in order to be compared with the police database and to be processed:

The patient's date of birth, gender.

Accident time, location. Type of the vehicle involved in the accident, the role of the injured, the accident's mechanism (if there was any reference in the data on health).

The attendance case of the injured has been classified on the basis of hospitalisation: out-patient, overnight, 1-3 days, or over 3 days attendance.

According to **AIS (Abbreviated Injury Scale)** injury severity has been determined for 6 body regions (head & neck, face, chest, abdomen-spine, extremity, external). Accordingly severity of different body regions' injuries are classified from 0 to 6. In compliance with the categorisation of the American Association for the Advancement of Automotive Medicine, the different scales are follows:

1 – Minor
2 – Moderate
3 – Serious
4 – Severe
5 – Critical
6 – Unsurvivable

MAIS (maximum of the AIS), the highest AIS value has been determined for each injured person. The AIS values were coded directly by medical staff.
ISS (Injury Severity Score) has been calculated in each injury case separately. Calculation of this generally applied indicator: the total amount of the AIS squared value calculated for the three most seriously injured body regions . This value may be 0-75. (If on some body region the AIS=6, then the ISS will be automatically 75.)

Furthermore, the number of persons **deceased** within 30 days in hospital have been recorded.

7.5.3. Description of the linking process

Starting data:

- **ACCIDENT data set**:
  KSH (police), 2005CA.dbf
  28783 records, out which 2665 records within catchment

- **HOSPITAL** data set
  KORHAZ6.dbf
  1294 records
  (only in the catchment, adult injured, not complete (e.g. falling by cycle not included, only 5 days of the week are concerned)

The data sets, clarified and coded were available in this item only for 2005. In the hospital the data have been coded for a longer period, as described in 7.5.2, but the problem is that the official data for 2006 are available only as of mid 2007. Therefore, at the time of the linkage we could use mainly the data for 2005.

Matching

According to its content, the HOSPITAL data set has been divided into two parts:

- K1-complete: the key fields (DATE, AGE, GENDER, PLACE) determinant from the aspect of matching are available
- K2-incomplete: some parts of the data are missing

The ACCIDENT data set has been divided into two parts:

- B1-in the catchment: if the accident occurred in the catchment area of the hospital
- B2-outside the area

The process of matching was implemented in three successive steps:
1) K1 – B1 (result: 474 data-series agree out of 1012 cases)
2) K2 – B1 remainder (result: 22 data-series agree out of 282 cases)
3) (K1+K2) remainder – B2 (result: 4 data-series agree out of 798 cases)
matched total: 500 events out of 1294 cases

Breakdowns:

Both data sets include	500 cases
Only the HOSPITAL data set includes	794 cases
Only the ACCIDENT data set includes (in the area of catchment)	2165 cases

The process of matching was carried out by computer, not manually. Experience showed that the difference allowed in the accident time (±1 hour) was too limited: the values in the hospital database are rather uncertain as they are based on declaration, or assessment. As a follow-up to this work, this tolerance should be increased, then the degree of matching of the two databases would certainly be greater.

Conclusions concerning the linking process

In the first step of matching, the agreement of gender + age + date + time + site has been investigated. The criterion of congruency is the precise agreement of the gender and the date, in the case of the age ±1 year, in the case of the time Ki-Bi<=1 hour, for the site the criterion was the geographical correlation, however in the process of correlation from these approximations only one reduction could be allowed for at once.

In the second step of matching, (there was no information on site) the precise agreement of age + gender + date has been required, however 1 hour time difference was allowed.

If 00h00 was entered in the HOSPITAL database (which in most cases implies „not filled in"), time agreement was approved.

In the data field „role of the road accident injured", only driver/front- and rear-seat passenger/pedestrian cases are differentiated, therefore no more detailed differentiation is possible.

In the B1 data set there are 3 records that could be matched (the injured appeared in the hospital records), but according to the data item „was the patient taken to the hospital", the road user was not hospitalised. This was considered as a recording mistake or a subsequent decision of the injured.

There were about 40 records where the injured was taken to hospital, but according to another data field the patient attended as an outpatient

There was one record where the injured died in the hospital according to the data field, nevertheless the MAIS value was not 6

It can be imagined, that following an injury which in itself is not considered as dangerous to life, an aged patient will decease later in the hospital, but the death may be caused by illness such as pneumonia.

In some cases the data referring to the role of the accident participant (recorded in both data sets) were different, in this case the Police data has been accepted

Whole days are missing from the ACCIDENT data set (e.g. 16.07.2005, 04.12.2005)

The structure of the result matrices is the same as the structure and the content of the „Methodology for SafetyNet Task 1.5".

7.5.4. Results:

During the **18 months of the survey 2107** persons injured in traffic accidents attended the hospital.

Role:	driver		1307	
	Front seat passenger		52	
	Rear seat passenger		75	
	Unknown passenger		342	
	Pedestrian		331	
	TOTAL		2107	
Nursing days:	Outpatient		1101	52,3%
	1 day		54	2,6%
	1-3 days		465	22,1%
	>3 day		487	23,1%
	TOTAL		2107	
	MAIS	1	985	46,7%
		2	688	32,7%
		3	319	15,1%
		4	67	3,2%
		5	40	1,9%
		6	8	0,4%
	TOTAL		2107	
	Deceased		9	0,42%
	ISS	0-15	1942	92,2%
		16-30	117	5,6%
		31-45	34	1,6%
		46-60	6	0,3%
		61-75	8	0,4%
	TOTAL		2107	
Motorcycle	373			
Passenger car	946			
Bus	32			
Heavy vehicle	17			
Trailer	1			
Special vehicle	0			
Tramway	2			
Trolleybus	5			
Suburban railways	0			
Train	0			
Bicycle	356			
Moped	40			
Animal	0			
Other	4			
Pedestrian	331			
TOTAL	2107			

Relation between vehicle type, injury severity and the period of stay in the hospital

		Nursing days				
vehicle	MAIS	outpatient	overnight	1-3 days	>3 days	Summary
car occupant	1	417	4	69	9	499
	2	84	14	149	50	297
	3	1	1	22	79	103
	4			1	29	30
	5				14	14
	6		3			3
car occupant Sum		502	22	241	181	946
motorcyclist	1	126	2	21	2	151
	2	63	1	36	24	124
	3	6	1	13	57	77
	4				11	11
	5				9	9
	6				1	1
motorcyclist Sum		195	4	70	104	373
pedal cyclist	1	156	4	8	4	172
	2	79	2	28	12	121
	3	9	6	6	34	55
	4				4	4
	5			1	3	4
pedal cyclist Sum		244	12	43	57	356
pedestrian	1	72	3	28	9	112
	2	30	7	54	23	114
	3	1		6	61	68
	4		1		20	21
	5				12	12
	6		3		1	4
pedestrian Sum		103	14	88	126	331
other	1	41		9	1	51
	2	14	2	12	4	32
	3	2		2	12	16
	4				1	1
	5				1	1
other Sum		57	2	23	19	101
	1 part	812	13	135	25	985
	2 parts	270	26	279	113	688
	3 parts	19	8	49	243	319
	4 parts		1	1	65	67
	5 parts			1	39	40
	6 parts		6		2	8
Summary		1101	54	465	487	2107

Relation between vehicle type and the severity scale of the body's regional severity:

Head&Neck	car occupant	motorcyclist	pedal cyclist	pedestrian	Other	Summary
1	395	30	53	80	24	582
2	199	41	41	78	13	372
3	6	5	8	6	1	26
4	5	3	1	7	1	17
5	11	5	4	10	1	31
6		1		3		4
Summary	616	85	107	184	40	1032

Face	car occupant	motorcyclist	pedal cyclist	pedestrian	other	Summary
1	61	10	9	14	1	95
2	14	6	7	20	1	48
3	2	2	1	2		7
4				1		1
Summary	77	18	17	37	2	151

Chest	car occupant	motorcyclist	pedal cyclist	pedestrian	other	Summary
1	145	26	19	33	12	235
2	101	29	11	25	5	171
3	33	10	6	16	2	67
4	7	2		2		11
5		1				1
6	2			1		3
Summary	288	68	36	77	19	488

Abd.Spine	car occupant	motorcyclist	pedal cyclist	pedestrian	other	Summary
1	78	29	14	29	6	156
2	13	5	1	3	4	26
3	17	16	2	7	2	44
4	9	5	1	4		19
5	2	1		3		6
6	1					1
Summary	120	56	18	46	12	252

Extremity	car occupant	motorcyclist	pedal cyclist	pedestrian	other	Summary
1	151	112	101	76	30	470
2	50	55	64	40	12	221
3	65	64	44	67	12	252
4	18	6	2	19		45
5	4	3		3		10
Summary	288	240	211	205	54	998

External	car occupant	motorcyclist	pedal cyclist	pedestrian	other	Summary
1	364	166	198	155	46	929
2	78	61	28	61	5	233
3	5	2		2		9
Summary	447	229	226	218	51	1171

Relation between vehicle type and injury severity

	MAIS						
Vehicle	1	2	3	4	5	6	Summary
car occupant	499	297	103	30	14	3	946
motorcyclist	151	124	77	11	9	1	373
pedal cyclist	172	121	55	4	4		356
pedestrian	112	114	68	21	12	4	331
other	51	32	16	1	1		101
Summary	985	688	319	67	40	8	2107

7.5.5. Conclusions

More than half of the injured (52,3%) have been attended as outpatients. 23,1% of the injured had been hospitalised for more than 3 days.

Accordingly, 80% of the injured suffered minor or moderate injuries (contusions, lesions, simple fractures, concussion).

5,5% of the injured had very severe, critical or fatal injury. 9 injured (0,4%) deceased. Of course, the latter means a death in the hospital and within 30 days.

Cycle accidents are frequently not included in the police database, and also the hospital data are deficient (e.g. the accident site), since very often neither the injured nor the doctor considers the falling with a bicycle on the public road as a traffic accident. In these cases the injured does not call out the police, and usually the hospital does not inform the authorities either. At the same time, relatively serious injuries may also occur in this category (e.g. shoulder fractures, dislocations).

Head, chest and abdominal injuries were in higher proportion in the case of injured car occupants, whereas pedestrians, motorcyclists and cyclists suffered more frequently of extremity-injuries.

The number of persons injured in passenger cars was almost three times higher than that of motorcyclists, cyclists and pedestrians separately. The accident victim rate in the case of these three latter groups was the same. Against them, the number of injured in other vehicles was insignificant.

Nevertheless, the number of critical and fatal cases involving pedestrians was the same as that concerning the car occupants (i.e. in proportion threefold majority). The rate of very serious and fatal injuries were also higher in the case of motorcyclists than motorists.

Accordingly, while in the case of car occupants the number of injured attended as out- and overnight patients were in majority in comparison with those hospitalised for several days, this rate is reversed if pedestrians are concerned, because almost twice as many injured had been attended in hospital for several days.

Representativeness of the results

We have compared the different injured categories for the catchment's area of the hospital and for the whole capital (Budapest). The distributions of the injured persons are as follows:

Vehicle categories	Sample		Budapest	
	injured	%	injured	%
Motorcycle	260	20,09	428	8,36
Car	546	42,19	2619	51,16
Bicycle	236	18,24	232	4,53
Pedestrian	194	14,99	1315	25,69
Other	58	4,48	525	10,26
All	1294	99,99	5119	100,00

It can be seen, that the sample was not representative even for Budapest. The percentages of injured motorcycle and bicycle riders are greater in the sample than in Budapest, and the percentages of injured car occupants and pedestrians are less. The representativeness is even less regarding the whole country.

The following are the differences between a trauma centre in the capital and that one in a county hospital:

a) No child under 14 will be admitted to the former (admission is usually possible to country centres)
b) There is no non-stop inspection over the 7 days of the week (5 days were investigated by the hospital)
c) There are out-patients' district dispensaries where the patient can present himself (in the country they are mostly substituted by surgeries linked to hospitals, so that they are also registered in the hospitals' database)

Studies in the future

Due primarily to likely lack of the representativeness of the sample of matched data, it is proposed to:

a) involve one countryside centre into the studies
b) carry out a prospective study in order to make possible recording of the predetermined data. (The character of the study detailed here – as it turns out from the particulars – has been retrospective.)

Another possibility can also be imagined for future studies. Besides taking into consideration the aspects of the protection of the personal data, an attempt should be made to study the databases of the central public health (GYÓGYINFOK) and of the police (KSH) on the basis of BNO-10. The coding of the accident cause is already compulsory for the hospital and outpatient cases, thus all the codes starting with “V” correspond to traffic accidents. Moreover, also the role of the injured can be derived from this. The TAJ number is personal data, but a crosscheck could be made on the basis of the date of birth and the starting date of hospitalisation.

Combined Police and Medical data set - HUNGARY

LENGTH of STAY matrix

		Police coding				
Road user type	Length of Stay	Fatal	Serious	Slight	not matched	Total
Driver	out-patient		11	79		90
	Overnight		1	4		5
	1-3 days		19	54		73
	>3 days		74	21		95
	not matched	39	255	802		1096
Passenger-front	out-patient		1	39		40
	Overnight	1		3		4
	1-3 days		3	23		26
	>3 days		12	5		17
	not matched	11	67	303		381
Passenger-rear	out-patient		2	19		21
	Overnight	1	1	2		4
	1-3 days			14		14
	>3 days		10	1		11
	not matched	4	52	254		310
Pedestrian	out-patient		1	16		17
	Overnight	1	2	3		6
	1-3 days		7	27		34
	>3 days		30	13		43
	not matched	29	111	238		378
not matched	out-patient				460	460
	Overnight				21	21
	1-3 days				175	175
	>3 days				138	138
Total		86	659	1920	794	3459

MAIS matrix

		Police coding				
Road user type	MAIS	Fatal	Serious	Slight	not matched	Total
Driver	1		10	83		93
	2		27	59		86
	3		54	14		68
	4		11	1		12
	5		3	1		4
	not matched	39	255	802		1096
Passenger-front	1		2	38		40
	2		2	27		29
	3		10	5		15
	4		2			2
	6	1				1
	not matched	11	67	303		381
Passenger-rear	1		1	21		22
	2		2	14		16
	3		4	1		5
	4		4			4
	5		2			2
	6	1				1
	not matched	4	52	254		310
Pedestrian	1		3	28		31
	2		8	27		35
	3		18	4		22
	4		7			7
	5		3			3
	6	1	1			2
	not matched	29	111	238		378
not matched	1				377	377
	2				283	283
	3				102	102
	4				18	18
	5				14	14
Total		86	659	1920	794	3459

7.6 Study carried out in the Netherlands

Report prepared by Niels Bos (SWOV)

Within the framework of the Safetnet project, a study was carried out into linking hospital data and police data on traffic casualties in The Netherlands. Both databases contain a number of key variables that enables matching of patients in one database with casualties in the other. This was done in order to estimate the level of underreporting and to verify the severity that the police assigns to a casualty. Police estimated severity not always appeared to be of the same severity in the hospital database, as was investigated by using the MAIS scores. Correction factors have been determined which, applied to the CARE accident data, should describe the development of *real numbers* of *severely injured* casualties. These factors are different by mode of transport. However, under-reporting and differences in distribution over relevant variables, such as year of accident and age of the casualty, prevent outcomes being very reliable.

At the moment, linkable Dutch hospital data is available at SWOV for the years 1997-2005 in a consistent database using uniform coding with ICD9-cm. In this study the years 1997-2003 are used. Linkable police data is available for the years 1976-2005, of which the CARE database contains the years 1991-2003.

7.6.1 Introduction

In this study the real number of hospitalised road casualties is derived from linking a hospital discharge file to an accident file of police reported road accident casualties. Based on the intersection of "distance based matches" and both rest files, an estimate is made for the number that was not reported in either database. This leads to a total number of casualties that can also be examined on their medical severity. This severity (MAIS) was determined on all injury codes that are in the hospital file and form an objective scale of severity. The severity will also be expressed in Length of Stay.

The linking study is extensively described in Dutch, in Reurings, Bos & Van Kampen (R-2007-8). In this report only an overview is presented and for details reference is made to the SWOV study.

For a correct assessment of developments in road safety and of the effects of road safety measures and interventions the availability of reliable figures on traffic injury outcomes is crucial. In recent years, the governments of many countries have started to set targets for their future national road safety. These targets are usually expressed in terms of numbers of traffic injury outcomes. Setting up realistic targets also depends to a large extent on the availability of accurate figures on current traffic injury outcomes. In the Netherlands, as in many other countries, fatal outcomes in traffic accidents are registered quite well, but this is generally not the case for hospitalised road traffic casualties.

Transport

Project co-financed by the European Commission, Directorate-General Transport and Energy

In order to determine the degree of underreporting of hospitalised road traffic casualties in Dutch police reports, an extensive study was carried out in which the records of two well-established databases were compared. The first database is the so-called AVV database of police-reported road accidents, which contains extensive information on the accident site, and on the vehicles and casualties involved in the accident. The second database is the so-called Prismant-Hospital Discharge database. This is the official database containing information on all patients admitted into Dutch hospitals. In this database reporting is mainly concerned with medical issues.

The police database only contains information on road traffic casualties and accidents, and is known to be incomplete (Polak, 1997). Moreover, both hospitalised and non-hospitalised casualties are recorded in this database. The hospital database, on the other hand, only contains information on hospitalised patients and is known to be (almost) complete. However, many of these hospitalised patients are not road traffic casualties, while road traffic casualties are not always recorded as such.

The aim of the present study was to asses the actual number of non-fatally injured but hospitalised road traffic casualties from these two databases. Thus, the group of interest consists of all casualties of a road accident (according to the international definition) who have been inpatients in a hospital as a result of the accident, and who did not die within 30 days after the accident. The determination of the maximum AIS score allows us to set a boundary in order to identify real seriously injured road traffic casualties among them. In contrast, road traffic casualties who die within 30 days of the accident (whether hospitalised or not) belong to the group of fatally injured road traffic casualties.

Considering the nature of the two Dutch databases, the seriously injured road traffic casualties can be classified into four subgroups: the hospitalised road traffic casualties recorded in both databases, those recorded in only one of the two databases, and those that are missing in both databases. Therefore, the group that has to be recovered consists of the four cells enclosed with double lines in Table 68, which contains all possible combinations.

Table 68: Possible combinations of presence or absence of hospitalised road traffic casualties in police and hospital databases.

	In hospital database	Not in hospital database	Not a hospitalised road traffic casualty
In police database	In both databases	Only in police database	Road traffic casualty but not hospitalised
Not in police database	Only in hospital database	In neither one database	
Not a hospitalised road traffic casualty	Hospitalisation not caused by a road traffic accident		

To identify these four subgroups of the hospitalised traffic road casualties, a linking procedure was applied which compares a number of key variables contained in the two databases on a record by record basis.

The linking procedure used in the present report is well suited for situations where unique identifiers like a personal ID number are not available for linking. A generalised distance function is defined which quantifies the similarity between pairs of records in the two databases. This quantified similarity can be used to assess the probability of the correctness of a match: the smaller the distance, the higher the probability that the two records refer to the same individual.

The linking procedure is probabilistic and conjunct. It is probabilistic because discrepancies between records are tolerated, including missing information. The procedure is conjunct because it simultaneously compares all the records in the first database with all the records in the second database[3], and therefore only requires two passes through the data. Other probabilistic methods for linking police and hospital records like GIRLS (Generalised Iterative Record Linkage System) are disjunct. In GIRLS records are grouped according to their scores on a key variable, and compared within each group. Matched records are then removed from the linking process. In the next pass the remaining records are grouped according to their scores on another key variable, and again compared within each group, etc., etc. This repeated grouping of records and removal of linked records creates order effects, and requires many passes through the data (see, e.g., SWOV, 2001 or Rosman et al., 1996).

In the conjunct method two records are matched when they are each other's nearest neighbours in terms of distance or similarity. Moreover, the difference between the distance of a matched pair of records and the distances to their two next best neighbours is used to quantify the selectivity (or exclusiveness or uniqueness) of the match. This selectivity measure provides a second diagnostic for the probability of correctness of a match.

The police and hospital databases, as well as the key variables used for linking, first are described in section 7.6.2. Then, in section 7.6.3 a generalised distance function is introduced which quantifies the similarity between records in the two databases, even if information is missing in one or both records. The linking procedure itself is described and applied to the Dutch police and hospital databases. In section 7.6.4, results are analysed in terms of the severity (Maximum AIS) and Length of Stay in hospital (LoS). In this section also the results that are required within the SafetyNet project are given. Finally, in section 7.6.5 conclusions will be drawn.

[3] Except for the time frame. Records are only compared in our study if they are within -1 to +3 days apart.

7.6.2 Description of data sources

7.6.2.1 The police database

In the police database for road accidents (VOR) all accidents in the Netherlands in which at least one person is injured are collected. This statistic should cover 100% of all relevant cases in The Netherlands, but we know that hospitalised casualties are underreported by about 40% (Polak, 2000). The police collects the data by filling in a paper report form. All these papers are collected by the Ministry of Transport (AVV) and transformed into an electronic format. Persons who successfully committed suicide (confirmed) are excluded from the database.

For the linking procedure we used data from the years 1997-2003. In these years, 324.717 persons were either injured or killed in a road accident according to the police. As in most of the European countries, the 30 day definition for fatalities is used: If a person dies within 30 days after the accident the person is counted as a road accident fatality. Uninjured persons involved in road accidents where excluded from the linking process. The casualties were coded by severity, of which only category 6 is defined as hospitalised:

Table 69: Number of casualties in the police database by severity.

	code	description	Number of cases (1997-2003)
Fatalities	0	Killed on the spot	4.431
	1	Killed later the same day	1.435
	2	Killed one day after	537
	3	Killed 2-5 days after	452
	4	Killed 6-10 days after	253
	5	Killed 11-30 days after	301
Hospitalised	6	Hospitalised	79.984
Slightly injured	7	Transferred to hospital, not hospitalised	90.738
	8	Transferred to hospital, hospitalisation unknown	11.304
	9	Not transferred to hospital	126.402
	10	Transfer and hospitalisation unknown	8.880
		SUM	324.717

7.6.2.2 The hospital discharge database

In the hospital discharge database (LMR) administrative and medical data of in-patients of all Dutch hospitals are collected. This statistic is supposed to cover 100% of all relevant cases. Outpatients are not recorded in this database. Data is collected by the hospitals and transferred to Prismant which prepares the hospital discharge files.

Project co-financed by the European Commission, Directorate-General Transport and Energy

We received records for patients that were injured in a traffic accident, but also of cases in which the cause was unknown (E-codes E800-E829+E928+ E958+E988). In this database the main diagnosis from hospital treatment is recorded. This is determined afterwards, at the moment that the patient is discharged. The hospitals use the international classification of diseases version 9 (ICD-9CM, 1980). The hospital discharge database is patient oriented, but if a patient is transferred from one hospital to another hospital he will be reported twice. Some variables make it possible to filter out patients that were treated in another hospital afterwards, however this is not a 100% check.

The data are published on an annual basis. Due to some late transmittals of hospitals, there is a delay of 6 months between the year of discharge and the year in which the data is published.

For the linking procedure data from 1997 up to 2004 were used in order to prepare a database with recorded patients who were admitted to a hospital in the years 1997-2003, so discharges in 2004 from casualties of accidents in 2003 were added. Similarly discharges in 1997 from accidents in 1996 were excluded from linking.

To limit the number of records in the hospital database to a reasonable amount, the linking procedure was carried out for each year separately. Fatalities and patients treated shorter than one day (day-treatment, opposed to clinical treatment) were included, in order to prevent avoidable mismatches. Records with an indication that it was the second or third treatment as a consequence of only one accident were removed, as well as records with an indication that the patient was treated in another hospital before. This excluded 6773 records from the linking process (3,3%). When applying this limitation the database contains 200.766 records.

Table 70: Number of patients in hospital database (admittance).

E-code	Description	number of admittances 1997-2003
E810-819, excluding 817	Motor vehicle accidents	87.382
E826-829 excluding 828	Accidents without motor vehicles	40.479
E817+E828	Not a moving vehicle	7.404
E820-825	Not on a public road	3.009
E800-807	Train accident	249
E958	Suicide	1.945
E928+E988	Not specified	60.298
	SUM	200.766

All these records were used in the linking process, however afterwards, records were excluded when they do not meet the definition of a road traffic accident. If a hospital record matches with a police record, it can be argued that if the police and the coders at AVV think it is a traffic accident it should be included. The people in the hospital are assumed to be less accurate in judging whether it was a train accident or a road traffic accident or whether the road was public or not.

