LIVE

Tools to injury prevention

Grant Agreement No MOVE/C4/SUB/2011-294/SI2.625804

D1
Data collection, analysis and recommendations

Status: Final

Prepared by:
António Couto, Carlos Rodrigues, José Pedro Tavares, Marco Amorim, Sara Ferreira

University of Porto

Date: **January 2014**

Project Duration: 27 July 2012 to 26 January of 2015

European Comission Road Safety EC co-funded projects

PREFACE

The research reported in this document has been conducted by the LIVE team, funded by the European Commission and is part of DG MOVE of the 7th Framework

The LIVE team consists of the University of Porto;

Co-ordinator: Sara Ferreira

António Fidalgo do Couto

Carlos Manuel Rodrigues

José Pedro Tavares

Marco Raul Amorim

This document was written by Sara Ferreira and Marco Amorim of the University of Porto, for the first task within the LIVE project.

Task 1 deals with the data treatment and analysis. Besides University of Porto, the LIVE project was supported by the partners in terms of providing data set that was fundamental to develop the project. The partners consist of the following entities;

Autoridade Nacional de Segurança Rodoviária (ANSR)

Hospital Geral de Santo António – Centro Hospitalar do Porto, EPE

Centro Hospitalar de Vila Nova de Gaia/Espinho, EPE

Centro Hospitalar São João, EPE

Hospital Pedro Hispano

Instituto Nacional de Emergência Médica (INEM)

ABSTRACT

The Deliverable describes the development within the LIVE project of a methodology for the record linkage between police data and hospital data, using the case study of the Porto region, Portugal.

The importance of this process is well recognized by institutions such as IRTAD (OECD/ITF) that are promoting the combination of various data sources to fully assess the consequences of road accidents and monitoring progress. The complexity of this process is mainly concerned with several issues found in the data sets; mistakes and missing values are frequently detected and there are only a few common data fields that can be matched by the linkage process.

The deliverable report explores the data issues usually present in this kind of data set that may influence the results of the linkage process. It is common in several countries/regions that each entity (e.g. hospital) presents a different data system, which imply several and distinct data treatments depending on data characteristics.

There is already some international work done in this field, which is briefly described in this deliverable report. As it was concluded in those studies, developments and/or improvements are still needed.

In Portugal, thus far, there is no database connecting the accident police data set and the injuries medical hospital information in which linked records may support several studies of the relationship between the accident sites' characteristics and the injury severities.

Under the LIVE project, for the first time, an approach was developed to link police and hospitals data sets covering the Porto region. In this work, a mixed deterministic and weight-based probabilistic method to link the police and hospital records was used. The results obtained lie within the range of rates found in other studies. Additionally, to improve the record linkage results, a validation process based on the emergency ambulance data was performed. Despite the missing values, it was possible to check 98% of the matched records as true matches. Finally, a preliminary investigation of bias after data linkage is described, showing that the variables selected for comparison indicate similar statistical values. In addition, a brief statistical analysis is presented using the linked records.

Considering that one of the main objectives of the LIVE project is to assess the application to the Portuguese context of the common definition already established by the European Union - the maximum abbreviated injury scale (MAIS), a description of the actual Portuguese background is presented. Ongoing efforts are being carried out to reach this final objective.

The road accident linkage process is clearly and deeply described in this report in order to be adapted, developed and applied in different contexts, aiming to promote future developments on police, hospital and emergency ambulance data either in Portugal or in other countries.

CONTENTS

ABREVIATIONS

AIS	Abbreviated Injury Scale
ANSR	Autoridade Nacionalde Segurança Rodoviária (National