The length of stay in the Dutch hospital file is defined as the number of nights stayed. Besides the regular clinical stays, there are also day-treatments, where normally the length of stay is 1, in rare cases 2 or 3. In order to be comparable with the studies in other countries we agreed to redefine the Length of Stay (LoS) as:

Table 71: Definition Length of Stay.

Dutch database		Definition used
type of treatment	length of stay in LMR	LoS
day-treatment	..	<1
clinical treatment	1	Overnight
	2	1
	3	2
	..	..
	N	N-1

Deriving the maximum AIS and ISS

In the Dutch hospital database the injuries of each patient are coded in 9 fields, by using ICD9-cm. This enables the application of software to determine the scores on the Abbreviated Injury Scale (AIS) per body region, as well as the Maximum AIS score and Injury Severity Scale (ISS). This was done using the software ICDmap90 (Johns Hopkins University, 2002).

In the hospital file used there is a maximum of 10 injury codes, of which at least one E-code indicates that it is a (traffic) accident. From identical codes, only one is stored. Code V714 (the stay is for "observation") is recoded to 89999, and then only ICD injury codes between 80000 and 99999 are kept, others are removed from the record. Crush injuries (8626, 'Multiple and unspecified intrathoracic organs, without mention of open wound into cavity') are recoded to the slighter 9221 ('contusion of chest wall'). This is because they are not as severely injured as ICDmap90 suggests (this can be seen on their length of stay and their type of discharge; there are no fatalities among them).

The record ID, patients age and up to 9 injury codes were exported as text file and analysed by the software. The ICDmap90 output[4] was linked back to the original data. Documentation on ICDmap90 suggests using the terms ICD/AIS and ICD/ISS in order to emphasize the difference between derived scores with direct chart-based scoring of AIS. In the remainder of this report we omit this "ICD/" when speaking of AIS, MAIS or ISS: they are all derived from ICD.

[4] parameters used: Ignore unknowns=on, ABCD=off, Severity mapping=low.
output variables: AIS injuries in AIS1990 (predot 1-12), AIS severity (AIS 1-12), Maximum AIS score (MAIS), Injury Severity Score (ISS), ISS body region (ISS_BR1-12).

If there is no injury present in the record, the values MAIS=0 and ISS=0 are returned (this can be a hospitalization for observation purpose, or an injury outside the range 800-959[5]). A value 9 is returned if there is an injury, but its severity cannot be classified by ICDmap90. Scores 1 to 6 denote increasing severity, from slight to untreatable.

As our experience with this software was rather limited, the quality and characteristics of the medical database itself is analysed. Relations to some other variables are investigated: The following issues will be dealt with:

A. The number of diagnose codes per patient;
B. The influence of the level of detail of the injury coding on MAIS and ISS;
C. The MAIS scores of fatalities, compared to survivors;
D. Finally, we assess the length of stay for cases in which MAIS equals 0 or 9 (page 132).

Of course, results for MAIS scores depend on the number of diagnoses that are available. Also the level of detail that was used in the particular database is important. Some hospitals do not always use the lowest possible level, or data providers limit the level of detail to the first 3 digits, truncating the forth and fifth digit.

An example ICD-9-CM Diagnosis

Table 72: ICD9 to AIS.

ICD9 code	description	AIS predot	AIS severity
893	Open wound of toe(s) 893 is a non-specific code that cannot be used to specify a diagnosis		1
893.0	Open wound of toe(s) without complication	810600	1
893.1	Open wound of toe(s) complicated	810600	1
893.2	Open wound of toe(s) with tendon involvement	840802	2

The AIS scores of these injuries are 1, except for the last one, 893.2, which scores 2. If the database or doctor only specifies an injury code 893, the derived AIS severity may be underestimated.

In order to explore the implication of these effects in the databases on the derived injury severity we investigated the level of detail of the Dutch hospital database and simulated truncation of the 5-digit code. We also simulated the effect of having a main diagnosis only, compared to the Dutch maximum of 9, see below.

[5] ICDmap90 software is able to map ICD9 codes in the range 800-959 to AIS injury codes and corresponding severities, except for 905-909 (late effects), 930-939 (foreign bodies) and 958 (early complications)). So patients having one or more codes in these exclusion ranges or above 960 may have an underestimate of their MAIS and ISS score. In the Dutch hospital file, the percentage of codes that cannot be mapped by ICDmap90 is 0,53% (main diagnosis only, average 1984-2005), so this is only a small problem.

Transport
Project co-financed by the European Commission, Directorate-General Transport and Energy

A Number of injury diagnoses

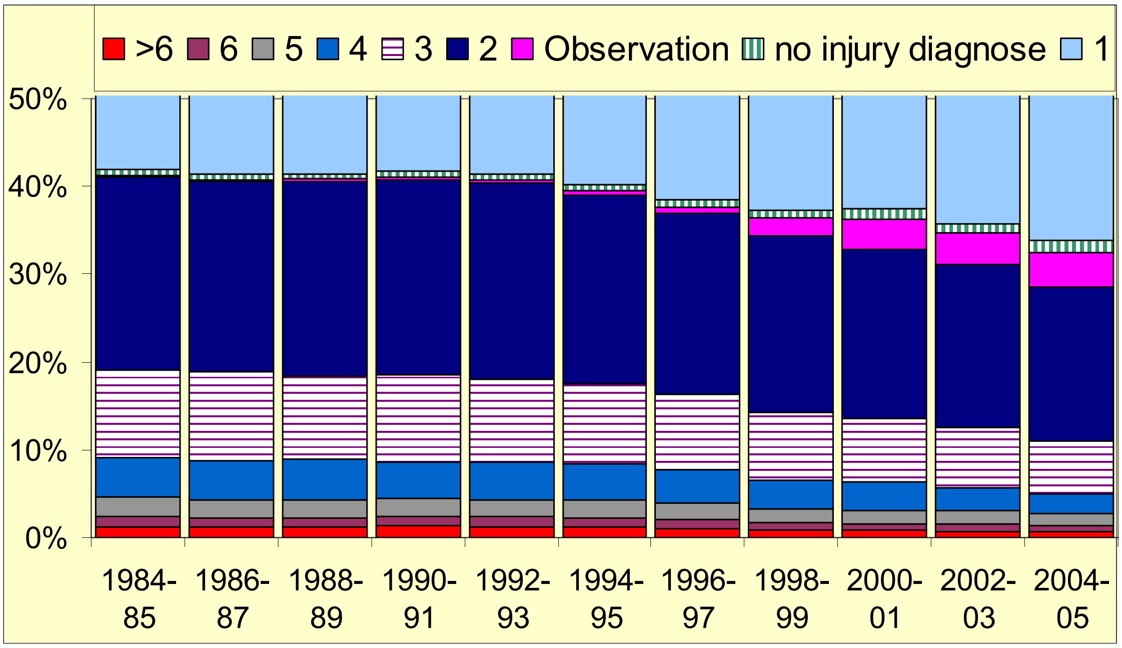
First we have examined the number of diagnoses over time.
We analysed the 1984-2005 discharge databases, for patients that have an external cause with E-codes E800-E829. This selection resulted in 758.733 injury codes (unique codes per patient). The annual number of unique injuries has decreased from 38.600 in 1984 to 30.900 in 2005. In the same period the number of patients also decreased from 22.000 to 19.400 and thus the average number of injuries decreases from 1,78 to 1,56 per patient.
When we look at the distribution of the number of injury diagnoses per person, we see an increasing number of patients that only has one diagnose. In the last two decades, this has increased from 58 to 66%. The number of patients having no injury code at all and the number of observations (V714) are increasing and treated separately in Table 73.

The observed (coded) number of injuries is different from Rosman et al. (1996), who used data from traffic accidents in Australia (1988). She found 10% of no diagnose/ observation, where we find only 1% in the same year and 5% for the current years of interest. The fraction having one main diagnose only is 42% with Rosman, 2 diagnoses 22% and all other fractions almost double the percentages found in the Netherlands. So compared to the Dutch database the number of patients without injury is low, and the number of patients having multiple injuries is also low. This may point to an incomplete reporting, however this can also be related to the number of coded diseases (diagnoses outside the range 800.00-999.99) or to the number of duplicate injury codes that enter the database. In the current analysis these are left out. In the 2003 database, the average number of unique diagnoses per patient is 2,77. This consists of 0,30 diseases, 1,62 injuries and 1,08 E-codes, among which 0,25 duplicate codes.

Table 73. Development of the number of Diagnoses. LMR1984-2005, filtered to patients having an E-codes in E800-E829.

Nr of	1984	1986	1988	1990	1992	1994	1996	1998	2000	2002	2004
diagnoses	-85	-87	-89	-91	-93	-95	-97	-99	-01	-03	-05
no injury diagn	0,6%	0,6%	0,7%	0,6%	0,6%	0,8%	0,9%	0,9%	1,2%	1,1%	1,3%
Observation	0,3%	0,2%	0,2%	0,3%	0,4%	0,5%	0,8%	2,1%	3,4%	3,7%	4,1%
1 =main	58%	59%	59%	58%	59%	60%	61%	63%	63%	64%	66%
2= 1 sub	22%	22%	22%	22%	22%	21%	21%	20%	19%	19%	17%
3	10%	10%	10%	10%	9%	9%	9%	8%	7%	7%	6%
4	4,5%	4,5%	4,5%	4,2%	4,3%	4,1%	3,8%	3,3%	3,2%	2,7%	2,4%
5	2,2%	2,0%	2,1%	2,0%	1,9%	2,1%	1,8%	1,6%	1,5%	1,5%	1,2%
6	1,1%	1,1%	1,1%	1,1%	1,1%	1,0%	1,0%	0,8%	0,8%	0,8%	0,7%
7	0,6%	0,6%	0,5%	0,7%	0,6%	0,6%	0,6%	0,5%	0,5%	0,4%	0,4%
8	0,3%	0,4%	0,3%	0,4%	0,4%	0,3%	0,2%	0,2%	0,2%	0,2%	0,2%
9	0,3%	0,2%	0,2%	0,3%	0,2%	0,2%	0,2%	0,1%	0,2%	0,1%	0,1%

Figure 18: Distribution of patients with their number of diagnoses for readability reasons the remaining 50% is left out of this figure.



From this data it can not be judged weather multiple injuries are less frequent in reality or that this a reporting issue, reflecting the quality of the database.

B Level of Detail of the recorded injury codes

When studying the level of detail of the ICD9-cm injury codes, we see that only a small minority of injuries is not coded at the lowest level available. Since 1992 all injuries are coded at the lowest available level.

Table 74: Percentage of injuries that is not coded at the lowest level. LMR1984-2005, filtered to E-codes E800-E829.

% not at lowest level	1984 -85	1986 -87	1988 -89	1990 -91	1992 -93	1994 -95	1996 -97	1998 -99	2000 -01	2002 -03	2004 -05
main diagnosis	1,08	0,78	0,52	0,17	0	0	0	0	0	0	0
All 9 diagnoses	1,03	0,67	0,42	0,13	0	0	0	0	0	0	0

However, some codes that are specified with all 5 digits, are not very specific. Compare for example the following codes:

ICD9	AIS	description
800.00	2	Fracture of vault of skull Closed without mention of intracranial injury unspecified state of consciousness).
800.04	5	Fracture of vault of skull Closed without mention of intracranial injury with prolonged [more than 24 hours] loss of consciousness and return to pre-existing conscious level)
800.30	4	Fracture of vault of skull Closed with other and unspecified intracranial hemorrhage unspecified state of consciousness)

It is relevant for the derived AIS score that the coded ICD9 code is sufficiently specific. Codes ending at '00' may show a less severe injury than a fully specified injury. Not in all cases it is as dramatic as in the example above, however its influence on derived severities should be considered as an explaining factor when using this kind of data and comparing across different hospitals, databases or countries.

In the table below, the structure of the injury code is specified with its occurrence in the data. The development over time is small; the four-digit codes ending at "0" decreased from 14% to 10% and the 5 digit code ending on "0x" increased from 14% to 18%. In total for about 30% of the injury codes a more specific ICD-code exists with probably a more accurate severity score.

Table 75: Number of injuries by structure of the injury code. LMR1984-2005, filtered to patients with E-codes E800-E829.

Digits	Structure	Number of codes	Distribution
3	xxx	8.873	1,2%
4	xxx.x	245.016	32,3%
	xxx.0	91.530	12,1%
5	xxx.xx	133.456	17,6%
	xxx.0x	124.273	16,4%
	xxx.x0	68.035	9,0%
	xxx.00	87.550	11,5%
	Total	758.733	100%

Apart from coding the injuries at an aggregated –non specific– level, the hospitals that provide data to road safety research may truncate the codes or only provide the main diagnose. In order to estimate the influence of this on the derived severity (MAIS or ISS) this is simulated further in this section.

B1 Influence on MAIS

In this section we simulate the influence of omitting detail of the ICD9 injury code and omitting subdiagnoses. We analysed the 2005 discharge database, that contains 31.176 records for E-codes E800-E829+E928+E958+E988.

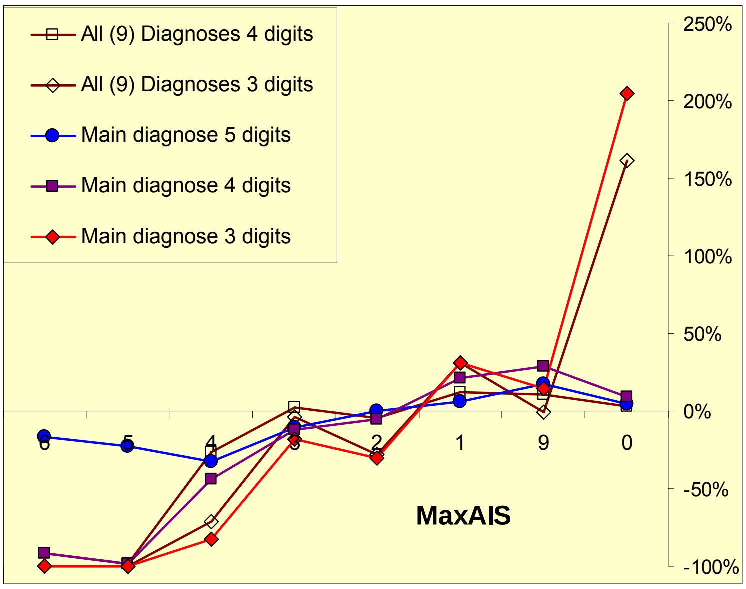
The most correct derivation of MAIS scores is of course found when using all injuries available in the most detailed coding (5 digits).

Table 76: Effects of truncation and limitation of injury codes on MAIS distribution. LMR2005, N=31.176.

%	Truncation			Main diagnosis only		
MAIS	All (9) diagnoses 5 digits	All (9) diagnoses 4 digits	All (9) diagnoses 3 digits	Main diagnose 5 digits	Main diagnose 4 digits	Main diagnose 3 digits
0 no injury	6,4	6,5	16,6	6,6	7,0	19,4
9 undetermined	5,2	5,7	5,1	6,1	6,7	5,9
1 minor	20,4	22,9	26,8	21,7	24,8	26,7
2 moderate	48,1	45,9	34,4	48,2	45,4	33,6
3 severe	17,2	17,6	16,5	15,5	15,2	14,1
4 serious	1,7	1,3	0,49	1,2	0,98	0,30
5 critical	0,94	0,02		0,72	0,02	
6 not survivable	0,04	0,00		0,03	0,00	
Sum	100%	100%	100%	100%	100%	100%

Comparing the share of patients of one MAIS group with the 'most correct' group (9 diagnoses 5 digits), we can see that the effect of truncation to 4 or even 3 digits removes all MAIS 5 or 6 cases (100% of the cases disappear). The effect of just using the main diagnosis is not too large. This is of course related to the fact that 70% of the patients only have the main diagnosis. Using only the main diagnose, truncated to 3 digits will let the number of MAIS=0 cases increase to 3 times its value, so will also the percentage (+200%, see Figure 19).

Figure 19: Effects of truncation and limitation of injury codes, differences with the most correct MAIS scores (all 9 codes, 5 digits).



B2 Injury Severity Scale (ISS)

On the same basis of AIS scores, the ISS can also be calculated. The Injury Severity Score (ISS) is an anatomical scoring system that provides an overall score for patients with multiple injuries. Each injury is assigned an Abbreviated Injury Scale (AIS) score, allocated to one of six body regions (Head, Face, Chest, Abdomen, Extremities (including Pelvis), External). Only the highest AIS score in each body region is used. The 3 most severely injured body regions have their score squared and added together to produce the ISS score.
The ISS is comparable to the New Injury Severity Score (NISS), in which the limitation to body region is cancelled; just the three most severe injuries are taken.

An example of the ISS calculation is shown below:

Table 77: Calculation of ISS from AIS scores. Example from www.trauma.org

Body Region	Injury Description	AIS	Square Top Three
Head & Neck	Cerebral Contusion	3	9
Face	No Injury	0	
Chest	Flail Chest	4	16
Abdomen	Minor Contusion of Liver Complex Rupture Spleen	2 5	25
Extremity	Fractured femur	3	
External	No Injury	0	
	Injury Severity Score		50

The ISS takes values from 0 to 75. If an injury is assigned an AIS of 6 (not survivable injury), the ISS score is automatically assigned to 75.

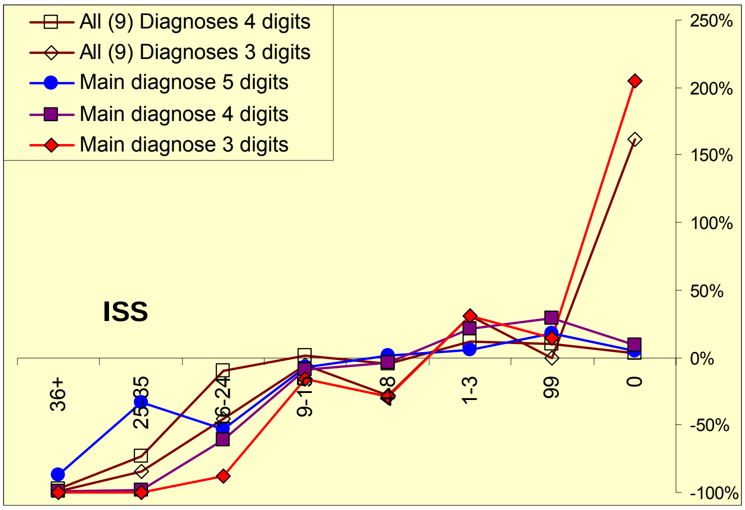
We analysed the same data by ISS, resulting in the following shifts over the ISS groups. For presentation we grouped ISS scores together at square values.

Table 78: Effects of truncation and limitation of injury codes on ISS distribution. LMR2005, N=31.176.

%	Truncation			Main diagnosis only		
ISS	All (9) diagnoses 5 digits	All (9) diagnoses 4 digits	All (9) diagnoses 3 digits	Main diagnose 5 digits	Main diagnose 4 digits	Main diagnose 3 digits
0	6,4%	6,5%	16,6%	6,6%	7,0%	19,4%
99	5,2%	5,7%	5,1%	6,1%	6,7%	5,9%
1-3	20,4%	22,9%	26,8%	21,7%	24,8%	26,7%
4-8	47,4%	45,3%	34,1%	48,2%	45,4%	33,6%
9-15	16,7%	17,0%	15,8%	15,5%	15,2%	14,1%
16-24	2,5%	2,3%	1,4%	1,2%	1,0%	0,3%
25-35	1,1%	0,3%	0,2%	0,7%	0,0%	
36+	0,2%	0,0%	0,0%	0,0%	0,0%	

Using main diagnosis only or using truncated values for the Injury codes, results in shifts to lower injury severities.

Figure 20: Effects of truncation and limitation of injury codes, differences with the most correct ISS scores (all 9 codes, 5 digits).



Comparing the share of patients of one ISS group with the 'most correct' group (9 diagnoses 5 digits), we can see that the effect of truncation to 4 or even 3 digits removes all ISS >= 25 cases (100% of the cases disappear). The effect of just using the main diagnosis is smaller, compared to truncation. This is of course related to the fact that 70% of the patients only have the main diagnosis. Using only the main diagnose, truncated to 3 digits will let the number of ISS=0 cases increase to 3 times its value, so will also the percentage (+200%, see Figure 20).

C Severity of fatalities, compared to survivors

It is known that not all fatalities have untreatable or critical injury (MAIS is 5 or 6). The distribution over the different MAIS-scores in the Dutch hospital file may be relevant for comparison with other databases; therefore a short analysis will be given here.

All casualties in the hospital file (1997-2005) having a traffic accident E-code (E810-E829) were selected.

Table 79: Fatalities by MAIS, casualties and percentage of fatalities with respect to all casualties per MAIS severity. LMR traffic 1997-2005.

	Fatalities		Fatality distribution by MAIS			
MAIS	within 30 days	after 30 days	within 30 days	after 30 days	Not killed	% fatalities
0	124	7	4,7%	3,0%	8.340	1,5%
9	74	2	2,8%	0,9%	3.714	2,0%
1	60	6	2,3%	2,6%	26.321	0,3%
2	196	33	7,4%	14%	92.570	0,2%
3	941	103	35%	44%	39.433	2,6%
4	454	33	17%	14%	3.418	12,5%
5	762	48	29%	20%	1.924	29,6%
6	50	3	1,9%	1,3%	41	56,4%
SUM	2661	235	100%	100%	175.761	1,6%

The injury of fatalities that die after 30 days is a little lower than those dying within 30 days. The percentage of casualties dying increases with severity, as was expected.

From a split by age, we see that the more severe MAIS codes can be found with younger casualties, where some elderly with MAIS1 and MAIS2 also die. The majority of elderly fatalities has a MAIS equal 3.

Project co-financed by the European Commission, Directorate-General Transport and Energy

Figure 21: Fatalities by age group and MAIS. LMR traffic 1997-2005.

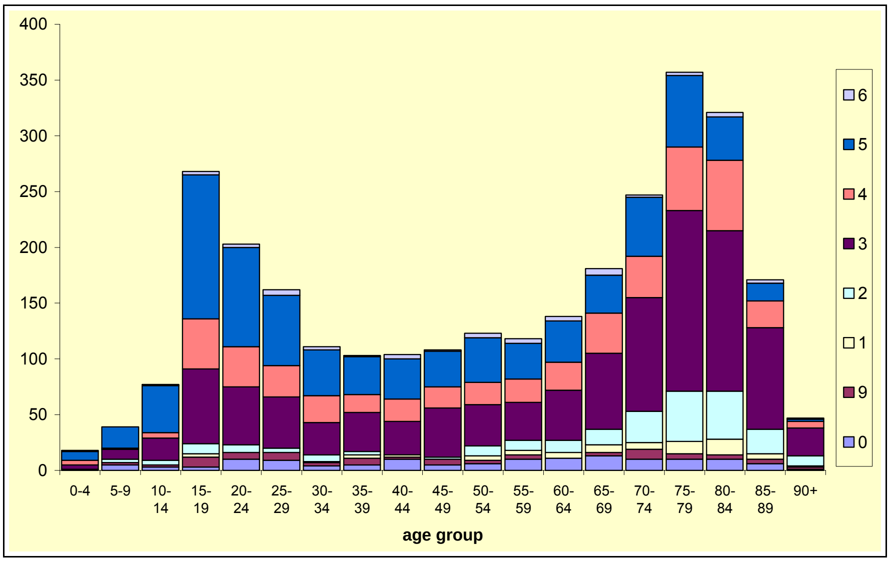
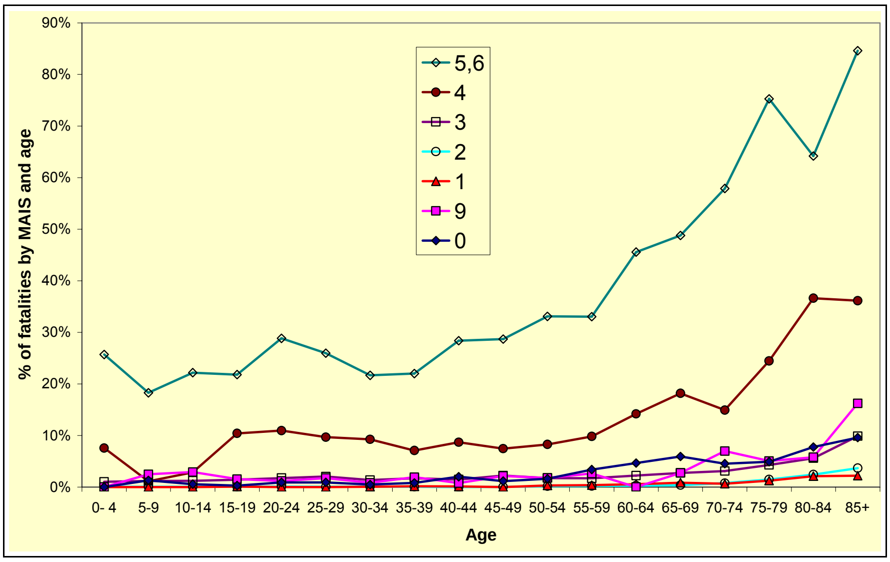


Figure 22: Morbidity (number of fatalities/total number of casualties) by MAIS score and age group. LMR traffic 1997-2005.



From Figure 22 it is obvious that traffic casualties have more chances to survive the younger they are. It is remarkable that the percentage of survivors is so large for MAIS3 cases, but this may be caused by the huge numbers of survivors, rather than by a small number of fatalities.

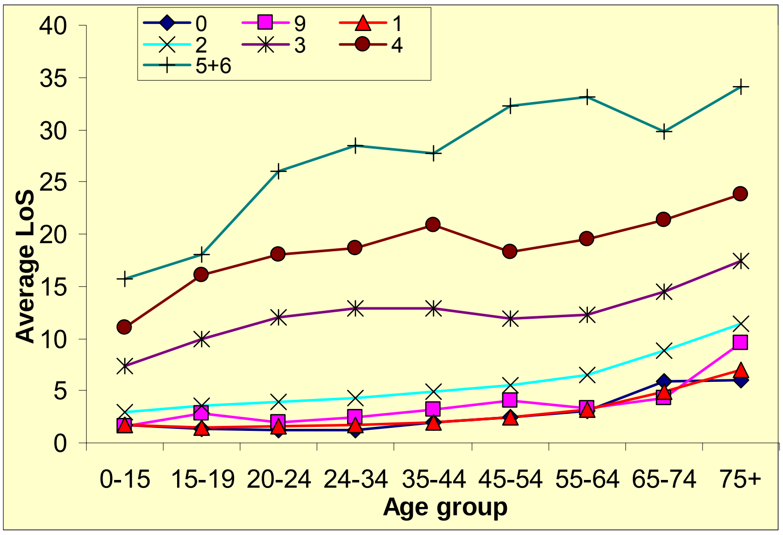
Project co-financed by the European Commission, Directorate-General Transport and Energy

D Relation between MAIS and LoS

For some casualties there is no information on their injury. However the length of hospital stay is known. In this study different measures for severity are assessed, so it is relevant to study the cases in which one of the measures cannot be determined. To be more specific: if the MAIS score cannot be determined (MAIS=0 or MAIS=9), the length of stay can give information on the severity.

In the figure below, the relation between the Maximum AIS score and the average Length of Stay is presented, for 9 age groups. LMR data for 1997-2005 have been used, omitting fatalities (at any LoS).

Figure 23: Average Length of Stay (LoS) by age for different MAIS-levels. LMR1997-2005, N=155.416.



The average length of stay for injuries that cannot be classified by ICDmap90 (MAIS9) is a little longer than MAIS1, but shorter than MAIS2. Lack of injuries in the range 800-959 (MAIS0) have the same length of stay as MAIS1 patients. This confirms that ICDmap90 is capable to score the most severe injuries and that no vital information is neglected by leaving out MAIS=0 and MAIS=9.

7.6.2.3 Variables used for the linking process

As the databases do not contain unique identifiers, casualties and patients need to be linked by other variables. The following variables where used as key variables:

Table 80: Key variables in the hospital and police database.

Common name	Hospital discharge database		Police database	
	Explanation	Variable	Explanation	Variable
Date	Date / hour when entering the hospital (no unknown in the database)	LMRepoch	Date/hour/minute when the accident happens (no unknown in the database)	VORepoch
Gender	Gender (no unknown in the database)	LMR_gender	Gender (7.627 unknowns in the database = 2,3%)	P_gender
Accident type	Type of the accident Ecode (ICD9)	E_code		-
Severity		-	Police indication of severity	P_severity
Birth	Date of birth in the hospital file	LMR_birth	Date of birth in the police file (24 records with unknown in the database)	P_birthdate
Region	Province where hospital is located. (no unknown in the database)	Prov_zh	Province where hospital is located according to the police (for hospital treated or Accident & Emergency treated persons) or unknown in cases of very slight injury (casualty not transferred to hospital or died on the spot)	P_Prov_zh

As mentioned above, all records of injured or killed persons from the police database were used. From the hospital discharge database only records of patients involved in an accident (road accident, suicide or unspecified accident) were used.

All key variables were prepared to use comparable values. The databases were arranged in ascending order of the variable date. To this end we presume that a very high percentage of all relevant road accident casualties are in a hospital within 4 days after the accident. Linking is done for all pairs within the timeframe of -1 to +4 days of the hospital admittance.

7.6.3 Description of linking process

In a procedure very similar to the one performed by Polak (1997, 2000, SWOV (2001)), it was attempted to match the police database with the hospital database.

The police database contains 324.717 road traffic accidents (average 127 per day), whereas the hospital database contains 200.766 hospitalisations (average 79 per day). Linking was carried out separately for each year. By simply joining every record of the police database with every record of the hospital database that meets the condition that the date difference must be in the range -1 .. +4 days. This means entry into hospital may be 1 day in advance of the accident or ultimately 4 days after. As a result, approximately 18 million links are assessed for each year (=365 * 127 * 5*79). Of course it is not possible to be medically treated before the accident actually happened, but this was done to enable linking of records in which recording problems have occurred.

The similarity of linked records of both databases is calculated by the distance function which is explained below. The quality of the link is calculated by the selectivity function. The selectivity points out if one can be very sure that this link is unique.

We speak about LINKING if we compare two records, one from each database. When we examine the pairs that are best linked (they have a lower distance than any other pair) we speak about MATCHING. By linking and matching we determined the intersection of both databases. By adding relevant estimates of the other cells of Table 68, an estimate can be made for the total number of hospitalised casualties.

7.6.3.1 The distance function

Since no personal ID-number, which could serve as a primary key, is recorded in Dutch police and hospital reports, a set of other characteristics has to be used for matching the respective records. These key variables (see Table 80) were used in the distance function. The values of the key variables are sometimes not correctly registered or they are missing. To quantify the similarity between two records of the police and hospital database, a generalised distance function has been defined. A very low distance close to zero indicates a very high probability that the person in the police database is the same person as the one recorded in the hospital database. If the distance is larger (because the key values are different in some variables) the probability that this linked pair refers to the same casualty is smaller.

The distance function can be described as follows:

> Let the hospital database contain N_1 records and the police database contain N_2 records, and let c_{ik} denote the value of record i on key variable k (k = 1,…., 6, we compare records on 6 key-variables), then the distance A_{ij}

between record i (i=1, …, N_1) in LMR and record j (j = 1,…, N_2) in VOR is defined as:

$$A_{ij} = \sum_{k=1}^{6} a_{ijk} = \sum_{k=1}^{6} \phi_k \ \delta(c_{ik}, c_{jk}). \qquad (1)$$

Very generally, in the term $a_{ijk} = \phi_k \ \delta(c_{ik}, c_{jk})$

$$\delta(c_{ik}, c_{jk}) = \left\{ \begin{matrix} 0 \ \text{ if } c_{ik} = c_{jk} \\ 1 \ \text{ if } c_{ik} \neq c_{ik} \\ \text{intermediate if } c_{ik} \text{ and/or } c_{jk} \text{ is missing} \end{matrix} \right\} \qquad (2)$$

and ϕ_k a weight factor for variable k. Although the values of c_{ik} and c_{jk} in (2) are defined for each key variable, they all have in common that they increase the distance between two records when the records contain unequal categories and/or missing information on a key variable.

The values of $\phi_k \ \delta(c_{ik}, c_{jk})$ for the following key variables were determined on the assumption that a distance of 100 corresponds to a probability of about 50% that two records refer to the same person.

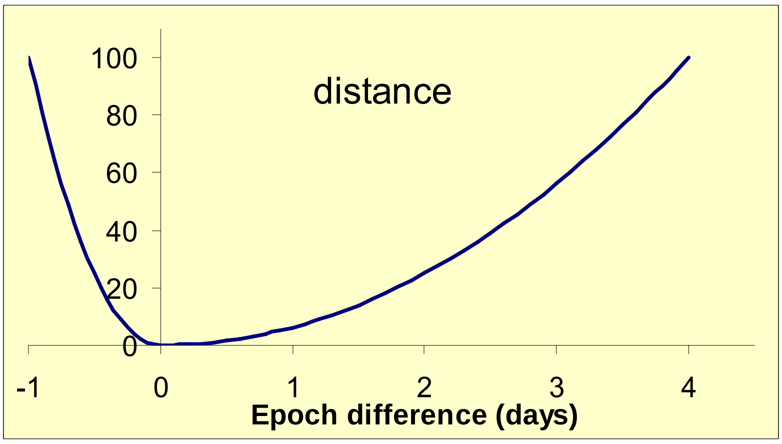
We now continue to list the distances assigned to the key variables from Table 80.

1. Epoch-difference (the difference between accident and hospital entry (date/ time)

a_{ij} = 100 * $(\alpha_i - \beta_j)^2/16$ if $\alpha_i \geq \beta_j$;

a_{ij} = 100 * $(\alpha_i - \beta_j)^2$ if $\alpha_i < \beta_j$;

Figure 24: Dependence of distance to the difference in time.



In which α_i is the epoch of hospital entry and β_i the epoch of the accident, both expressed in days. This distance is constructed in such a way that it equals 100 for a time difference of -1 and +4 days.

2. Date of birth

a_{ij} = 220 * 0 = 0 if all 8 digits are equal;
a_{ij} = 220 * 0,2 = 44 if all digits but one are equal;
a_{ij} = 220 * 0,5 = 110 if all digits but two are equal;
a_{ij} = 220 * 0,45 = 99 if the date of birth is unknown in one of the records;
a_{ij} = 220 * 1 =220 if the dates differ on more that 2 digits.

3. Gender

a_{ij} = 90 * 0 = 0 if known and equal;
a_{ij} = 90 * 0,5 = 45 if one of both is unknown;
a_{ij} = 90 * 1 = 90 if different.

4. Region of hospital

a_{ij} = 50 * 0 = 0 if the provinces are equal;
a_{ij} = 50 * 1 = 50 if the provinces are not equal;
a_{ij} = 50 * 1 = 50 if unknown in which province the hospital is, or if no hospital admittance in the police file;

5. Accident type (E-code), only in hospital file

a_{ij} = 100 * 0,9 = 90 if E-code equals 817.*, 828.*, 958.* of 988.*;
a_{ij} = 100 * 0,5 = 50 if E-code equals 820.* to 825.*;
a_{ij} = 100 * 0,55 = 55 if E-code equals 928.9*;
a_{ij} = 100 * 0 = 0 in all other cases: 810-816, 818, 819, 826, 827, 829

6. P_severity, only in police file, see coding from Table 69.

a_{ij} = 50 * 0 = 0 if P_severity equals 0, 2, 3, 4, 5, 6, 9 or 10;
a_{ij} = 50 * 0,7 = 35 if P_severity equals 1 or 8;
a_{ij} = 50 * 1 = 50 if P_severity equals 7.

At first it looks strange to give a distance of 0 if P_severity is 0, 9 or 10, while these values indicate that the casualty was not hospitalised. However in these cases there is no hospital, nor a province of the hospital, known which leads to a distance of 50 on variable 4.

If a pair of records has a difference on more than one variable, the distances are added (see Equation (1)).

For each pair of records that has been formed by joining the two databases, the distance is calculated. Small distances correspond to similar and matching pairs. Each record in one database forms a number of pairs with records from the other database. The records from the other database are called neighbours here.

Table 81: Example distances from a police record to two hospital records.

				Distance	Distance
key	police	hosp1	hosp2	P-h1	P-h2
date	22-1-2002	23-1-2002	23-1-2002		
hour	23	2	3		
minute	35				
Epoch	37278,98	37279,08	37279,13	0,06	0,13
Birth	**23-3-1980**	23-4-1980	**23-3-1980**	44	0
Gender	**Male**	**male**	**male**	0	0
Region	**5**	**5**	6	0	50
Ecode		812	813	0	0
Severity	6=hospitalised			0	0
Distance				44,06	50,13

The preferred pair in this example is the one with the correct region, but an acceptable typing error in the date of birth.

The goal in the remaining part of this section is to find the best fitting neighbours (matching) and to tell something about the quality of the match (a combination of distance and selectivity), compared to other neighbours.

7.6.3.2 The selectivity function

By calculating the distance for two records in the two databases the similarity of the records can be quantified. If e.g. a record in the police database finds a very similar record (small distance) in the hospital database, they probably belong to the same person. However, it is also important to check if there are other hospital records which also have a small distance to the initial record in the police database. If others are present, you do not know which one is the true match and the uniqueness of the initial pair can be criticised.

The selectivity of a matched pair is the minimum of the differences in distance to its next best neighbour. If the distance to the next best neighbour is large, the selectivity is high, also the uniqueness is perfect. If the selectivity is low it is very unsure which pair refers to the same person. This is the case with twins having an accident together. However from our anonymised view we do not know if they are really twins, only that in the administrations of police and hospital records are very similar.

7.6.3.3 The matching procedure

The linking and matching procedure was programmed in SAS. In this section the method of how to find matches between the hospital database and the police database is described briefly. For matches there must be a high probability that they refer to the same person in each database.

First the distances are calculated for each of the 18 million pairs that are joined annually.

Second, the best neighbour and the next best neighbour are determined by using the distance function. This is done twice, once starting from the police records and once starting from the hospital records. Starting from a police record, the smallest distance to any of the hospital records is determined (within the timeframe -4 to +1 days) , as well as the record number, the same for the smallest distance but one.

Equal distance

If a record has a certain distance to its best neighbour and has the same distance to its next best neighbour, it is uncertain which record must be taken for the match.

The distance that is assigned for the epoch difference between the VOR and LMR record is a number with a lot of decimals. As the accident time is very unlikely to be exactly the same this problem mainly exists with two records in the hospital file (only hour is available) that link to one record in the police file. In these cases the first record is taken. The selectivity will be zero (as the difference in distance to the next best neighbour is zero) and the quality of the match is poor. There is only a small number of these equal distances.

Matching

In the third step the real matching is performed. The records meeting the following conditions will be assigned to each other, which means that they belong to the same person:

1. if the record in the first database points to its best neighbour in the other database and this record also points back to the record in the first database as its best neighbour.

After the complete database has been matched by the rule above, the remainder of the database continues with:

2. if the record in the first database points to its best neighbour in the other database and this record also points back to the record in the first database as its next best neighbour.

The remainder continues with:

3. if the record in the first database points to its next best neighbour in the other database and this record also points back to the record in the first database as its best neighbour.

4. if the record in the first database points to its next best neighbour in the other database and this record also points back to the record in the first database as its next best neighbour.

Now many records from one database have been matched with records from the other database. Records that have not been selected in the four rules above belong to the rest-files 'police not hospital' and 'hospital not police' respectively.

7.6.3.4 Selectivity and quality of a match

After the matching has been done as described above, the selectivity of each matched pair is determined.

First the difference between the distances of a record to its best neighbour and to its next best neighbour are calculated. This is done for records of both databases. The lowest difference of a matched pair is then stored as their selectivity in both records of the matched pair.

Not all matched pairs really belong to the same person. Matches at a large distance do differ much on the key variables, but there seems to be no other closer record. This may be caused by underreporting of road accidents by the police. In order to separate true matches from weak matches, a combined criterion of distance and selectivity was used.

Table 82 summarises the relation between distance and selectivity for all matched records in the Dutch police and hospital databases of 1997-2003. Of the 200.000 records in the hospital database, 112.000 were matched with records in the police database, the latter consisting of 327.000 records.

To simplify inspection of the results, in Table 82 the observed distances for the matched records have been divided into seven classes ranging from very similar (distance class 0-0.1) to very dissimilar (distance class 220+). For the same reason, the observed selectivity values have been classified into five classes ranging from very low (selectivity class 0-10) to very high (selectivity class 130+).

Table 82: Frequencies of distance and selectivity classes for matched records in Dutch police and hospital databases (1997-2003, excluding fatalities and day treatment).

		selectivity class					
		0-10	10-30	30-80	80-130	130+	Total
distance class	0-0.1	244	47	1.306	13.467	17.956	33.020
	0.1-35	64	26	373	3.118	4.312	7.893
	35-55	349	147	5.510	9.850	396	16.252
	55-100	1.909	1.094	5.329	2.547	581	11.460
	100-160	7.356	5.033	4.851	502	8	17.750
	160-220	15.295	5.570	1.555	3	0	22.423
	220+	1.198	835	153	2	0	2.188
	Total	26.415	12.752	19.077	29.489	23.253	110.986

As can be seen in Table 82, of all the distance classes the one containing the smallest distances (0-0.1) has the highest frequency (33,020 records). Moreover, 95% of the matched record pairs in this distance class have a selectivity of 80 or more. In the second distance class (0.1-35), the selectivity is larger than 80 in 94% of the cases. On the whole, larger distance classes are associated with lower selectivity classes. Almost all matched record pairs with distances larger than 100 have selectivity values smaller than 80.

It may safely be assumed that many of the matched pairs in Table 82 do not actually refer to one and the same casualty. Incorrectly matched records are certain to arise as a result of random matching. In the ideal situation where police and hospital records would contain no coding errors or missing values, correctly matched records could be differentiated (almost) perfectly from incorrectly matched records, just by evaluating their distances and selectivity values. Correctly matched records would all be characterised by near zero distances combined with large selectivity values, while incorrectly matched records would all be associated with large distances combined with small selectivity values. In this ideal situation, only the upper right and lower left cells of Table 82 would contain nonzero frequencies (representing the correctly and incorrectly matched record, respectively), while the frequencies in the upper left and lower right cells would all be equal to zero. Incorrect matches would for example arise when patients from other than traffic accidents would match to slightly injured casualties from the police database that did not attend hospital.

As Table 82 shows, in reality the distinction is not so clear-cut, although the larger nonzero frequencies in Table 82 do tend to be concentrated in the upper right and lower left cells of the table. Therefore, in the next section a method is presented to differentiate the correctly matched records in Table 82 from the incorrectly matched records.

**Table 83: Matching quality status of cells in Table 82
1 (high quality) up to 6 (uncertain quality).**

		selectivity class				
		0-10	10-30	30-80	80-130	130+
distance class	0-0.1	6	6	1	1	1
	0.1-35	6	6	2	2	2
	35-55	6	6	3	3	3
	55-100	6	6	4	4	4
	100-160	6	6	5	5	5
	160-220	6	6	6	6	6
	220+	6	6	6	6	6

In this method, matched records are assigned to one of six matching quality classes. The matching quality classes are defined in Table 83. Records of matching quality class 1 have a high probability of having been correctly matched. At the same time, the correctness of the matching of records assigned to class 6 ranges from uncertain to improbable. In the next section, this classification of matched records is used to obtain estimates of the number of correctly matched records, and thus of the frequency in the first cell of Table 68 "in both databases" (intersection of police and hospital file). Matches of poor quality, as well as records that are not matched at all, will appear in the cells "Only in hospital database" and "Only in police database" respectively.

In the practical application we used matches of quality 1, 2 and 3 as part of the intersection of both databases, whereas we omitted matches with quality 4 5 and 6. An exception was made for matches between police records and hospital records that had an Ecode of a non-traffic accident, so quality 4 is included for Ecodes 817+ 828+ 928+ 958+ 988). The underlying idea is that if the police say it is a traffic accident this information is assumed to be true, so if the hospital file says that it was a suicide attempt, the hospital information is overruled by the police information. This split of the quality 4 matches is indicated by quality 4*.

This brings us to a total of 63.779 matches (57% of all matches). The pour quality matches will be added to the rest-files 'police not hospital' and 'hospital not police' respectively.

7.6.3.5 Plausibility check of the matched records

As no set of records with proofed quality of linking is available, the linking procedure can not be validated perfectly. To see if the linking procedure is calculating reasonable results, an alternative way of checking the plausibility of the matched records was chosen. Although available in both databases, the mode of transport is not one of the key variables. This gives the opportunity to see if the mode of transport corresponds within a matched pair.

Table 84: Distribution of modes of transport in matches of quality 1-3+4* for accidents involving motorised vehicles. 89% of the cases have the same mode (omitting unspecified modes of transport).

1997-2003 Mode in VOR	Mode hospital file LMR								Total
	Pedes trian	Bicycle	Moped	Motor cycle	Car	Lorry/ truck	Other	Not specified	
Pedestrian	3.058	236	31	8	105	89	11	197	3.735
Bicycle	913	7.156	117	22	312	38	7	449	9.014
Moped	150	350	8.543	581	187	8	280	439	10.538
Motorcycle	31	12	236	3.884	75	1	10	166	4.415
Car	624	226	91	68	20.820	142	19	2.063	24.053
Lorry/truck	23	5	1	1	100	242	2	33	407
Other	10	9	8	5	46	2	36	17	133
Total	4.809	7.994	9.027	4.569	21.645	522	365	3.364	52.295

Table 85: Distribution of modes of transport in matches of quality 1-3+4* for accidents *not* involving motorised vehicles. 87% of the cases have the same mode (omitting unspecified modes of transport).

1997-2003 Mode in VOR	Mode hospital file LMR								Total
	Pedes trian	Bicycle	Moped	Motor cycle	Car	Lorry/ truck	Other	Not specified	
Pedestrian	209	52	1	0	0	0	4	8	274
Bicycle	53	3.097	6	0	6	0	4	33	3.199
Moped	7	292	100	0	3	0	5	8	415
Motorcycle	0	6	0	2	0	0	0	0	8
Car	3	51	2	0	8	0	2	2	68
Lorry/truck	0	2	0	0	0	2	0	0	4
Other	1	5	0	0	0	1	11	7	25
Total	273	3.505	109	2	17	3	26	58	3.993

Many of the off-diagonal numbers can be explained by the confusion of transport mode in cases of bicyclists walking with the bike or drivers/passengers of motor vehicles being hit when getting off or standing next to their vehicle. If in the hospital file the mode is unspecified, we see that the numbers follow the distribution of the specified modes.

When we compare these results for the poor quality matches (4*+5+6), we observe that these matrices are much more randomly distributed than the high quality matrices.

Table 86: Distribution of modes of transport in matches of quality 4*+5+6 for accidents involving motorised vehicles. 41% of the cases have the same mode (omitting unspecified modes of transport).

1997-2003 Mode in VOR	Mode hospital file LMR								Total
	Pedes trian	Bicycle	Moped	Motor cycle	Car	Lorry/ truck	Other	Not specified	
Pedestrian	395	104	118	67	199	22	17	95	1.017
Bicycle	407	810	770	367	894	56	57	269	3.630
Moped	309	346	1.829	438	877	38	125	307	4.269
Motorcycle	79	49	167	400	250	21	17	70	1.053
Car	726	646	1.453	963	3.654	166	160	685	8.453
Lorry/truck	16	13	40	32	65	20	6	24	216
Other	7	11	21	15	34	2	3	6	99
Total	1.939	1.979	4.398	2.282	5.973	325	385	1.456	18.737

Table 87: Distribution of modes of transport in matches of quality 4*+5+6 for accidents *not* involving motorised vehicles. 26% of the cases have the same mode (omitting unspecified modes of transport).

1997-2003 Mode in VOR	Mode hospital file LMR								Total
	Pedestrian	Bicycle	Moped	Motor cycle	Car	Lorry/truck	Other	Not specified	
Pedestrian	53	815	8	1	1	0	8	7	893
Bicycle	89	4.182	20	1	14	4	32	47	4.389
Moped	68	2.949	37	2	12	2	24	29	3.123
Motorcycle	8	672	5	1	3	1	6	13	709
Car	140	7.071	21	2	33	4	55	72	7.398
Lorry/truck	6	195	0	0	1	0	3	1	206
Other	3	73	1	0	0	0	1	0	78
Total	367	15.957	92	7	64	11	129	169	16.796

This confirms that matches of quality 1-4* can be considered as correct.

Other options for a quality check, such as

- linking the police records from one year to the hospital records from a different year (this should only give a small number of quality matches),
- linking with other techniques,

were not yet performed.

7.6.4 Results and Discussion

In the previous sections the intersection between the two databases was determined. Now we return to Table 68 in order to fill the upper left cell. Additional details in this table will be added, as can be seen in Table 88.

The databases contain details of road traffic fatalities and casualties (severe as well as minor injuries). As fatalities and minor injury records do by definition not belong to the 'hospitalised road traffic casualties' their number is only included in Table 88 (**labelled NotHRTC**) and excluded in almost all other tables and further analyses.

Casualties in the hospital database were not always hospitalised: some of them received day-treatment, whereas the majority stayed overnight or more nights. The ones that stayed one or more nights are hospitalised according to the Dutch definition. However, if we want to compare internationally, we need to explore common definitions of a ‘severe road traffic casualty’. Different types and criteria will be addressed here, based on the Length of Stay and the Maximum AIS.

We will split the numbers NotHRTC apart from our basic Table 68, in order to omit these in the remainder of the report. This provides five groups of casualty records:

1. police and hospital – casualties in police and hospital records;
2. hospital, not police – casualties in hospital records but not in police records;
3. hospital, not a traffic accident – casualties in hospital records that have been used in the linking process, but were excluded from the label 'hospitalised road traffic casualty' because the casualty died within 30 days, received only day-treatment, was not a traffic casualty (but other or unspecified external cause);
4. police, not hospital – casualties in police records but not in hospital records;
5. police not hospitalised – casualties in police records that haven been used in the linking process but were excluded from the label 'hospitalised road traffic casualty' because the casualty died within 30 days or had only minor injury.

Groups 1, 2 and 4 also contain records that do not fulfil the definition of a hospitalised road traffic casualty. In group 1 (intersection) 2015 cases concern fatalities[6], 1532 day-treatment. Group 2 contains 513 fatalities and 2957 cases of day-treatment.

For the fourth group of records an assumption is made, while the information from the police is doubted: 4% of the total number of hospitalised casualties (79.984 according to the police) is assumed to follow the definition 'hospitalized, reported by police but not recognized in hospital'. All others (total – 4% - linked – fatalities - day treatments = 27.069 cases) are assumed not to reflect a hospitalisation. Reasons can be that they have the wrong severity assigned (not hospitalised) or that some information on key variables is incorrect in either one of the databases, so that a match is prevented. By taking the maximum number of records on the hospital database (group 2), these cases are caught without the risk of double counting.

The results of the linkage can be summarised as follows:

Table 88: Linking between 324.717 police records and 200.066 hospital records (1997-2003).

1 police and hospital (60.232 casualties + 3.547 NotHRTC)	2 hospital, not police (63.354 casualties + 3.470 NotHRTC)	3 Hospitalised NotHRTC (70.163 casualties)
4 police, not hospital (3.205 casualties + 27.069 NotHRTC)	6 neither police nor hospital (estimated 2.826)	
5 Police Not HRTC (230.664 casualties)		

The cells within the double lines form the total number of hospitalised road traffic casualties (129.617 casualties).

A 6th group of neither police nor hospital can be seen inside the double lines. With a Capture-Recapture methodology an estimate has been made of the

[6] 2015 fatalities, of which 1779 dead in both files, 114 in police file only and 122 in medical file only.

number of not reported casualties that fulfil the definition of hospitalised. We estimated the number of 'neither police nor hospital' casualties (2826) from the assumption that the following ratios are equal

$$\frac{\text{neither police nor hospital}}{\text{police not hospital}} = \frac{\text{hospital not police}}{\text{police and hospital}} \qquad (3)$$

For the real numerical application the reader is referred to Reurings et al. (2007) as some records (5437) from group 2, which were linked with pour quality, have been assigned as true links by the so-called Footprint method. The number has been split over the modes of transport by the same distribution as in the police file.

Other national studies have not included such estimates, however they take into account all police reported records where we made an assumption for 'police not hospital' (group 4, see above). We may only apply this assumption if we also correct for not reported cases.

We will see later that depending on a boundary value set by MAIS or Length of Stay only a part of these casualties will follow the new definition of Severely Injured.

The results of the linkage are summarised by road user type and casualty severity in Table 89, while the national totals from police reported casualties are shown in Table 90.

Table 89: Linkage results by road user type (1997-2003).

		police		not police	
		hospitalised	slight		Total
hospital	Car/van occupant	21.176	4.590	12.570	38.336
	Motorcyclist	3.847	831	3.266	7.944
	Moped	9.230	2.477	11.146	22.853
	Pedal cyclist	10.323	2.732	32.006	45.061
	Pedestrian	3.620	781	3.693	8.094
	Other	539	86	672	1.297
	Subtotal	48.735	11.497	63.354	123.586
not hospital	Car/van occupant	1.479		1.309	2.788
	Motorcyclist	222		198	420
	Moped	599		530	1.129
	Pedal cyclist	644		570	1.214
	Pedestrian	213		188	401
	Other	48		31	79
	Subtotal	3.205	0	2.826	6.031
Total		51.940	11.497	66.180	129.617

Table 90: National reported totals, 1997-2003, by road user type and police severity.

Road user	hospitalised	slightly injured	Total
Car/van occupant	36.972	107.339	144.311
Motorcyclist	5.539	10.595	16.134
Mopedist	14.992	50.936	65.928
Pedal cyclist	16.048	53.199	69.247
Pedestrian	5.326	11.577	16.903
Other	1.107	3.678	4.785
All	79.984	237.324	317.308

In The Netherlands we are used to express a reporting rate as 79.984/129.617 = 61,7%. This is similar to a general factor of 1,62, working on the number of hospitalised only. However in this international approach we want to establish a set of factors based on hospitalised and slight injuries as well. Furthermore we want to deselect some of the 129.617 casualties as they do not all fulfil severity conditions as Length Of Stay≥boundary$_{LOS}$ and MAIS≥boundary$_{MAIS}$.

7.6.4.1 Results for Length of Stay

First the data are analysed by the Length of Stay (LoS) in hospital. The overall results of the linkage are shown in Table 91, omitting fatalities from the police database. The proportion of casualties who were not reported by the police is lower among the more severely injured, but the difference is less than has been found in other countries. The Length of Stay is reported for all hospital records.

Table 91: The linkage results, by Length of Stay.

Note that only the marked cells are considered
as 'hospitalized road traffic casualty'

N	Length Of Stay (LoS)	police police 'severity'=		not police	Total	% not reported by police
		hospitalised	slight			
hospital	overnight	4.919	2.140	8.761	15.820	55%
	1	10.522	3.802	15.649	29.973	52%
	2	5.128	1.318	6.753	13.199	51%
	3	3.402	733	4.157	8.292	50%
	≥4	24.764	3.504	28.033	56.301	50%
	Fatalities within 30 days	118	4		122	
	Day treatment	857	675	2.956	4.488	
not hospital	Police not in hospital file	3.205		2.826	6.031	
	Not hospitalised	27.069	225.148		252.217	
Total		79.984	237.324	69.136	386.444	

Project co-financed by the European Commission, Directorate-General Transport and Energy

In order to create a table in which all casualties are distributed over Length of Stay and Police severity, we need to redistribute two groups of casualties over unknown properties.

First we distribute 'Hospital, not police' over the severities that the police most probable would have assigned. We assume that for a casualty with a certain LoS the same proportion would be labelled hospitalised as the records that could be matched. From 8761 overnight casualties in the hospital database we estimate that a proportion of 4919/(4919+2140)=70% would be judged hospitalised by the police (6105 cases), and so on.

Secondly, for the ‘police, not hospital’ casualties (3205) and for the casualties reported 'in neither database' (2826), no Length of Stay information is available. As we excluded most not linked police casualties from our results, we assume that these 6031 casualties have the same distribution as any other hospitalised casualty, so we assume that a proportion of 15820/(15820+29973+13199+8292+56301)=13% stays overnight (772 cases). So the total number of casualties, rated hospitalised by the police that stays overnight is 4919+6105+772=11796. Table 92 presents the results of both redistributions.

Table 92: Estimated results, by Length of Stay.

	Casualties by police severity			Factors		Cumulative factors	
Length of Stay	hospitalised	slight	Total	hospitalised	slight	hospitalised	slight
Overnight	11.796	4.796	16.592	0,147	0,020	1,322	0,101
1	23.480	7.956	31.436	0,294	0,034	1,174	0,080
2	11.144	2.699	13.843	0,139	0,011	0,881	0,047
3	7.227	1.470	8.697	0,090	0,006	0,741	0,036
≥4	52.070	6.979	59.049	0,651	0,029	0,651	0,029
Total	105.717	23.899	129.617	1,322	0,101		

The results show that, corresponding to each casualty reported as hospitalised by the police (see Table 90), 23480/79984=0.294 casualties were in hospital for 1 day, and 0.651 casualties for 4 or more days. Such conversion factors can be used to estimate casualty totals from police casualty totals, although changes over time in hospital procedures may mean that the factors depends upon the period chosen.

For example, if serious casualties were to be defined as those staying 3 or more days in hospital then the actual total could be estimated as:

N_{LoS3+} = 0,741 x number of hospitalised casualties reported by the police +

0,036 x number of slight casualties reported by the police

Matrix 1 by Length of Stay

We have now determined correction factors for police records to the real number of hospitalised casualties for different lower boundaries of LengthOfStay. Now we want to do the same per road user type (mode). Table

93 shows the data where police severity and length of stay in hospital are crossed per mode.

Table 93: matrix 1, numbers by mode, police severity and length of stay.

N			Police police "severity"=		not police	
	Mode	Length Of Stay	hospitalised	slight	unknown	SUM
Hospital	car/van	overnight	2.971	1.174	2.298	6.443
		1	5.471	1.786	3.489	10.746
		2	2.378	529	1.327	4.234
		3	1.430	236	730	2.396
		>3	8.926	865	4.727	14.518
	motorcycle	overnight	261	120	369	750
		1	586	219	825	1.630
		2	364	90	391	845
		3	307	66	240	613
		>3	2.329	336	1.442	4.107
	moped	overnight	790	384	1.565	2.739
		1	1.571	718	2.650	4.939
		2	881	270	1.282	2.433
		3	651	194	793	1.638
		>3	5.337	911	4.857	11.105
	pedal cycle	overnight	603	343	3.966	4.912
		1	2.029	815	7.618	10.462
		2	1.095	333	3.292	4.720
		3	754	186	2.124	3.064
		>3	5.842	1.055	15.005	21.902
	pedestrian	overnight	255	111	474	840
		1	696	227	899	1.822
		2	340	79	392	811
		3	223	46	231	500
		>3	2.106	318	1.697	4.121
	other	overnight	39	8	89	136
		1	169	37	169	375
		2	70	17	69	156
		3	37	5	40	82
		>3	224	19	305	548
	subtotal		48.735	11.497	63.354	123.586
not hospital	car/van	unknown	1.479		1.309	2.788
	motorcycle	unknown	222		198	420
	moped	unknown	599		530	1.129
	pedal cycle	unknown	644		570	1.214
	pedestrian	unknown	213		188	401
	other	unknown	48		31	79
	subtotal		3.205		2.826	6.031
	SUM		51.940	11.497	66.180	129.617

Once more we need to use assumptions about missing information on LoS for records within 'police, not hospital' and 'in neither database', as well as about a severity that the police would have assigned for records in 'hospital, not police'. The same assumptions are used as above and the resulting conversion factors by road user type are presented in Table 94.

Table 94: Conversion Factors based on Length of Stay and mode for different cut-off boundaries.

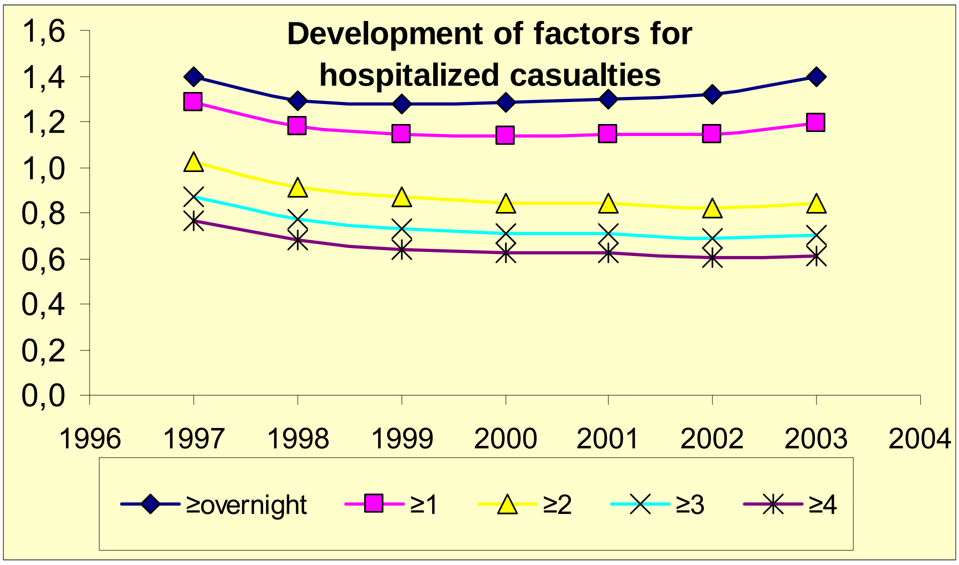
Cumulative factors	Length of Stay									
	≥overnight		≥1		≥2		≥3		≥4	
	hosp.	slight	hosp.	slight	hosp.	slight	hosp.	slight	hosp.	slight
car/van	0,918	0,067	0,784	0,050	0,549	0,024	0,448	0,017	0,388	0,014
motorcycle	1,246	0,138	1,143	0,116	0,909	0,075	0,779	0,059	0,683	0,049
moped	1,310	0,092	1,173	0,075	0,916	0,047	0,779	0,037	0,685	0,030
pedal cycle	2,331	0,167	2,112	0,138	1,618	0,084	1,379	0,065	1,218	0,055
pedestrian	1,314	0,129	1,195	0,107	0,922	0,067	0,791	0,053	0,709	0,046
other	0,857	0,052	0,775	0,043	0,544	0,024	0,445	0,016	0,390	0,014
ALL	1,322	0,101	1,174	0,080	0,881	0,047	0,741	0,036	0,651	0,029

Further in this report these factors will be applied to the CARE data, to derive the number of severely injured. The bottom line in Table 94 will be referred to as 'ALL factors'.

Time dependency of the correction factors for LoS

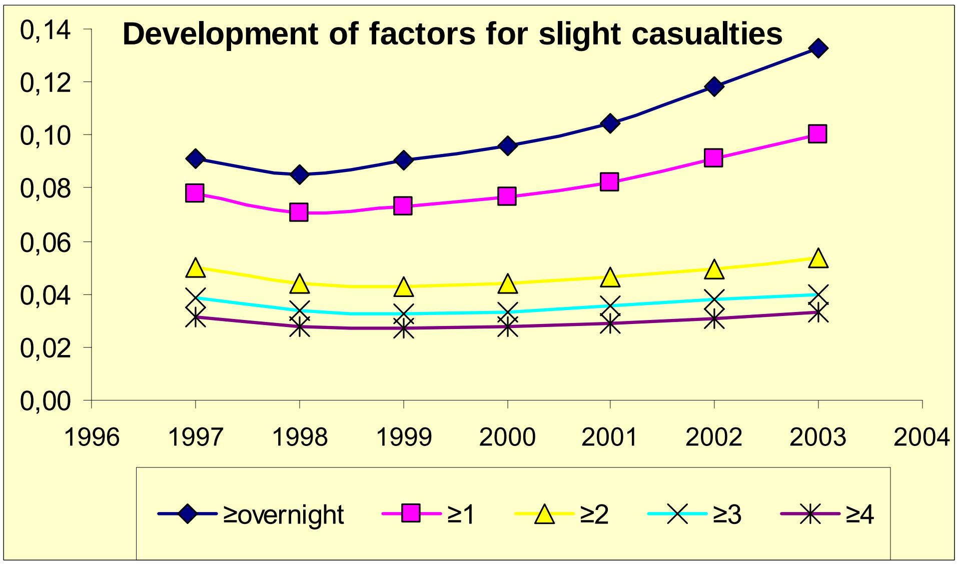
It is possible that these factors develop over time, so we split the data by year instead of mode and examined the annual factors. The basic annual data is not given here, only a graphical presentation of the resulting factors is given. We observe rather stable factors to be applied to hospitalised casualties. Factors for higher boundaries (small factors) tend to decrease.

Figure 25: Conversion factors to LoS, to be applied to hospitalised police reported casualties.



The factor to be applied to the number of slight casualties appears to increase. This may be the effect of a decreasing reporting rate for slight casualties (the number of reported slight casualties has dropped with 27% in the 7 years studied), but is also influenced by the Length of Stay. For longer hospitalizations the factor is rather stable. If we use average factors, the results for the development of both factors will probably compensate.

Figure 26: Conversion factors to LoS, to be applied to slight police reported casualties.



Project co-financed by the European Commission, Directorate-General Transport and Energy

7.6.4.2 Results for Maximum AIS

When we want to investigate the effects of a cut-off at certain MAIS level, we need to split the data as presented in Table 71 et seq not by Length of Stay, but by MAIS instead.

The MAIS scores have been assigned from the ICD injury codes for each case. The overall linkage results are shown in Table 95. The column '% not reported by police' is based on the three others, e.g. at MAIS=1 47%=8399/17846. As with the data in Table 90 that was based on Length of Stay, the proportion of casualties who were not reported by the police is lower among the more severely injured.

Table 95: The linkage results by MAIS.

N		Police police 'severity'=		not police		% not reported by police
	MAIS	hospitalised	slight	unknown	Total	
Hospital	0 (not known)	1.969	761	2.583	5.313	49%
	9 (unknown)	1.098	338	1.134	2.570	44%
	1	7.107	2.340	8.399	17.846	47%
	2	24.066	6.226	34.177	64.469	53%
	3	12.384	1.673	15.313	29.370	52%
	4	1.252	110	1.157	2.519	46%
	5	843	48	573	1.464	39%
	6	16	1	18	35	51%
	Fatality	118	4		122	
	Day treatment	857	675	2.956	4.488	
Not hospital	unknown	3.205		2.826	6.031	
	Not hospitalised	27.069	225.148		252.217	
Total		79.984	237.324	69.136	386.444	

As with Length of Stay, in order to create a table in which all casualties are distributed over MAIS and Police severity, we need to distribute the two groups casualties over unknown properties.

First we distribute 'Hospital not police' over the severities that the police most probable would have assigned. We assume that for a casualty with a certain MAIS the same proportion would be labelled hospitalised as the records that could be matched. From 8399 MAIS=1 casualties in the hospital database we estimate that a proportion of 7107/(7107+2340)=75% would be judged hospitalised by the police (6319 cases), and so on.

Secondly, for the 'police, not hospital' casualties (3205) and for the casualties reported 'in neither database' (2826), no MAIS information is available. As we excluded most not linked casualties from our results, we assume that the 6031 casualties have the same MAIS distribution as any other hospitalised casualty, So for example we assume a proportion of 17.846/123.586=14% to have MAIS=1 (871 cases). So the total number of casualties, rated hospitalised by the police with MAIS=1 is 7107+6319+871=14296. Table 96 presents the total results of this redistribution.

Table 96: Results by MAIS and police severity.

	Casualties by police severity			Factor		Cumulative factor	
	hospitalised	slight	Total	hosp.	slight	hosp.	slight
0 (not known)	4.091	1.481	5.572	0,051	0,006	1,326	0,099
9 (unknown)	2.090	605	2.695	0,026	0,003	1,275	0,093
1	14.296	4.420	18.717	0,179	0,019	1,249	0,090
2	54.364	13.250	67.615	0,680	0,056	1,070	0,072
3	27.308	3.495	30.803	0,341	0,015	0,391	0,016
4	2.439	203	2.642	0,030	0,001	0,049	0,0012
5-6	1.491	81	1.572	0,019	0,0003	0,019	0,0003
Total	106.080	23.537	129.617	1,326	0,099		

Due to small numbers the groups for MAIS=5 and MAIS=6 have been taken together.

The results show that, corresponding to each hospitalised casualty reported by the police (see Table 90), 27.308/79.984=0,341 casualties were in hospital with MAIS=3, and 0,049 with MAIS 4, 5, or 6. Such conversion factors can be used to estimate casualty totals from police casualty totals, although changes over time in hospital procedures may mean that the factors depends upon the period chosen.

For example, if serious casualties were to be defined as those having at least MAIS=2 then the actual total could be estimated as:

N_{mais2+} = 1.070 x number of hospitalised casualties reported by the police +
0.072 x number of slight casualties reported by the police

Comparing the redistributions for MAIS with LoS

From 49.633 cases there is no information for the police estimate of hospitalized or slight injury (129.617 – 79.984). Where the Length of stay approach assigned 25.729 of them to hospitalized (51,8%), the MAIS approach assigned 26.092 cases to hospitalized (52,6%). So both approaches result in a comparable estimate.

Matrix 2 by Maximum AIS

We have now determined correction factors for police records to the real number of hospitalised casualties for different lower boundaries of Maximum AIS. Now we want to do the same for the road user type (mode of transport). **Table 97** shows the data where police severity and maximum AIS are crossed per mode.

Table 97: Matrix 2, numbers by mode, police severity and MAIS

N			Police police "severity"=		Not police	
	mode	MaxAis	hospitalised	slight	unknown	SUM
Hospital	car/van	0	1.431	548	1.201	3.180
		9	674	212	417	1.303
		1	4.508	1.358	2.687	8.553
		2	9.147	2.007	5.515	16.669
		3	4.539	433	2.278	7.250
		4	509	20	294	823
		5	359	11	172	542
		6	9	1	7	17
	motorcycle	0	70	22	94	186
		9	61	13	44	118
		1	334	129	375	838
		2	2.088	515	2.026	4.629
		3	1.082	135	611	1.828
		4	113	11	68	192
		5	99	6	46	151
		6			2	2
	moped	0	176	89	350	615
		9	157	50	203	410
		1	867	378	1.352	2.597
		2	4.814	1.517	6.524	12.855
		3	2.786	404	2.368	5.558
		4	224	21	220	465
		5	205	18	126	349
		6	1		2	3
	pedal cycle	0	187	84	771	1.042
		9	146	49	398	593
		1	944	332	3.384	4.660
		2	5.687	1.653	17.679	25.019
		3	2.929	562	9.109	12.600
		4	300	42	471	813
		5	124	10	188	322
		6	6		6	12
	pedestrian	0	77	17	138	232
		9	50	12	56	118
		1	329	108	499	936
		2	2.081	499	2.083	4.663
		3	937	127	798	1.862
		4	93	15	89	197
		5	53	3	31	87
		6				0
	Other	0	28	1	29	58
		9	10	2	16	28
		1	125	35	102	262
		2	249	35	350	634
		3	111	12	149	272

		4	13	1	16	30
		5	3		10	13
		6				0
	subtotal		48.735	11.497	63.354	123.586
Not hospital	car/van	unknown	1.479		1.309	2.788
	motorcycle	unknown	222		198	420
	moped	unknown	599		530	1.129
	pedal cycle	unknown	644		570	1.214
	pedestrian	unknown	213		188	401
	other	unknown	48		31	79
	subtotal		3.205		2.826	6.031
	SUM		51.940	11.497	66.180	129.617

Again we need to use assumptions about missing information on MAIS for records within 'police, not hospital' and 'in neither database' as well as about a severity that the police would have assigned for records in 'hospital, not police'. The same assumptions are used as above and the resulting conversion factors by road user type are presented in Table 98.

Table 98: Conversion factors based on MAIS and mode for different cut-off values.

cumulative	Maximum AIS									
	≥1		≥2		≥3		≥4		≥5	
factors	hosp.	slight	hosp.	slight	hosp.	slight	hosp.	slight	hosp.	slight
car/van	0,826	0,055	0,638	0,036	0,233	0,007	0,038	0,001	0,015	0,0002
motorcycle	1,200	0,131	1,078	0,110	0,371	0,022	0,060	0,002	0,027	0,0008
moped	1,259	0,086	1,120	0,072	0,402	0,015	0,054	0,001	0,023	0,0005
pedal cycle	2,266	0,154	2,037	0,132	0,768	0,033	0,068	0,002	0,020	0,0004
pedestrian	1,260	0,122	1,117	0,102	0,377	0,021	0,051	0,002	0,016	0,0004
other	0,803	0,048	0,645	0,033	0,225	0,008	0,032	0,001	0,010	0,0001
ALL	1,249	0,090	1,070	0,072	0,391	0,016	0,049	0,001	0,019	0,0003

Further in this report these factors will be applied to the CARE data, to derive the number of severely injured. The bottom line in Table 98 will be referred to as 'ALL factors'.

The results emphasise that pedal cyclist casualties are recorded less fully by the Dutch police than casualties among other road user groups, as these factors are much higher.

Time dependency of the corrections factors for MAIS

It is possible that these factors develop over time, so we split the data by year instead of mode and examined the annual factors. The basic annual data is not given here, but a graphical presentation of the resulting factors is given.

We observe rather stable factors to be applied to hospitalised casualties. Factors for higher boundaries (small factors) are also stable, but very small.

Figure 27: Conversion factors to MAIS, to be applied to hospitalised police reported casualties.

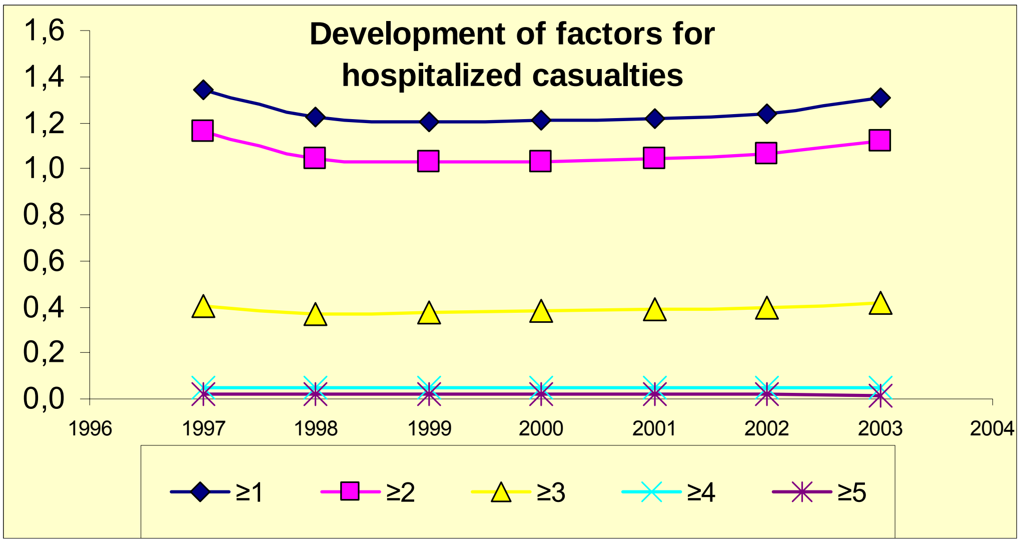
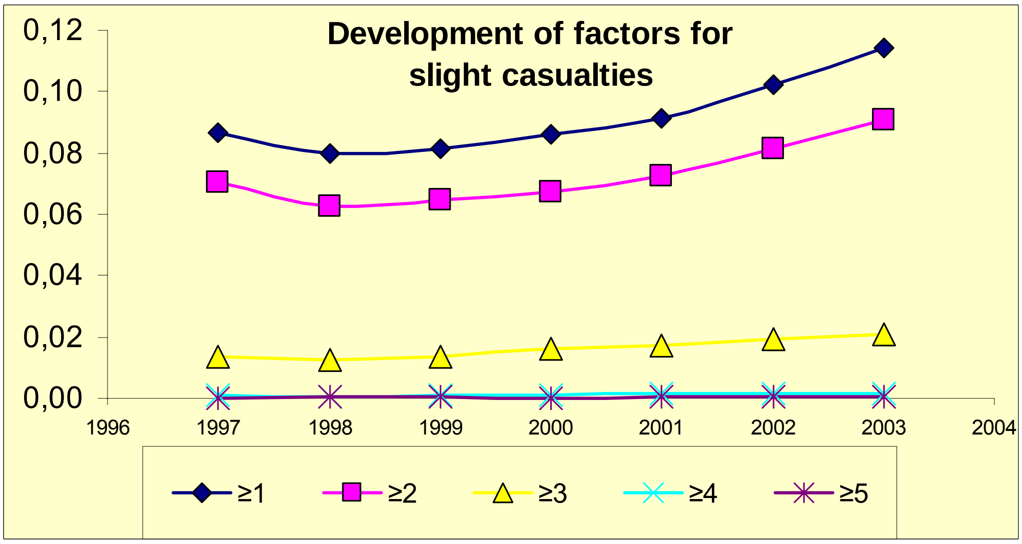


Figure 28: Conversion factors to MAIS, to be applied to slight police reported casualties.



7.6.4.3 Overview and discussion

The linking between the Dutch hospital database and the police database of traffic casualties resulted in an estimate of 18.500 hospitalised traffic casualties annually, distributed over the different cells of Table 99.

Table 99: Distribution of police and hospital records over intersection and rest files.

police and hospital 46%	hospital, not police 49%
police, not hospital 2,5%	neither police nor hospital 2,2%

When applying a certain cut-off at a lower boundary on LoS or MAIS, this distribution will not really change. So, if only the hospital file is used, 95% of the records consist of real information; only 4,7% is estimated. This means that the distribution of casualties by variables that are in the hospital database (such as LoS, MAIS, mode, age, day of week, time of day) can be determined quite accurately, close to the real distribution.

If only the police file is used, known information accounts only for less than 50%. Therefore, variables that are only in the police file (such as area type [built up area, urban area, motorway], crash opponent, intersection/junction, weather condition) can not be determined very accurately.

One of the aims of this project is to develop factors on the police file, in order to estimate the real numbers and their distributions. With several assumptions for missing information, correction factors have been developed that can be applied to the CARE database (i.e. the police file), resulting in an estimate of the real number of severely injured.

We can distinguish different groups of variables which may be expected to be more or less reliable in the estimate of the real number.

1) Variables that are present in both databases.
   Some variables have equivalents in the other database. This group can be split into 2 subgroups:
   a) variables that were used in the linking process (date/time, age, gender, severity, region),
   b) variables that were used to calculate the correction factors (severity, mode),

   For group b, of course the outcome reflects the real distribution for that variable, because the factors were developed to do so. The variable severity (police severity, MAIS or LoS) is a special one in this group, as it is also used to set a minimum on the severity, by filtering out only cases above certain lower boundary. In the analysis of results some assumptions have been made on the relations between these variables.

   Variables from 'group a' which were not used in the calculation can give us a validation of the outcomes (year, age, gender, day of week, time of day). Only if the numbers/distributions obtained in such a validation are within a reasonable similarity with the real distribution, one may expect the distributions on variables that are in one file only to be reliable. In the next sections we will compare the real distribution (from 95% hospital data) with the calculated results (police file x correction factors) for the variables year and age.

2) Variables that exist in one file only.

For variables which are available in only one database it is very difficult to say if the observed distribution is comparable to the real distribution, as we do not know this real distribution. If, for example, all accidents at intersections were reported correctly by the police, all missing information should be assigned to sections. An average correction factor as is calculated here would take the same fraction for both types and thus overestimate the number of intersection accidents and underestimate the section accidents.

The variables Year and Age are explored below.

7.6.4.4 The application to the police file 1991-2005

As an example we present the number of severe casualties by mode of transport according to different boundary levels. We used the correction factors by mode of transport, but without time dependence (see Table 94 and Table 98). We need to judge whether the results seem reliable and we compare the sum of all modes with the application of correction factors without mode of transport (see Table 92 and Table 96).

Figure 29: Car occupants by MAIS2+, MAIS3+, LoS2+, LoS3+.

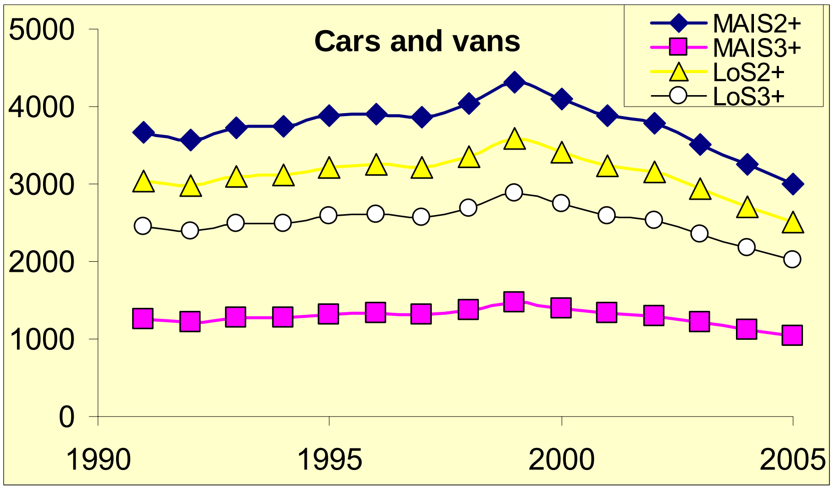


Figure 30: Motorcycles by MAIS2+, MAIS3+, LoS2+, LoS3+.

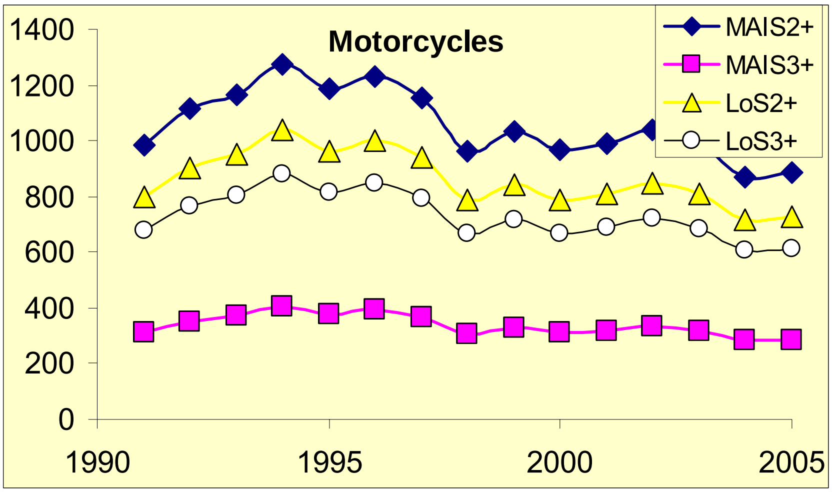


Figure 31: Mopeds by MAIS2+, MAIS3+, LoS2+, LoS3+.

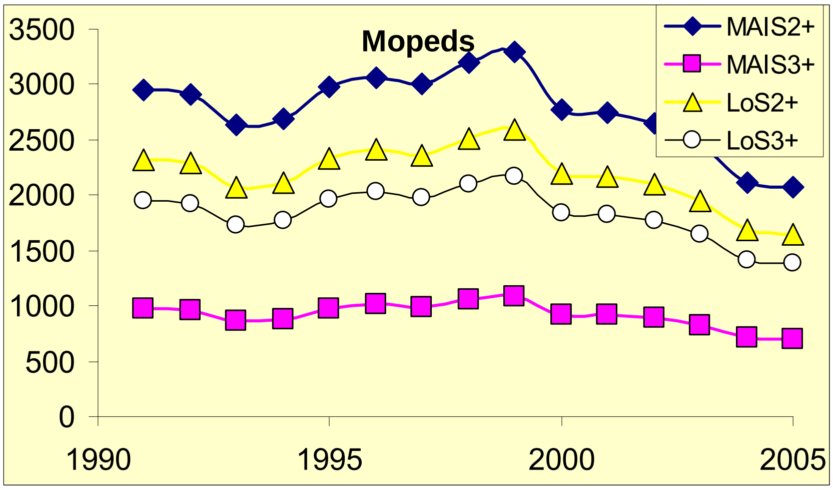
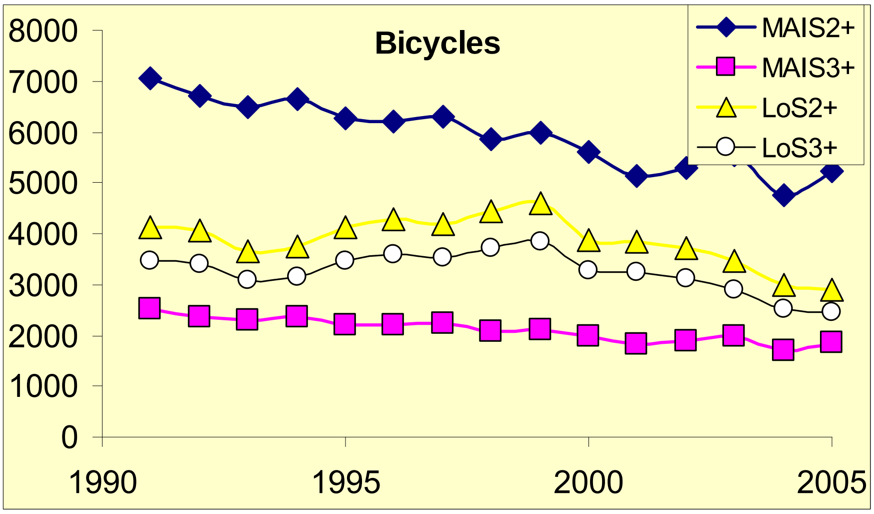


Figure 32: Bicycles by MAIS2+, MAIS3+, LoS2+, LoS3+.



Transport
Project co-financed by the European Commission, Directorate-General Transport and Energy

Figure 33: Pedestrians by MAIS2+, MAIS3+, LoS2+, LoS3+.

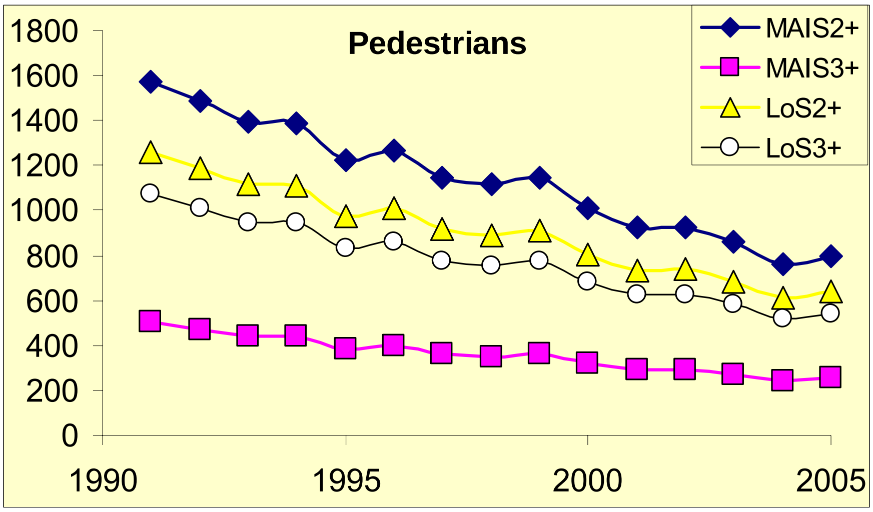
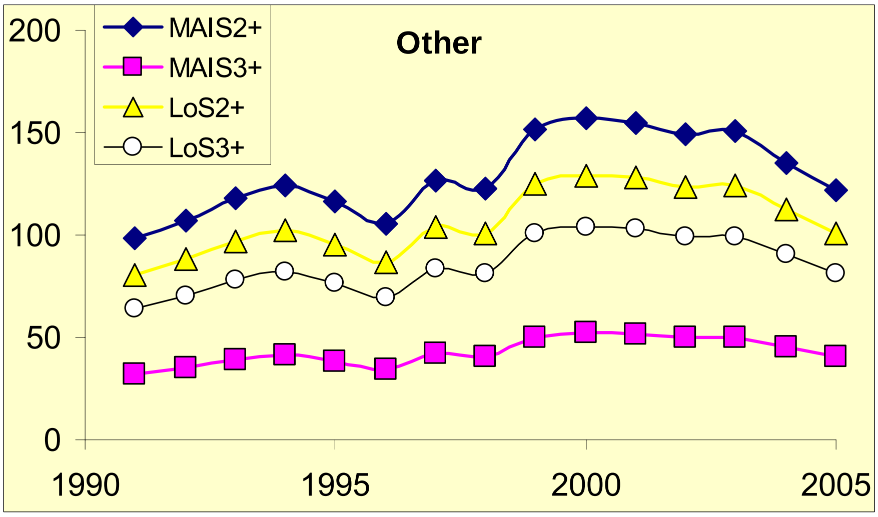


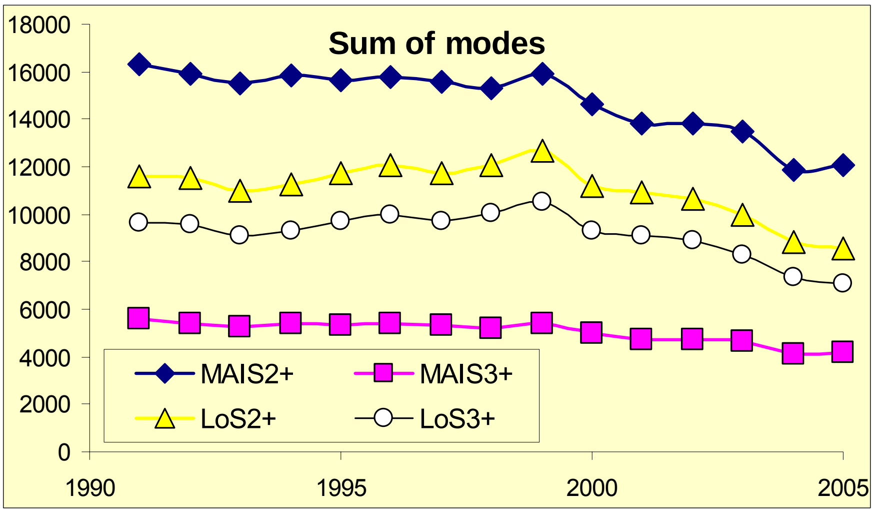
Figure 34: Other modes by MAIS2+, MAIS3+, LoS2+, LoS3+.



As the basis for each graph is the same (the number of police reported hospitalised and slight casualties by mode of transport) and the factors are constant for each year, the patterns observed within each graph are the same. The level is the only difference. We can see that the lower boundaries of MAIS2+ and MAIS3+ form the most extreme values. LoS2+ and LoS3+ are in between.

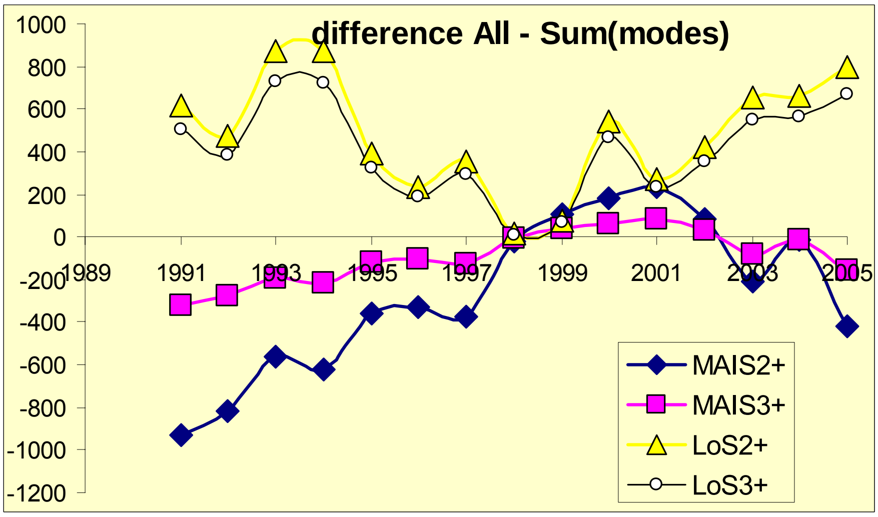
We can add up the numbers by mode of transport to reach the total number of casualties per severity boundary, see Figure 35.

Figure 35: Sum of modes by MAIS2+, MAIS3+, LoS2+, LoS3+.



Then we can compare the results if we would have applied the ALL factors from Table 92 and Table 96.

Figure 36: Differences between sum of modes and the application of ALL factors, irrespective of modes.



From Figure 36 it can be seen that there is a difference between the application of ALL factors and application of factors for each mode separately. The difference is acceptable in the period that the linking was performed and on which basis the factors were determined (1997-2003). However in years before and after, the difference is larger. For factors on LoS, the application of ALL factors lead to higher totals than the sum of the separate modes. With MAIS, the application of ALL factors in general leads to lower totals than the sum of separate modes.

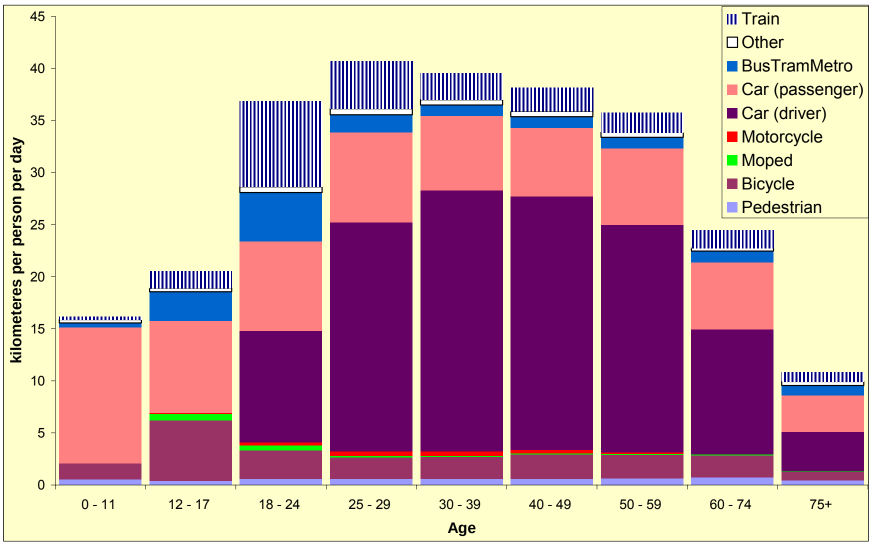
7.6.4.5 Age distribution Hospital - Police

In the second experiment to explore the reliability of the results obtained, we looked at the age distribution. The age of the casualty was not used in the calculation of the correction factors. There were two main reasons not to do so:

1) If the numbers were split by age group, too few would remain to calculate reliable correction factors.
2) There is evidence that the age of traffic participants is correlated to their mode of transport.

From a point of mobility the use of a certain mode of transport can be addressed to specific age groups quite well: cycling is dominated by teenagers, moped riding is done by 16-20 year olds, and public transport is mainly used by people in their twenties. The car is dominant at every age, but the young and the very old often travel as a passenger.

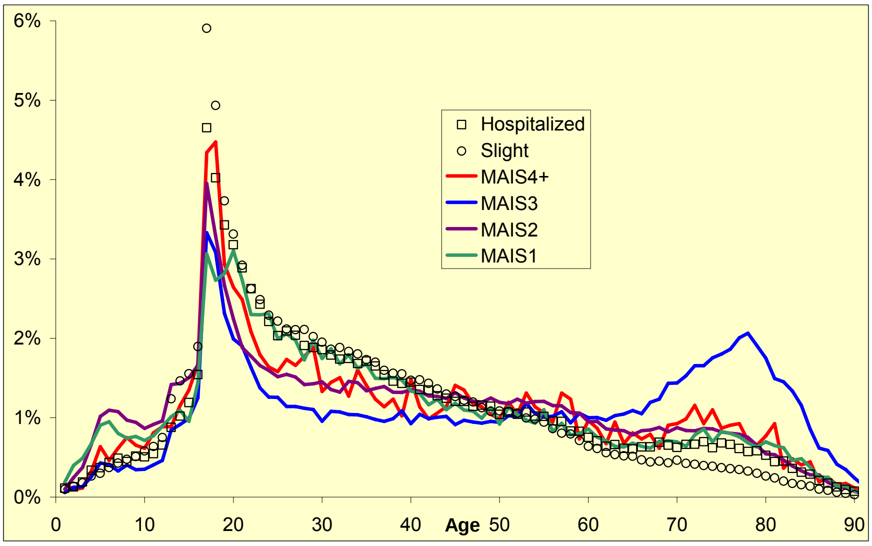
Figure 37: Number of travelling kilometres per person per day in The Netherlands, Sources AVV(MON), CSB(OVG) 1994-2006.



However, the insights that the accident rate is so different for beginners compared to that of experienced drivers and that the elderly are very fragile when involved in an accident help to understand that age and mode of transport are not very much correlated in terms of traffic casualties.

In a comparison of the age-distribution of hospitalised and slight injuries in the police database with the age-distribution in the hospital file by MAIS, it can be observed that age distributions in both files are not the same. The hospital records are assumed to represent the real distribution, as the real numbers are built for 95% from hospital records.

Figure 38: Age distribution of police records and hospital records for different severities.



The age distribution of MAIS3 casualties is the most different compared to the other severities (much lower for 20-50 year olds, much higher above the age of 60). The strong peak that is observed in the police records for 17 and 18 year olds is not seen so strongly in any MAIS distribution of the hospital file. The high MAIS2 level for 5-14 year olds is not observed in the police file.

Due to the smaller number of cases the pattern for MAIS4+ is the most varying one.

Table 100: Number of cases by mode and severity. Police and hospital files 1997-2005.

Source	severity	pedestr	bicycle	moped	motorcycle	car/van	other	SUM
Police file	Hospitalised	6457	20069	18159	6909	45035	1368	97997
	Slight	13511	63702	58534	12724	124168	4169	276808
Hospital file	MAIS4+	433	1583	781	453	1436	500	5186
	MAIS3	2629	18000	4736	2330	7285	2416	37396
	MAIS2	6539	35692	11984	6066	17101	4856	82238
	MAIS1	1430	6635	2441	1085	9442	1477	22510

The difference in age distribution may be caused by different age distributions per mode of transport. There are various ways to show that this is not the case. As an example the factors between the age distributions are plotted for cyclists below.

Project co-financed by the European Commission, Directorate-General Transport and Energy

Figure 39: Factor between the age distribution of hospital records and hospitalised police casualties by MAIS level. LMR+VOR 1997-2005.

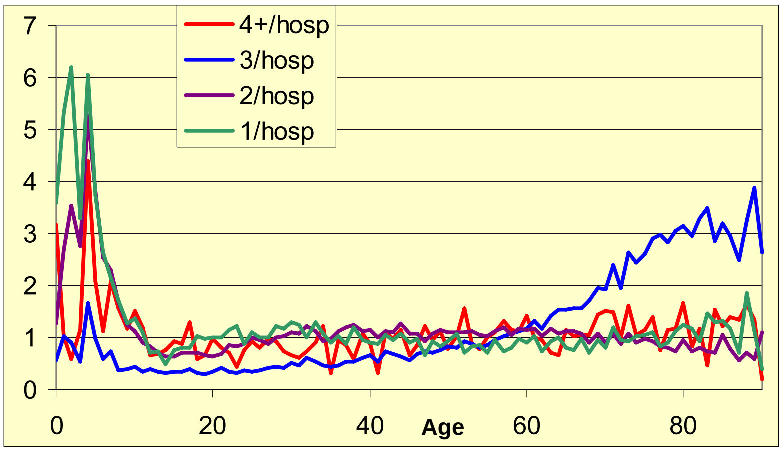
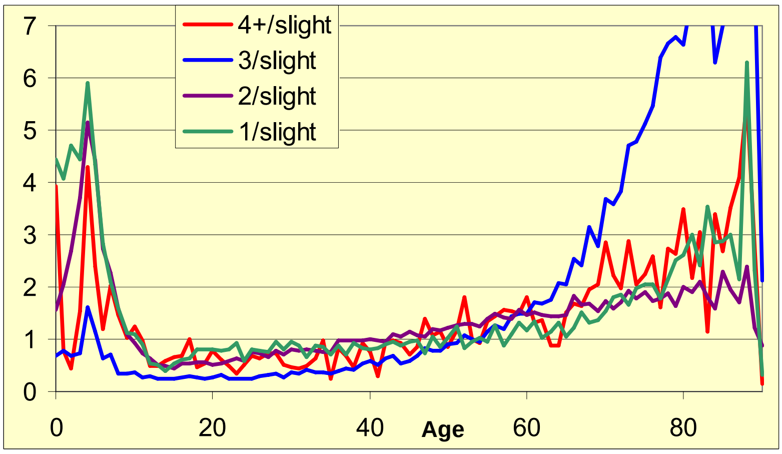


Figure 40: Factor between the age distribution of hospital records and slight police casualties by MAIS level. LMR+VOR 1997-2005.



This illustrates clearly that any factor on the police records will never result in an age distribution that is similar to the hospital file, if the age is not among the components used to determine the factors. Especially the numbers of casualties above 60 and below 10 years old are too low; for the other ages the number is too high.

7.6.5 Conclusions

Linking police reported traffic casualties and hospital patients having accident injuries was successfully carried out. On average 46% of the hospitalised casualties was found in both database, 49% was in the hospital database only, 2,5% in the police database only. It was estimated that 2,2% of casualties was reported in neither database.

Transport
Project co-financed by the European Commission, Directorate-General Transport and Energy

Severity estimates from the hospital file, expressed in MaximumAIS and Length Of Stay enabled to judge the severity assigned by the police. The more severe the injury, the better the reporting by the police.

As traffic is becoming safer, the number of fatalities to be studied is decreasing, which leads to statistical interpretation problems. A small difference with the expected trend may not be statistically significant. Analysis on severely injured casualties had the problems of underreporting and not being representative, next to the international comparability component. These linking studies and severity assessments from the hospital injuries enables to define sharp boundaries to groups of which underreporting coefficients are well known. With a boundary on MAIS3+, the group of Killed and Severly Injured (KSI) will be approximately 6 times lager than the group of killed only. However with a boundary on MAIS2+ the group of about 15 times larger. In The Netherlands it is technically and statistically possible to define these boundaries in a reliable way, so a definition of MAIS2+ is recommended for The Netherlands as study group. Unfortunately this appeared not feasible in the international setting.

Correction factors have been derived to calculate the real number of severely injured casualties (above several severity boundaries) from the police casualty data.

Problems have been observed in the application of these factors as not all resulting distributions follow the expected distribution. An example on age was given to illustrate this problem. Time dependency of the factors is another problem that needs to be addressed in any next study. Special attention should be given to the reliability of the derived factors. By disaggregating the data to age group and year, the numbers in each cell might become too small to estimate the factors with the required accuracy.

Compared to other countries, in the Netherlands we were very lucky to have complete national hospital files for a larger number of years (1997-2003). Other countries, being less fortunate with the numbers of hospitals and years, already experienced the small number problems right now, only using the disaggregation by mode of transport.

For international comparison, problems were encountered in the definition of the Maximum AIS, as not all countries have the same ICD version, the same number of diagnoses, and the same level of detail of the available injury codes. Special attention in this national report was given to the quality and characteristics of the medical file, such as the number of diagnoses and the level of detail of the diagnoses. By simulation of truncation and omitting diagnose codes valuable reference has been set.

The difference of severity is influenced both by the ICD version and the AIS version that was used. In The Netherlands ICD9-CM was used and a conversion to AIS1990, while other countries used ICD10 and a conversion to

AIS1998. For an analysis of the compound result on injury severity (MAIS) see the study from the United Kingdom.

These difficulties in defining a comparable severity in all European countries prevented the calculation of correction factors in all countries. Further research on medical coding of injury and possible conversions between them is required before any European traffic injury severity boundary can be set.

7.6.6 References

Hook, E.B., Regal, R.R., (1995). *Capture–recapture methods in epidemiology: methods and limitations*. Epidemiol. Rev. 17 (2), 243–264.

ICDMAP-90 User's Guide. The Johns Hopkins University & Tri analytics, Inc. (1998-2002).

Polak, P.H. (1997). *Registratiegraad van in ziekenhuizen opgenomen verkeersslachtofers; Eindrapport*. R-97-15. SWOV, The Netherlands.

Polak, P.H. (2000). *De aantallen in ziekenhuizen opgenomen verkeers-gewonden, 1985 – 1997; Koppeling van gegevens van de verkeersongevallen-registratie en de registratie van de ziekenhuizen*. R-2000-26. SWOV, The Netherlands.

Reurings, M.C.B., Bos, N.M., van Kampen, L.T.B. (2007). *Berekening van het werkelijk aantal in ziekenhuizen opgenomen verkeersgewonden 1997-2003, Methode en resultaten van koppeling en ophoging van bestanden.* R-2007-8. SWOV, The Netherlands.

Rosman, D.L., Knuiman, M.W., Ryan, G.A. (1996), *An Evaluation Of Road Crash Injury Severity Measures*, Accident Analysis & Prevention, Vol. 28, No. 2, pp. 163-170.

SWOV (2001). *A new linking procedure for the determination of the total number of hospitalised road traffic casualties by comparing police and hospital reports*. The Netherlands.

Wittes, J., Colton, T., Sidel, V., (1974). *Capture–recapture methods for assessing the completeness of case ascertainment when using multiple information sources.* J. Chronol. Dis. 27, 25–36.

7.7 Study carried out in Spain

Report prepared by Catherine Pérez (ASPB)

7.7.1 Introduction

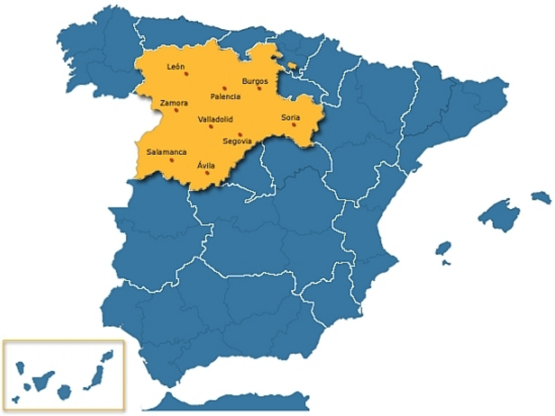
This section describes the national study carried out in Spain to achieve the aims of Task 1.5 of the SafetyNet IP. The objective of the task is to estimate the actual numbers of casualties from the CARE database. The general methodology has been previously described, therefore this section presents only the specific details of the Spanish study.

As a first step, we explore the feasibility of knowing whether it would be possible to carry out a probabilistic record linkage with the National Accident Police Registry (Dirección General de Tráfico, DGT) and the National Hospital Discharge Registry (HDR). Both registries have a national coverage. As a result of the first phase we selected an Autonomous Region where the study was feasible: Castilla y León. We describe the characteristics of this region and the methodology used, then present the results for this region.

Characteristics of Castilla y Leon

Castilla y Leon is the largest autonomous region in Spain, and is located in the upper centre of the Iberian Peninsula. It has an area of 94,233km^2 which represents 18,8% of all Spain and 2.523.020 habitants, 5,7% of all Spanish population. It is divided into 9 provinces. There is one city with more than 300.000 population, three between 100.000 and 200.000, and five between 50.000 and 99.999 habitants. In 2004 there were 1.440.056 vehicles registered, 5% of all vehicles registered in Spain. There are 18,890 km of roads, 11% of the total.

Figure 41: Map of Castilla y León



The National Traffic Authority (Dirección General de Tráfico, DGT) reported for year 2005, 9,857 road victims. Among them 384 were fatalities, 2,207 severe injured and 7,266 slight injured. In 2004 there were 2,367 hospitalisations due to road injuries.

In 2004 there were 21.0 road fatalities per 100.000 males and 5,4 road fatalities per 100.000 females, higher than the national estimates. (National Standardised mortality ratios are 17.2 and 4.9 respectively). During the same year there were 125.3 hospitalisations due to traffic injuries per 100.000 males and 43.9 per 100.000 females. (National estimates are 118.5 and 45.5 respectively).

7.7.2 Description of data sources

Hospital Discharge Register (HDR) database

The Hospital Discharge Register (HDR) database was provided by the autonomous department of health of Castilla y Leon. Data included records of hospitalisations due to road injuries from 1st July 2005 to 31st December 2005.

Case definition of road casualty:
We consider a road casualty if it fulfils the following criteria:

1. Suffering a injury defined by the International Classification of Diseases, 9 revision, Clinical Modification. (ICD-9-CM: 800 TO 959.9) **and**
2. Primary or Urgent hospitalisation (in front to scheduled, which would indicated that the hospitalisation is related to complications or rehabilitation process, and not due to a recent accident) **and**

a. Type of funding: road traffic insurance company **or**
b. External cause of injury: road traffic accident (E-Code: 810-819, and 826)

No severity level is provided by ICD-9-CM. A conversion from ICD-9-CM diagnosis to AIS was carried out using the method developed by MacKenzie and implemented in ICDMAP90 software.

Police data (DGT)

The Police (DGT) data on fatalities or injured in Castilla y Leon was provided by the National Traffic Authority for the year 2005. Only cases from the second half year were included for the linkage process.

Representativeness of data from Castilla y Leon

In this section we assess how representative are data from Castilla y Leon of road accident casualties in the rest of Spain. We use the national DGT registry for year 2005 and the national HDR for year 2004. Police data shows that among Castilla y Leon injured there are slightly more females, less children and youth, and more adults over 44 years old (Table 101). Regarding the vehicle, there are more car users and less motorcycle and moped users.

For the area where happened the crash, there is no difference for fatalities, but the proportion of non urban casualties is 11% higher than the rest of Spain for severe casualties and there are 21% more for slight casualties. Complementary the proportion of urban severe and slight casualties is lower among Castilla y Leon crashes.

Table 101: Characteristics of road victims in Castilla y Leon and rest of Spain by level of severity. DGT Spain, 2005.

	Fatal			Serious			Slight		
	Castilla y Leon	Rest of Spain	Total	Castilla y Leon	Rest of Spain	Total	Castilla y Leon	Rest of Spain	Total
N	384	3 473	3 857	2 207	20 237	22 444	7 266	103 684	110 950
GENDER									
Male	76.6	79.0	78.7	69.4	72.2	71.9	61.8	62.3	62.3
Female	23.4	20.2	20.5	30.4	26.5	26.9	37.3	35.6	35.7
Unknown		0.9	0.8	0.3	1.2	1.1	0.9	2.1	2.0
AGE									
<= 14	3.9	2.3	2.4	3.5	4.2	4.1	5.8	4.7	4.7
15 - 29	25.3	31.4	30.8	35.2	40.7	40.1	38.1	43.9	43.5
30 - 44	24.5	27.3	27.0	26.4	27.2	27.1	27.2	28.5	28.4
45 - 59	19.0	18.1	18.2	17.9	14.6	14.9	15.9	13.9	14.0
60 - 74	15.1	13.6	13.7	11.1	8.9	9.1	9.5	6.7	6.9
>75	12.2	7.3	7.8	5.9	4.5	4.6	3.5	2.4	2.5
VEHICLE									
Car	64.3	54.3	55.3	57.1	46.8	47.8	70.1	56.8	57.7
Motorcycle	5.2	11.2	10.6	8.5	12.6	12.2	3.2	9.4	9.0
Moped	3.1	6.4	6.1	6.4	16.2	15.3	5.1	15.3	14.6
Bicycle	2.3	1.8	1.8	2.1	2.0	2.0	1.3	1.6	1.6
Bus	1.3	0.5	0.6	0.8	0.7	0.7	1.1	1.8	1.8
Truck or lorry	10.4	8.8	8.9	11.7	7.3	7.8	10.4	6.3	6.6
Other	2.1	2.5	2.4	2.6	1.8	1.9	1.9	1.2	1.2
Unknown	11.2	14.6	14.2	10.8	12.4	12.3	6.9	7.5	7.5
AREA									
Non Urban	85.2	84.7	84.7	77.5	67.2	68.2	68.1	47.2	48.6
Urban	14.8	15.3	15.3	22.5	32.8	31.8	31.9	52.8	51.4

The lethality (number of deaths per 1000 victims) for non urban accidents is similar in Castilla y Leon than the rest of Spain. But the lethality for urban accidents is clearly higher (Table 102).

Table 102: Lethality (number of deaths per 1000 victims) in Castilla y Leon and rest of Spain. DGT, Spain 2005

	Castilla y Leon	Rest of Spain	Total
Non urban	46,8	44,9	45,1
Urban	19,9	8,6	9,1
Total	39,0	27,3	28,1

There is no reason to believe that severity of crashes is higher in Castilla y Leon than in the whole Spain. On the contrary, it suggests that there might be a significant underreporting of severe and especially slight casualties. Most slight casualties occur in urban areas, where motorcycles and mopeds are very popular in Spain. It could also explain the differences by type of road user.

Among hospitalised casualties there is a slightly lower proportion of females, and youth from 15 to 29 years (Table 103). Regarding severity among Castilla y Leon records there is a higher proportion of unknown severity.

Table 103: Characteristics of HDR road injured in Castilla y Leon and rest of Spain by level of severity. HDR Spain, 2004.

	Castilla y Leon	Rest of Spain	Total
N	2 196	29 812	32 008
GENDER			
Male	73.7	71.5	71.6
Female	26.3	28.5	28.4
AGE			
<= 14	10.7	10.1	10.1
15 - 29	34.9	37.6	37.5
30 - 44	22.4	21.7	21.7
45 - 59	15.1	13.4	13.6
60 - 74	11.8	10.7	10.8
75 - 99	5.1	6.5	6.4
MAIS			
1-Slight	10.2	12.3	12.2
2-Moderate	49.5	52.4	52.2
3-6 Severe	25.9	29.3	29.0
Unknown	14.4	6.1	6.5

In conclusion, we can assume that characteristics of casualties are not significantly different from the rest of Spain, except for urban casualties, which are more likely to be underreported than in the rest of Spain. We assume that data from Castilla y Leon can be used to estimate the number of severe casualties in the whole country.

7.7.3 Description of the linking process

Common variables available for the linking process were gender, age, date of the accident and autonomous region. A simple way to assess feasibility is to multiply the number categories for each variable (Roos and Waida, 1991). For instance, 2 (gender) * 100 (age) * 365 (days)= 73,000. This number must be greater than the total number of records summing up both databases (155.000 police records + 40,000 hospital records = 195,000). It showed to be clearly insufficient to identify true pairs of records.

A more complex way to assess feasibility of linkage considers not only the distribution of categories of variables, but also the information contained in them (Cook, 2001). It is based on calculating the weight (Wt) needed to achieve a specific probability that two records are a true pair. Once this weight (we call it as reference weight) has been determined, we check whether there is enough information to achieve it. The reference weight is calculated with this expression:

$$W_t = \log_2 (p / (1-p)) - \log_2 (E / (A \times B-E))$$

p = probability that a pair is correct
A = size of database 1
B = size of database 2.
E = number of correct pairs

The value of the weight to contrast with W_t will be the attainable minimum weight when all the variables agree exactly (Wmin). This will be obtained from the combination of the most frequent categories of each variable. For instance if we had only sex and the position in the vehicle, the minimum weight would be for those records in which the category of the variable sex was “man” and the one of the position was “driver” because they are the most frequent. Once determined, it will be compared with W_t (Newcombe, 1988).

If $W_{min} >= W_t$ → Sensibility or Specificity of the process > 95%

If $W_t - 3 <= W_{min} <= W_t - 1$ → 90%<= Sensibility and Specificity of the process < 95%.

Table 104: Example of weights obtained for different distribution of frequencies for the variable gender.

	Freq.	W_{min}	Freq.	W_{min}	Freq.	W_{min}	Freq.	W_{min}	Freq.	W_{min}
Male	0,1	3,31	0,2	2,31	0,5	0,99	0,6	0,72	0,9	0,14
Female	0,9	0,14	0,8	0,31	0,5	0,99	0,4	1,31	0,1	3,31

If our population were composed of 10% male and 90% female, the minimum weight would be for the category “female”, and would be 0.14. While the number of women is greater than men, the minimum weight will give this category, and will be increasing gradually. When the number of men and women are equal, the minimum weight is equal for both categories, and is of 0.99. When the number of men is greater to the one of women, i.e. 60%, the minimum weight gives the category “male”, and is of 0,72. Increasing the presence of males to 90%, the minimum weight is 0,14.

As a second step we checked through this process whether information on road casualties provided directly by some health departments of the autonomous regions would be feasible. Health departments provided more information to allow the linkage, such province, than the Ministry of Health at national level. Table 105 shows the feasibility considering only number of variable’s categories. For all autonomous region it would be feasible.

Table 105: Distribution of cases and categories among autonomous regions

Period Variables	Year 2005 Gender, age, date of accident/hospitalization.				
Autonomous regions	Cases HDR^a	Cases DGT^a	HDR + DGT	x Common Variables	Feasibility
Baleares	1.000	5.000	6.000	2 x 100 x 365=73.000	✓
Galicia	2.500	9.000	11.500	2 x 100 x 365=73.000	✓
Castilla-León	3.000	10.000	13.000	2 x 100 x 365=73.000	✓
C.Valenciana	4.000	13.000	17.000	2 x 100 x 365=73.000	✓

^{a}: *Estimated cases.*

We assume that the injured are hospitalised in the same autonomous region where the accident occurred

If we calculate the reference and minimum weights for these autonomous regions only Castilla y Leon and Galicia are feasible. Table 106 shows the results of reference weight (W_t) and the minimum weights (Wmin) for different autonomous regions. Those in black show cases where it is possible to reach values of Sensitivity and Specificity between 90% and 95%. In no case, with these values it is possible to reach values of Sensitivity and Specificity greater than 95%. Therefore we decided to carry out the national study in the autonomous region of Castilla y Leon.

Table 106: Reference weights and minimum weights to assess the feasibility of probabilistic record linkage

	ESPAÑA (2.001)	BALEARES	GALICIA	CAS-LEÓN	C.VALENCIANA
W_{t} (p=0,95)	20,34	15,35	16,21	16,54	16,91
Minimum Weight (Wmin).					
Gender + age + hospitalisation date	14,04	10,29	11,83	11,83	11,11
Gender + age + hospitalisation date + hospital province	16,44	10,43	13,33	13,95	11,25
Gender + age + hospitalisation date + hospital reference area	*	11,74	13,52	14,46	13,83

* Not available

The method used for linking records was a mix, that use probabilistic, with the aid of deterministic to generate blocking and with a final manual review. The probabilistic linkage process consists in matching two or more records which are believed to belong to the same individual. It is based in two probabilities: the probability of matching given that both records belong to the same individual and the probability of matching by chance. The less probable is a value of the variables, the greater is the weight assigned. The process is done by the software WCONNECTA developed by the Agència de Salut Pública de Barcelona (ASPB) (Cirera et al, 2000 and Arribas et al, 2004).

The matching process implies these phases:

1. Data preparation
2. Selection of linkage variables
3. Evaluation of process feasibility
4. Computation of simple weights
5. Restriction of comparison pairs (blocking)
6. Comparison stage (matching)
7. Simple weights assignment
8. Computation of composite weights
9. Decision stage (linking)
10. Threshold determination
11. Review of dubious pairs

1. Data preparation

During July-December of 2005, 1,636 people were admitted a public hospital (HDR) of Castilla y Leon as a result of the injuries suffered in a traffic accident. During the same period, 6,970 victims of different severity were reported by police (DGT).

2. Selection of linkage variables

Variables used for linkage were age, gender, date of the crash and province of accident / province of hospital. The absence of a common identifier for both databases resulted in the need to link both sources of information with the probabilistic method, using the information of the common variables in both databases. We assume that the hospitalisation is done the same day of the crash, and in the same province of the crash.

3.Evaluation of process feasibility

The feasibility study has been previously explained in section 7.7.1. with the aim of justifying why we selected the autonomous region of Castilla y Leon.

4. Computation of simple weights

Prior to linking records it was necessary to compute weights that will become useful later in the linking phase. These weights are based on two probabilities, the probability of matching given that both records belong to the same individual, and the probability of matching by chance. The less probable is a value of the variable, the greater is the weight assigned (Jaro, 1995).

For each category within each variable there are three possible values based all of them in the distribution of the variables to be compared among both files and taking into account missing values. A value will be assigned to the pair if they coincided, a value will be assigned if they do not coincide and zero if one of the two values are missing.

5. Restriction of comparison pairs (blocking)

Once weights had been computed, it is necessary to compare the information obtained for the variables common to both files. This first step, known as the blocking phase (Jaro, 1995) consisted in forming blocks in order to reduce the comparisons number. In our case we form blocks with those HDR records for which the date and time of patient attendance was within three days after the crash occurrence reported in the police files.

6. Comparison stage (matching)
Within each block, two level comparisons were made: firstly, the contents of the common variables for both files for each HDR record with each DGT record were compared.

7. Simple weights assignment
Out of every between-variable comparison a weight value was assigned.

8. Computation of composite weights
At a second stage, a composite weight as the sum of the individual weights obtained in between-variable comparisons was generated, allowing the comparison between records.

Figure 42: Distribution of weights

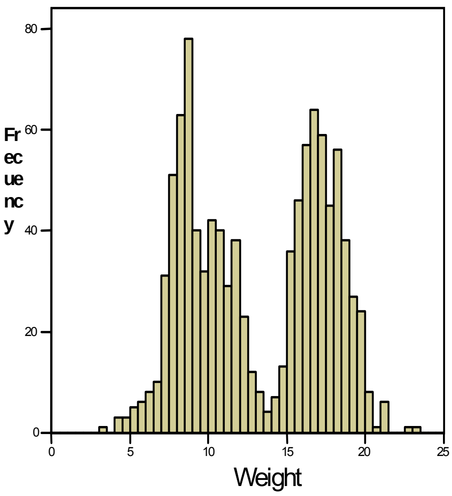


Figure 42 indicates that the distribution of the variable WEIGHT is bimodal. It seems reasonable to think that some characteristic in the registries exists that divides them in two groups, each one with a normal distribution.

We would suppose that those records that have not agreed with their real pair will have a relatively low weight, and in some few cases - based on the little frequency - not yet being real, the weight can get to be relatively high. On the other hand, those that have agreed with their pair, in most of cases will have assigned a relatively high weight, except those with the most common characteristics, that they will have the smaller weight. Under these assumptions, we can think that the characteristic that divides the records in two groups is being or not a real pair. The observation of the figure 7.7.2 indicates that a good cut off to differentiate the two distributions could be the corresponding one to weight 13,5 (W_L).

9. Decision stage (linking)

An HDR record was matched to a police record when it was the record with the highest composite weight after its comparison with all the remaining records in the selected block. If there are two or more records with the same weight for one HDR record, it remains unlinked, because it is impossible to distinguish which record corresponds to the same person.

Out of 1693 HDR records, 1016 have been linked (63%), although we need to decide which of them are really true pairs. In order to decide which are true pairs, we analyse both the number of variables that agree and the weight of each pair.

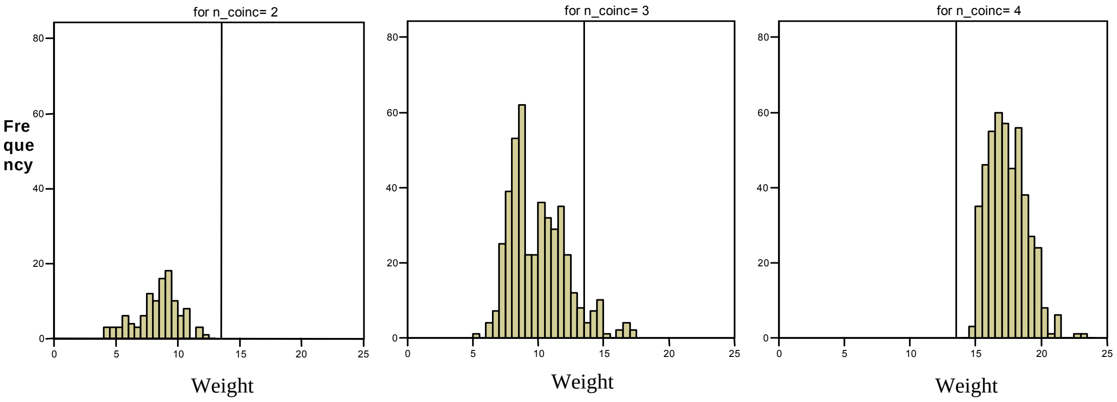
Table 107 shows the distribution of the number of coincidences. Two of the 1,016 records obtain the maximum weight with a record with which they only have in common or the age, or the sex, or the day of the accident or the province where the accident happened. That is its “better pair”, but it is not sufficient to trust that it is its true pair. On the other hand, 45.6% of the cases were associated to a record which agreed in all the variables.

Table 107: Distribution of number of coincidences.

	N	%
1	2	0,2
2	112	11,0
3	439	43,2
4	463	45,6
Total	1.016	100,0

Figure 7.7.3 shows the distribution of the weights assigned according to the number of coincident variables. The weighs cut-off W_L is also indicated. Using this point, all the records in which all the variables have agreed and 7% are considered correct of which the three variables have agreed. They do not consider any of the pairs with less correct than three coincident variables.

Figure 43: Weight distribution according the number of coincident variables



10. Threshold determination

Using these weights, two threshold values were defined: the lower-threshold limit, under which all records with such weight value would be considered not corresponding to the same individual, and the upper-threshold limit, above which a record would be considered to belong to the same individual.

11. Review of dubious pairs

For those pairs with a weight value between the two threshold limits, a manual review of the data by three reviewers was established, using additional information, in order to decide if the linkage was accepted. The software W-conecta includes and adaptation of the sequential review process used in the field of quality control. That is when dubious cases are reviewed, considering the number of correct or incorrect pairs, the software indicates how the thresholds must be modified. If the chosen interval there are too many correct records the program suggest to lower the upper threshold, and if there are too many incorrect records to higher the lower threshold.

The intervals decided to be reviewed were established observing the histogram and the distribution of weights. Review of dubious cases was done by three people to assure objectivity. A conservative criteria was chosen: the linkages between records considered correct by two or more persons, were considered correct linkages. The same criteria was used for incorrect pairs of records.

We will consider as linked 493 pairs. This represents 48.5% of the HDR records that had assigned to some police record, and 30% of the total of HDR records Castilla y Leon.

Finally, it is necessary to discard the possibility that the connected data are biased towards the cases of more peculiar characteristics. Table 7.7.8. shows the basic characteristics of the individuals in the three groups: total hospitalised, hospitalised which have assigned any police record (63%) and total of pairs that are considered truly linked (30%).

Table 108: Main characteristics of the individuals hospitalised due to road injuries. Percentage, quartiles and values minimum and maximum.

	Total hospital records	Linked records	Real pairs
N	1.636	1.016	493
GENDER (%)			
Male	70,5	70,7	70,8
Female	29,5	29,3	29,2
AGE			
Mín	0	0	2
P25	21,0	23,0	24,00
P50	35,0	35,5	35,00
P75	55,0	56,0	55,50
Máx	97	97	97
PROVINCE (%)			
Avila	4,6	5,4	7,1
Burgos	22,7	21,2	20,9
León	19,8	18,8	18,3
Palencia	8,6	9,4	10,5
Salamanca	11,6	12,2	14,2
Segovia	4,4	4,7	3,7
Soria	3,9	5,3	5,9
Valladolid	16,8	15,2	12,4
Zamora	7,6	7,9	7,1
LENGTH OF STAY			
Mín	0	0	0
P25	2,0	2,0	2,0
P50	5,0	5,0	6,0
P75	11,0	11,0	12,0
Máx	109,0	89	82
TYPE OF DISCHARGE (%)			
Home	85,7	83,2	81,1
Transfer	10,5	11,6	12,8
Voluntary discharge	,8	1,0	1,0
Fatality	3,0	4,1	5,1

The results obtained in the previous analyses suggest to consider as weight threshold (W_L) the value of 13,5. The pairs with a weight equal or greater than this are considered true linked pairs (corresponding to the same individual).

Validity of record linkage

As there are no unique identifiers or a subsample of records with unique identifiers that could identify the linked pairs as real pairs, two fictitious data bases for the calculation of the values of Sensitivity and Specificity have been built up.

The databases were created using the following procedures: initially we decide the percentage of cases of the source that we expect to link with the records of source B. In our case, we suppose that we would expect to be able to link 70% the records of the HDR with any record of the DGT.

The fictitious database contains the HDR records used for the linkage (n=1636). Out of these records 70% have been selected randomly that will also be included in the data base B. There is a variable that identifies each record (ID) and another that indicates if the registry comes also from data base B.

The fictitious database B contains 70% of the records used for the linkage that we have selected randomly (n=1.187) and in addition, have added 5,800 records corresponding to other years which the year of the accident has been modified until arriving to quadruplicate the size of the data base A (from the data of previous years, we know that the record number of the data base of the HDR is approximately one fourth of the record number of the DGT). A new ID has been assigned to the new cases different from those assigned previously to the records of source A.

Once the two databases have been created, we proceed to link them. Records are classified as linked and unlinked according to the value of the weight threshold (W_L) established in the real connection. The ID variable is used to judge the real status of the pair of registries. Thus, we will have four possible situations (Table 109):

Table 109: True and false positives and negatives according the coincidence with the identifier (ID)

		Real pair (same ID)	No real pair (different ID)
Result of linkage process	linked $W \geq W_L$	True Positives (TP) 1,129	False Positives (FP) 20
	Not linked $W< W_L$	False Negatives (FN) 58	True Negatives (TN) 429

True Positives (TP): Number of records pairs linked correctly.
False Positives (FP): Number of records pairs wrongly linked.
True Negatives (TN): Number of records unlinked pairs correctly.
False Negatives (FN): Number of records wrongly unlinked pairs.

From the previous parameters, we can obtain different measures to evaluate the accuracy the record linkage:

- ***Sensitivity*** (S): S=TP/(TP+FN)

The number of pairs of records true linked pairs divided by the total number of correct pairs of records. It is interpreted as the probability that a concordant pair of records has been connected by the process.

- ***Specificity*** (E): E=TN/(FP+TN),

The number of unlinked correctly divided the total number of pairs of incorrect records. It is interpreted as the probability that a discordant pair does not connect in the process.

- ***Positive Predictive Value*** (PPV). VPP=TP/(TP+FP)

The number of correct linked records divided by the total number of pairs of linked records. It is interpreted as the probability that a pair linked is a really pair. The PPV is useful as an indicator accuracy of the linkage process.

- ***Match Rate*** (MR). MR=(TP+FP)/(TP+FN)

The total number of linked records pairs divided by the total number of pairs of correct records.

- From our data, out of 1,636 HDR records, 1,149 were linked with a weight equal or higher than 13.5. It yields a **positive predictive value** of 98.3%.

- Out of 1,187 records of both databases, 1,129 were linked, which yields a **sensibility** of 95,1%.

- Out of 449 records of A database 429 were not linked, which yields a **specificity** of 95,5%.

- The **match rate** was 0,97.

According to this test there would be 20 false positives and 58 false negatives. This corresponds to 0.3% and 0.8% respectively of the police database, and 1.2% and 3.5% of the HDR database. We conclude that the likelihood that a pair of linked records belongs to the same person and accuracy of the procedures are high.

7.7.4 Results

In Castilla y Leon, during July-December of 2005, Police reported 6,970 people injured in traffic accidents, and public hospitals reported 1,636 persons hospitalised due to road injuries. Four hundred and ninety three records were linked between both databases. This corresponds to 7,1% of police records and 30% of hospital records. As expected the proportion of serious casualties is higher among linked records (Table 7.7.11.)

Table 110: Linked and unlinked police and hospital records distribution.

	Police		Hospital	
	n	%	n	%
Yes	493	7.1	493	30.1
No	6477	92.9	1143	69.9
Total	6970	100.0	1636	100.0

Table 111: Linked and unlinked police records by severity.

Police	Linked			Unlinked			Total		
severity	n	Col %	Row %	n	Col %	Row %	n	Col %	Row %
Fatal	13	2,6	5,3	234	3,6	94,7	247	3,5	100.0
Serious	326	66,1	20,3	1283	19,8	79,7	1609	23,1	100.0
Slight	154	31,2	3,0	4933	76,2	97,0	5087	73,0	100.0
Unknown				27	0,4	100,0	27	0,4	100.0
Total	493	100		6477	100.0		6970	100.0	100.0

The distribution of casualties by MAIS is similar among linked and unlinked hospital records. The proportion of linked records increases with severity, from 30.9% for MAIS 1 to 50% for MAIS 6.

Table 112: Linked and unlinked hospital records by MAIS (Maximum Abbreviated Injury Severity).

MAIS	Linked			Unlinked			Total		
	n	Col %	Row %	n	Col %	Row %	n	Col %	Row %
0	20	4.1	17,4	95	8.3	82,6	115	7.0	100.0
1	54	11.0	30,9	121	10.6	69,1	175	10.7	100.0
2	242	49.1	29,7	572	50.0	70,3	814	49.8	100.0
3	99	20.1	31,8	212	18.5	68,2	311	19.0	100.0
4	51	10.3	33,3	102	8.9	66,7	153	9.4	100.0
5	20	4.1	40,0	30	2.6	60,0	50	3.1	100.0
6	1	0.2	50,0	1	0.1	50,0	2	0.1	100.0
9	6	1.2	37,5	10	0.9	62,5	16	1.0	100.0
Total	493	100.0	30,1	1143	100.0	69,9	1636	100.0	100.0

The proportion of linked records increased also with longer length of stay.

Table 113: Linked and unlinked hospital records by length of stay.

MAIS	Linked			Unlinked			Total		
	n	Col %	Row %	n	Col %	Row %	n	Col %	Row %
Overnight	15	3.0	24.6	46	4.0	75.4	61	3.7	100.0
1-3	149	30.2	25.9	426	37.3	74.1	575	35.1	100.0
>3	329	66.7	32.9	671	58.7	67.1	1000	61.1	100.0
Total	493	100.0	30.1	1143	100.0	69.9	1636	100.0	100.0

The distribution of gender among linked records is similar to hospital only records, but the proportion of females is 5% lower than police only records.

Regarding age, linked records shows a higher proportion (7.3%) of children under 14 years old than police only records (4.5%), but lower than hospital only records (14%). Youth from 18 to 35 years are more represented among police only records (44.8%), and less among hospital only (29.6%) and linked records (39.6%). On the other side, elderly are more represented among hospital only records (15.1%) and linked records (15,6%).

Table 114: Gender and age group according to data source

	Police only		Police ∩ Hospital		Hospital only	
	N=6477	%	N=493	%	N=1143	%
Gender						
Male	4141	63.9	344	69.8	805	70.4
Female	2274	35.1	148	30.0	338	29.6
Unknown	62		1	0.2		
Age						
0-13 years	290	4,5	36	7,3	160	14,0
14-15 years	109	1,7	8	1,6	41	3,6
16-17 years	199	3,1	10	2,0	40	3,5
18-35 years	2900	44,8	195	39,6	338	29,6
36-50 years	1433	22,1	94	19,1	221	19,3
51-65 years	814	12,6	73	14,8	170	14,9
66-98 years	602	9,3	77	15,6	173	15,1
999	130	2,0				

Information about the type of road user or of vehicle is not available for hospitals. It should be recorded with the E-code of the ICD-9-CM. But usually it is not recorded, or it is recorded only as traffic crash, without specifying anything else. A previous study (Pérez et al, 2006) showed that in Spain the E-code only gives useful information on these variables for 21% of HDR at national level. Therefore it is not possible to compare and to derive conversion factors for road user or vehicle.

Among the linked records there is a higher proportion of pedestrians, pedal cyclists and motor cyclist , and a lower proportion of car occupant (Table 115 and Table 116).

Table 115: Road user distribution among linked and unlinked records

	Police only		Police Linked	
	n	%	n	%
Car occupant	4389	67,8	286	58,0
Pedestrian	440	6,8	64	13,0
Pedal cyclist	112	1,7	13	2,6
Motor cyclist	636	9,8	68	13,8
Other	894	13,8	62	12,6
Unknown	6	0,1		

Table 116: Type of vehicle distribution among linked and unlinked records

	Police only		Police Linked	
Type of vehicle	n	%	n	%
Car	4389	67,8	286	58,0
Motorcycle	290	4,5	37	7,5
Moped	346	5,3	31	6,3
Bicycle	112	1,7	13	2,6
Bus	19	0,3	4	0,8
Truck or van	706	10,9	45	9,1
Other	169	2,6	13	2,6
Unknown	446	6.9	64	13.0

Some of the differences found between linked and unlinked data might be due to the record linkage process. Cases with less common characteristics receive a higher weight than those that are more common, such 18 to 35 year old men or car occupants. It is more likely to find several records with the same characteristics.

Conversion factors

Table 117 and Table 119 show the number of cases reported by police and hospital and the estimated cases according police severity by length of stay and MAIS. Table 118 and Table 120 show the conversion factors derived from these data. (Results are also presented in the general section of the report).

Table 117: Police and hospital reported cases and estimated cases by length of stay

Length of Stay	Police severity						Estimated cases			
	Fatal	Serious	Slight	Not known	Not in police data-base	Total	Fatal	Serious	Slight	Total
Overnight	7	3	5		46	61	28	12	20	61
1-3 nights	3	80	66		426	575	12	309	255	575
>3 nights	3	243	83		671	1000	9	739	252	1000
Sub total	13	326	154		1143	1636	43	1082	511	1636
Police not hospital	234	1283	4933	27			234	1283	4933	6450
Total	247	1609	5087	27	1143	8113	277	2365	5444	8086

Table 118: Conversion factors by length of stay

Length of Stay	Conversion factors		
	Fatal	Serious	Slight
Overnight	0.115	0.008	0.004
1-3 nights	0.047	0.192	0.050
>3 nights	0.037	0.459	0.050
Sub total	0.175	0.672	0.100
Police not hospital			
Total	1.122	1.470	1.070

Table 119: Police and hospital reported cases and estimated cases by MAIS

MAIS	Police severity						Estimated cases			
	Fatal	Serious	Slight	Not known	Not in police data-base	Total	Fatal	Serious	Slight	Total
1+2	4	178	114		693	989	2471	18781	5314[1]	7439
3	0	81	18		212	311	0	254	57	311
4	3	40	8		102	153	9	120	24	153
5	4	14	2		30	50	10	35	5	50
6	0	0	1		1	2	0	0	2	2
0+9	2	13	11		105	131	10	66	55	131
Sub total	13	326	154		1143	1636	276	2353	5457	8086
Police not hospital	234	1283	4933	27						
Total	247	1609	5087	27	1143	8113				

[1] Includes proportional distribution by severity plus "Police not hospital" cases, assuming that these cases are of low severity.

Table 120: Conversion factors by MAIS

MAIS	Conversion factors		
	Fatal	Serious	Slight
1+2	1,001	1,167	1,045
3	0,000	0,158	0,011
4	0,036	0,075	0,005
5	0,040	0,022	0,001
6	0,000	0,000	0,000
0+9	0,041	0,041	0,011
Total	1,119	1,462	1,073

7.7.5 Conclusions

In this report we have shown that although it has been impossible to carry out a national study to estimate conversion factors to address the issue of underreporting road casualties, it is feasible to carry out the study with only one autonomic region. The validity of the record linkage procedure is quite good, and the conversion factors derived seem to yield reasonable estimates at national level.

The proportion of police records linked has been very low (7,1%). This proportion was expected because we are linking with hospitalised records, which by definition have some level of severity, and not all casualties. We know from a Road Injuries Information System based on hospital emergencies, that in Barcelona 7,8% of road casualties are hospitalised and 3,2% are transferred to another hospital, some of them are later on hospitalised.

Conversion factors derived from this study would not be appropriate for slight and for urban casualties, as slight casualties are underreported in the Castilla y Leon data. We tried to estimate conversion factors from a record linkage done in Barcelona. But it is not representative of the national urban data, because in Barcelona local police report exhaustively traffic crashes and therefore conversion factors were very small. This is not the situation in many cities in Spain, although the quality of reporting is improving because many cities are setting up urban road safety plans.

Some limitations need to be considered. First of all, due to lack of availability data we only used a 6 months database from only one autonomous region. It would be convenient to repeat the study including a year or even several years of data and at least one or two more autonomous regions where the study would be feasible and were there is a good balance of urban and non urban crashes.

Secondly, we do not have information about the coverage of HDR in Castilla y Leon. For the whole Spain, we know that for the recent years the HDR has a good coverage of public hospitals, around 99%, but do not include private hospitals. We can assume, however that serious injuries in general attend a public hospital. Nonetheless it should be assessed.

Thirdly, and related to coverage when studying a region, it can happen that in some cases the persons injured in a collision in Castilla y Leon might be transferred to a neighbourhood autonomous region because there is a large hospital closer. This can occur in for instance in provinces closer to Madrid. In this case police records would not be able to be linked to hospital databases, and conversion factors would give lower estimates.

Finally it would be very useful to derive conversion factors from different features such gender an age, and type of road user. The E-code is recorded better In some autonomous regions.

In conclusion, the methodology of the study carried out to obtain conversion factors to estimate serious casualties is appropriate, but in order to apply these conversion factors we recommend to repeat the study with a broader database.

Acknowledgements
To Eva Cirera and Marc Marí-Dell'Olmo. To the Department of Information and Health from Castilla y León and the National Traffic Safety Authority (DGT).

7.7.6 References

Arribas P, Cirera E, Tristan-Polo M. (2004). *Buscando una aguja en un pajar: las técnicas de conexión de registros en los sistemas de información sanitaria.* Med Clin (Barc). 2004 Feb 15;122 Suppl 1:16-20

Cirera E, Plasencia A, Ferrando J and Arribas P (2000). *Probabilistic linkage of police and emergency department sources of information on motor-vehicle injury cases: a proposal for improvement.* J Crash Prevention and Injury Control 2000; 2(3):229-237.

Cook LJ, Olson LM, Dean JM. *Probabilistic record linkage: Relationships between file sizes, identifiers, and match weights*, 2001;40:196-203.

Roos LL, Wajda A. *Record linkage strategies*. Methods of information in Medicine, 1991;30:117-123

Jaro M A. *Matching and Record Linkage*, in Business Survey Methods, eds. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, New York: Wiley, 1995 pp. 355–384.

MacKenzie, E. J., Sacco, W et al. (1997). *ICDMAP-90: A users guide.* Baltimore, The Johns Hopkins University School of Public Health and Tri-Analytics, Inc.

Newcombe HB (1988). *Handbook of record linkage: methods for health and statistical studies. Administration and business*. Oxford: Oxford University Press,.

Pérez C, Cirera E, Borrell C, Plasencia A on behalf of the work group of the Spanish Society of Epidemiology on the Measuring of the Impact on health of road traffic accidents in Spain. *Motor vehicle crash fatalities at 30 days in Spain.* Gac Sanit. 2006; 20(2):108-15.

7.8 Study carried out in the United Kingdom

Report prepared by Jeremy Broughton and Maureen Keigan (TRL)

7.8.1 Introduction

The UK study has consisted of linking the road accident data from Scotland for 1997-2005 with medical data from the Scottish Hospital In-Patient System. The linkage had been carried out previously for road accidents occurring between 1980 and 1995 (Stone, 1985; Keigan et al, 1999). For the SafetyNet project, the procedures and software were updated and applied to data from the 1997-2005 period. Various comparisons are made of the results from the current study with the results of these earlier studies.

The general concept of linking and comparing road accident and medical databases has been applied in a number of previous studies and can be implemented in various ways. The study reported by Simpson (1996), for example, used medical data recorded by clerks working in A&E Departments of a sample of 16 hospitals in Great Britain. Collecting data in this way is relatively expensive, however, so the specific approach adopted for the SafetyNet collaboration was selected on the basis of the funding available and the level of experience that existed in the eight national teams.

The United Kingdom comprises England, Wales, Scotland and Northern Ireland. Scotland is the most northerly of these countries, with almost 9% of the overall population and total of registered vehicles. While Scotland has had a devolved government for several years, the same traffic laws apply throughout the United Kingdom, the 8 Scottish police forces operate in the same way as those in the rest of the country and they are subject to the same operational pressures. Table 121 compares the casualties in Scotland in 2005 with the UK total.

Table 121: Proportion of UK casualties in Scotland, 2005

	UK total		% in Scotland	
	killed	injured	killed	injured
Pedestrians	699	33249	9.4%	9.0%
Pedal cyclists	152	16558	10.5%	4.6%
Motorcycle users	584	24669	5.8%	4.2%
Car users	1756	182797	8.8%	6.1%
Others	145	18567	11.0%	8.6%
All road users	3336	275840	8.6%	6.3%

7.8.2 Description of data sources

The STATS19 file

All road accidents involving personal injury and at least one vehicle occurring on the highway ('road' in Scotland) that were reported to the police within 30 days are recorded in the National Road Accident database (STATS19). Details of accident circumstances and the vehicles and casualties involved are recorded in annual files.

Transport

Project co-financed by the European Commission, Directorate-General Transport and Energy

A file of all casualties in Scotland was extracted from the national STATS19 files for the years 1997 to 2005. SHIPS data were supplied for 1995-2005, but the ICD9 system of injury coding used in 1995 and 1996 was superseded in 1997 by the ICD10 system so the main results are from 1997-2005.

The STATS19 variables extracted for the matching process consisted of police force area, date and time of the accident, casualty type, age and gender, casualty severity, road user and type of vehicle involved. The type of road user and vehicle involved were combined to create the following road user classes which were used for matching.

Code	Road user
1	Driver of a motor vehicle
2	Passenger of a motor vehicle
3	Rider of a motorcycle
4	Passenger of a motorcycle
5	Pedal cyclist
6	Pedestrian
7	Unknown

The SHIPS file

The Scottish Hospital In-Patient System (SHIPS) data were supplied by the Healthcare Information group of NHS National Services Scotland. It was released to TRL under a confidentiality statement for users of NHS patient data. Episodes (casualties) with an 'Emergency - Road Traffic Accident' type of admission or a specified Road Traffic Accident diagnostic code on the hospital discharge that were admitted and discharged within the years 1997 to 2005 were selected.

TRL has been advised that the operational procedures for SHIPS were unchanged between 1997 and 2005, and indeed for many years previously. Thus, any changes that are identified when the data are analysed by year cannot be attributed to changes in the data collection procedure. They must caused by changes in the number and nature of casualties, or the criteria used for admission as a hospital in-patient.

A large number of variables are provided within this file including hospital code, age and gender of the patient, admission type, length of stay and six diagnosis codes. The hospital code is a unique alpha-numeric five character code. The age of the in-patient is provided and has been calculated from date of birth. The admission type notes the reason for admission and includes, for example, whether it is from an emergency, a transfer or the waiting list. The length of stay used in this report refers to total length of stay and is either for one admission or has been accumulated from a number of 'stays' in connection with the diagnoses. Thus, the earlier problems (Stone, 1984) of linking patient records for the same person with more than one admission or with additional transfer records from the same accident have been eradicated by the suppliers of the data.

The files sent to TRL now hold information on a Continuous Inpatient Stay basis. A continuous inpatient stay (CIS) is a continuous period of time spent as an inpatient or day case in hospital regardless of any transfers between specialities or hospitals. From 1997 the diagnosis codes are based on the International Classification of Diseases (ICD) codes (World Health Organisation, 1992) using the ICD10 format.

Length of Stay in the SHIPS data is in effect the number of nights spent in hospital. 0 days is recorded for a patient admitted and discharged on the same day, 1 day for a patient discharged the day after admission (irrespective of time on either day) etc. It is thus possible to adopt exactly the definition of Length of Stay set out in Section 2.2.

Some of the variables included in the SHIPS data have been used to generate fields for use in matching the two datasets. The ICD10 V codes relate to external causes of morbidity and mortality; codes V01 - V99 define transport accidents and have been used to determine the mode of transport of the casualty. These codes, provided in the SHIPS data as one of the six diagnosis codes, are as follows.

ICD10 V code	Mode of transport
V01 - V09	Pedestrian
V10 - V19	Pedal cyclist
V20 - V29	Motorcyclist
V30 - V39, V80 - V86	Other motor vehicle
V40 - V49	Car occupant
V50 - V59	Light goods vehicle
V60 - V69	Heavy goods vehicle
V70 - V79	Public service vehicle occupant
V87	Other vehicle
V89, V98 - V99	Vehicle type unknown

The V codes also have a fourth character subdivision which has been used, where possible, to assign the casualty as a driver or passenger of the vehicle. In those records where a V code was not present a road user class of unknown has been assigned.

A look-up table was developed to achieve common definitions of road user. The seven classes of road user defined using STATS19 variables were taken and assigned to the V codes in ICD10. This set of codes has been used previously for linking data from these two sources for years prior to 1996. The rationale for using them again was that this would enable comparisons to be made of proportions of records that have been linked for datasets from previous years, thus ensuring confidence in this revised matching method.

The hospital code present in the SHIPS data provided a link to the police force area held within STATS19 as follows. As would be expected, the larger police forces have more hospitals within their areas.

Police Force Area	Hospital code letter
Northern	W, R, H, Z
Grampian	N
Tayside	T
Fife	F
Lothian and Borders	B, S
Central	V
Strathclyde	C, G, A, L
Dumfries and Galloway	Y

7.8.3 Description of the linking process

The principles used in this method have been developed from the work carried out by Nicholl (1980) and Stone (1984). In the first instance a 10 per cent sample of Hospital In-Patient Enquiry (HIPE) records were matched with STATS19 cases for England and Wales. This produced a matching procedure with a success rate of about 50% using distance from a hospital with an accident and emergency department, date and time (for accidents that occurred on the day before admission), gender, age to produce a score for each match. This work recommended that the E-code from the ICD9 code, which represents class of road user, should be made available for further work to reduce the number of possible matches.

The study reported by Stone (1984) is a comprehensive matching of police accident records and 100% sample of hospital in-patients records for Scotland for the year 1980. A matching algorithm using a similar basis to Nicholl in terms of matching variables and degrees of tolerance was developed. The bonus for this study was that the E-code describing class of road user was available and a matching proportion of 67% was achieved. Scottish Hospital In-patient data has been matched to STATS19 using this method for the years 1980 to 1995.

The main variables used in the matching process for the data reported here were police force area/hospital code, date of accident/admission, casualty/in-patient age and road user class. In addition, some STATS19 variables were allowed to vary according to the tolerance level; these are casualty severity, date of accident, police force area, age and road user class. Gender was included as a matching variable and was always an exact match.

Method

The data were imported into a MS Access database and queries have been conducted to perform the matching process. The tolerance levels have been applied in ascending order and once a record has been matched then it is withdrawn from further matching. The lower tolerance levels are considered the best match, for example, a match at tolerance level 1 requires all of the matching fields to have the same value.

Tolerance values

Using an Access update query for each of the 30 tolerance levels, the matching fields from the table containing the STATS19 records are linked with the corresponding field in the SHIPS data table. For some tolerance levels the STATS19 matching variables are permitted to vary, for example, in tolerance levels 7, 8 and 9 the age of the casualty in the STATS19 record may be plus or minus 1. Also, where the casualty severity is given a value of -1 a match with a fatal casualty in STATS19 is permitted. These queries are run to update the SHIPS record with the STATS19 unique identity reference and the tolerance level field. The tolerance values are summarised in Table 122; these are numeric ranges which are discussed following the Table. The procedures and values were established in a rigorous series of tests (Stone, 1984) and have been used in several STATS19-based linkage studies (Stone, 1984, Keigan et al., 1999, Broughton et al, 2001).

Table 122: Summary of tolerance levels

Level	Police Force area	Age	Road user class	Casualty severity	Length of stay	Hour	Day
1	0	0	0	0	0	0	0
2	0	0	7	0	0	0	0
3	0	0	0	-1	0	0	0
4	0	0	7	-1	0	0	0
5	0	0	0	0	0	2	1
6	0	0	7	0	0	2	1
7	0	1	0	0	0	0	0
8	0	1	7	0	0	0	0
9	0	1	0	-1	0	0	0
10	0	1	7	-1	0	0	0
11	0	0	0	0	0	4	1
12	0	0	7	0	0	4	1
13	9	0	0	0	0	0	0
14	9	0	7	0	0	0	0
15	0	0	0	1	5	0	0
16	0	0	7	1	5	0	0
17	0	2	0	0	0	0	0
18	0	2	7	0	0	0	0
19	0	0	5	0	0	0	0
20	0	0	0	0	0	4	1
21	0	0	7	0	0	4	1
22	0	0	0	1	0	0	0
23	0	0	7	1	0	0	0
24	0	0	0	0	0	0	1
25	0	0	7	0	0	0	1
26	0	0	0	1	0	1	1
27	0	1	0	1	0	0	0
28	0	1	7	1	0	0	0
29	0	3	0	0	0	0	0
30	0	3	7	0	0	0	0

There are certain special codes in Table 122:

- For Police Force area a value of 9 allows a match with defined neighbouring areas, 0 is the exact match value.
- Road user class unknown (value 7) allows an unknown road user class in SHIPS to match with any road user class in STATS19, also at tolerance level 19 the value 5 for road user class allows a pedal cyclist or pedestrian in SHIPS to match with either a pedal cyclist or pedestrian in STATS19 data. The exact match value is 0.
- Casualty severity of -1 permits a match with a fatal casualty in STATS19. A value of 1 allows a match with a slight casualty in the STATS19 data. The exact match value is 0.
- Where the length of stay is quoted as 5 a SHIPS casualty with a stay of 5 days or less may be matched with a slight casualty.
- The date of the accident is recorded in both datasets. However, time is only recorded in the STATS19 data so the hour is permitted to change in STATS19 by plus or minus n (1, 2 or 4) hours. The exact match value is 0.
- When the time is within n hours of midnight then the day is allowed to be the previous day and if n (1, 2 or 4) hours after midnight then the following day may be matched. The exact match value is 0.

7.8.4 Results

The SHIPS dataset includes 47,297 records for the years 1997 to 2005; these were matched to the STATS19 data using the 30 tolerance levels, and a total of 26,625 (56%) matches were achieved. All of these matches have been allocated on a one-to-one basis using the first appropriate match. In previous matching studies, a number of multiple matches were obtained that needed manual sifting to identify the best match using the local authority area. This system of tolerance levels inevitably leads to matches at the higher levels being less precise; however, it maximises the number of matches in cases where several potential matches have similar details.

The proportions of matches achieved at the different tolerance levels in 1997-2005 are given in Table 123. Results from the previous matching for 1993 and 1995 are included to examine the consistency of the new linkage with the original linkage procedure.

Table 123: Proportion of SHIPS records matched to STATS19

	ICD9			ICD10								
	1993	1995		1997	1998	1999	2000	2001	2002	2003	2004	2005
		Old method	New method									
A	3615	3610	3972	3469	3457	3191	2995	3016	2879	2595	2535	2488
B	5399	5321	6096	5767	5751	5448	5341	5359	5178	4914	4837	4702
A/B	67.0%	67.8%	65.2%	60.2%	60.1%	58.6%	56.1%	56.3%	55.6%	52.8%	52.4%	52.9%

A=number of matched SHIPS records, B=number of SHIPS records

In addition, new data for 1995 was obtained (the original 1995 dataset was no longer available at TRL) and the new matching procedure was applied; the data allow the results of the previous and the new method of matching to be compared directly. The number of SHIPS records for 1995 in this dataset was 6096 compared to 5321 previously obtained, the matches at the individual tolerance levels are broadly similar apart from level 24 where the proportion doubles from 2 to 4.1%, this tolerance level allows a SHIPS casualty to match with a STATS19 record on the following day (all other variables are exact). The overall proportion matched using the new procedure was slightly less at 65.2% compared to 67.8%, which is indicative that the method of linking has been applied consistently.

However, there is a steady fall in the proportion of SHIPS records that have been matched to STATS19 records during the years 1997 to 2005. This could indicate an overall decline in the proportion of road accident casualties reported to and by the police. Alternatively, certain attributes of an accident may make it more or less likely to be reported, and the range of casualties may have changed over this period so as to reduce the proportion that is likely to be reported to and by the police. For example, an increasing proportion of casualties could be admitted to hospital for observation, but found to be uninjured.

The proportion of matches achieved when the road user class is unknown in SHIPS and allowed to match with any class of road user in STATS19 is lower for the period 1997 to 2005 than for the earlier years. This may be a consequence of the change from using ICD9 to ICD10.

The ICD10 codes appear to determine the class of road user more precisely than the ICD9 codes. The proportion of SHIPS casualties where the road user class was unknown from the ICD9 code was almost 3 times higher than for the years where ICD10 codes were used.

The earlier version of the matching procedure ended with a manual comparison of the unmatched records, which identified further potential matches that lay just outside the system of tolerance levels. This was a laborious and time-consuming process that involved a degree of subjectivity. This final step was not included with the new matching procedure, which partly explains the slightly lower matching levels achieved by the new procedure – especially for 1995.

It is planned to systematically assess a subset of unmatched records to see whether the system of tolerance levels can be improved. In particular, there are indications that police reporting of age may have tended to deteriorate over time, so it may be appropriate to relax the age-related tolerances.

The AIS scores for each casualty are estimated from the ICD codes in the SHIPS file. The SHIPS data from 1997 used the ICD10 system, whereas the ICD9 system had been used previously, so a new mapping from ICD to AIS was needed. The mapping from ICD10 to AIS98 developed at the University of Navarra (Apollo, 2006) was adopted.

Whether MAIS is estimated directly or indirectly via such a mapping, unexpected results can often be found, such as fatal casualties with low MAIS scores. This may appear to call into question the validity of the linkage, although such results are also found in studies which do not rely on record linkage. Consequently, a number of such cases were examined in detail, and in each case the outcome was consistent with the ICD codes, i.e. there was no cause to doubt the linkage. Four SHIPS cases with MAIS 2 that were linked to a STATS19 fatality are presented below:

> Male pedestrian aged 20, MAIS 2, ICD10 injury codes included: Intracranial injury, unspecified injury of abdomen, lower back and pelvis and cardiac arrest. Length of stay was 0 days.
>
> Male pedestrian aged 83, MAIS 2, ICD10 injury codes included: fracture of vault of skull, unspecified injury of thorax, atrial fibrillation and flutter and respiratory failure (unspecified). Length of stay was 3 days.
>
> Male driver aged 19, MAIS 2, ICD10 injury codes included: other unspecified injuries of multiple body regions and cardiac arrest. Length of stay was 1 day.
>
> Male driver aged 34, MAIS 2, ICD10 injury codes included: intracranial injury and dependence on respirator. Length of stay was 19 days.

Trends

Keigan et al. (1999) presents various tables of results for 1980-95 that allow comparisons to be made with results of the new linkage, and overall trends since 1980 will now be examined. Some comparisons can be made exactly, others must be approximate because of the groups used in the earlier report.

First, Figure 44 examines the proportion of STATS19 casualties in Scotland that could be linked to SHIPS records. The data for 1980-95 come from the previous linkage, the data for 1997-2005 come from the new linkage and there has been no linkage for 1996. Between 53 and 60% of serious STATS19 casualties each year were linked to SHIPS records; the mean proportion in 1997-2005 is 57.5%, slightly higher than the 1987-95 mean of 56.8% although there has been a downward trend since 1999 of about 0.7% p.a.

Figure 44: The proportion of STATS19 casualties linked to SHIPS

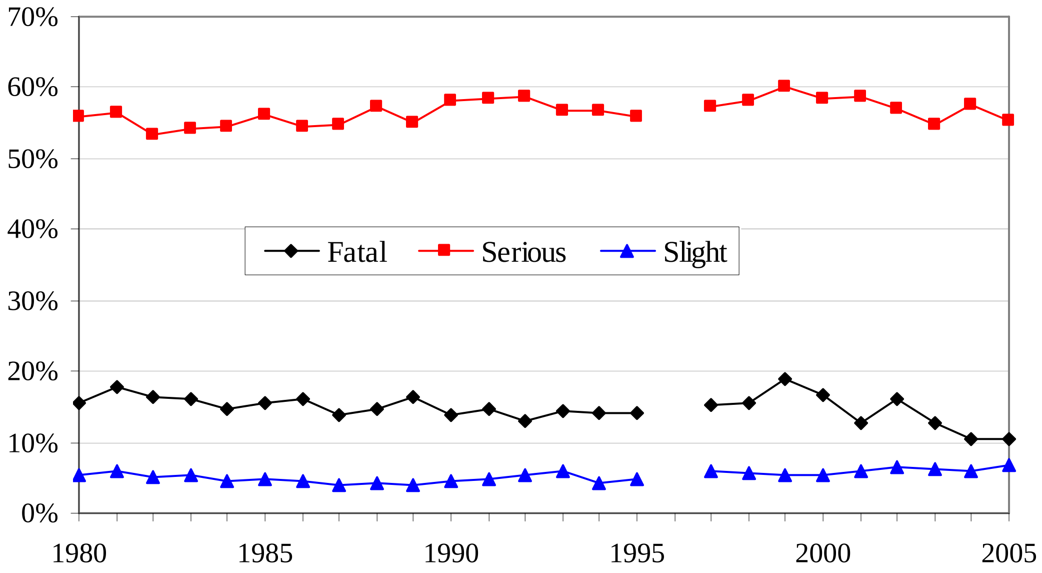


Figure 4 appeared in Section 4.1 of the main report. It compares the distribution of Length of Stay in the linked casualty data with three ranges: 0 days (admitted and left hospital on the same day), 1-3 days and over 3 days. This uses the definition of Length of Stay that was applied in the earlier study, for consistency, so the data from the current study have been recalculated. There are clear overall trends, with a shift towards shorter stays in hospital. The changes between 1991-95 and 1997-99 fit broadly within the overall trends, so it appears that the results of the new linkage are consistent with those of the original linkage in terms of Length of Stay.

Figure 5 also appeared in Section 4.1 of the main report, and presents the corresponding comparison for MAIS. This comparison is affected by the switch from ICD9 to ICD10. A new mapping from ICD10 to AIS98 developed at the University of Navarra (Apollo, 2006) is used, and it produced an appreciable proportion of MAIS 9 (unknown) scores. These appear to be generally minor injuries, so they have been included with MAIS 1 to prepare the Figure. The Figure shows major increases in the proportion of casualties with MAIS 1 between 1991-95 and 1997-99, and corresponding reductions with higher MAIS. If, less plausibly, the MAIS 9 scores are distributed pro rata among the known codes then the changes are a little less definite, but it is clear that the combination of the ICD10 codes and the new mapping has tended to yield lower MAIS scores. There is no way of telling whether the earlier or the later system yields the more reliable results, but this Figure does indicate that results based on mapping ICD9 codes to MAIS should not be compared with results based on mapping ICD10 codes.

The relationship between MAIS and Length of Stay is examined in Table 124 for the 1997-2005 period (MAIS 9 is included with MAIS 1, serious and slight casualties only are included). For example, 15% of MAIS 1 casualties left hospital on the same day in 1997-99, compared with 17% in 2003-05. The distributions vary markedly by MAIS, and there is a trend for shorter hospital stays at most MAIS levels.

Table 124: Distribution of casualties by Length of Stay at each MAIS level

		Length of Stay		
MAIS		overnight	1-3 days	>3 days
1	1997-99	15%	74%	11%
	2000-02	17%	74%	10%
	2003-05	17%	74%	9%
2	1997-99	4%	51%	45%
	2000-02	5%	51%	45%
	2003-05	5%	48%	47%
3	1997-99	1%	13%	86%
	2000-02	1%	12%	87%
	2003-05	1%	15%	84%
4-6	1997-99	1%	3%	96%
	2000-02	2%	6%	93%
	2003-05	1%	9%	90%

This analysis is extended in Table 125 to compare the distributions for the 1997-2005 period by road user type; pedal cyclist and other casualties are omitted because of the smaller casualty numbers. The Table shows that car occupants tend to be discharged slightly earlier than pedestrians at each level of MAIS, while motorcyclists tend to spend longer in hospital.

Table 125: Distribution of casualties by Length of Stay by road user type, 1997-2005

		Length of Stay		
MAIS	Road user type	overnight	1-3 days	>3 days
1	car occupant	18%	72%	10%
	pedestrian	13%	77%	10%
	motorcyclist	9%	79%	13%
2	car occupant	5%	51%	43%
	pedestrian	4%	48%	49%
	motorcyclist	4%	46%	51%
3	car occupant	1%	14%	85%
	pedestrian	1%	13%	86%
	motorcyclist	0%	11%	89%
4-6	car occupant	3%	6%	92%
	pedestrian	0%	7%	93%
	motorcyclist	0%	2%	98%

Cost-benefit analysis

One way of understanding the relative contribution of a particular group of casualties to the national total is via cost-benefit analysis. For example, Table 126 uses the British Government's cost-benefit value of prevention of road accidents to show the relative contribution of fatal, serious and slight casualties to the national cost of injuries. It includes the number of serious* casualties (MAIS≥3) estimated using the conversion factors from the STATS19/SHIPS linkage (Table 8). The official values are averages per severity, not by MAIS, and it is assumed in the Table that the value of a serious* casualty is twice the official value of a serious casualty. This is probably conservative, as this category includes those who died more than 30 days after the accident and those who suffered long-term disability.

Table 126: Cost-benefit value of prevention of road accidents, Great Britain, 2005

	Cost per casualty (£m)	Number of casualties	Cost (£m)	
Fatal	1.42846	3201	4573	38%
Serious	0.16051	28954	4647	38%
Slight	0.01238	238862	2957	24%
Total			12177	100%
Serious*	0.32102	6945	2230	18%

Cost per serious casualty estimated as twice cost per serious casualty*

The lower threshold of the serious injury category in Great Britain is set relatively low, but it appears that non-fatal casualties with MAIS≥3 account for at least one half of the burden of serious casualties, and one fifth of the overall injury burden.

7.8.5 Conclusions

The matching of SHIPS and STATS19 records that had previously been carried out at TRL for 1980-95 has been successfully extended to the 1997-2005 period. This involved developing an ACCESS database that applied the principals of the earlier matching procedure. The checks described above suggest that the two procedures are broadly consistent. Table 123 showed that the proportion of SHIPS records that could be matched to STATS19 has fallen steadily since at least 1993, and it is planned to investigate this further. One possible explanation would be that changes to reporting procedures and standards may mean that the system of tolerance levels needs to be revised.

An important question that must be considered is whether Scotland may be considered representative of the UK as a whole in terms of accident reporting, so that the conversion factors derived from Scottish data may be generalised to the rest of the country. The UK Department for Transport is currently carrying out a similar study to match English STATS19 records to in-patient data from the Hospital Episodes System, the equivalent for England of the SHIPS system in Scotland, so a well-founded answer may well emerge in due course.

For the present, one may observe that Scotland is typical of the UK in terms of traffic law and traffic conditions, and the police in Scotland are subject to the same operational pressures as the police in the rest of the country. On the other hand, although Scotland does contain major urban areas, overall it contains a higher proportion of rural and remote areas than do England and Wales. In terms of hospital admission and clinical procedures, Scottish hospitals operate within the framework of the National Health Service, although budgets are devolved and this may lead to some variation across the country.

Scotland accounts for about 9% of the UK's fatal casualties and 6% of all casualties, and the study has matched data from 9 years, so there is no reason on statistical grounds that the results should not be considered representative. Overall, one may tentatively argue that the conversion factors derived from Scottish data may be generalised to the rest of the country.

7.8.6 References

Apollo Project (2006).
http://www.unav.es/preventiva/traffic_accidents/pagina_5.html
University of Navarra – Apollo Project.

Association for the Advancement of Automotive Medicine (1998). *The abbreviated injury scale, 1998 revision.* Des Plaines, USA.

Broughton J, Keigan M and James F J (2001)*. Linkage of hospital trauma data and road accident data.* TRL Report 518. Wokingham: TRL Limited.

CODES. Reports available from
http://www-nrd.nhtsa.dot.gov/departments/nrd-30/ncsa/CODES.html

Keigan M, Broughton J and Tunbridge R J (1999). *Linkage of STATS19 and Scottish hospital in-patient data – analyses for 1980-1995.* TRL Report 420. Wokingham: TRL Limited.

Nicholl JP. *The Use of Hospital In-patient Data in the Analysis of the Injuries Sustained by Road Accident Casualties.* Supplementary Report 628, Wokingham: TRL Limited.

Simpson H F (1996). *Comparison of hospital and police casualty data: a national study*. TRL Report 173: Wokingham: TRL Limited.

Stone R D (1984). *Computer linkage of transport and health data*. Laboratory Report LR1130, Wokingham: TRL Limited.

World Health Organisation (1992). *International Statistical Classification of Diseases and Related Health Problems. Tenth Revision*. Geneva.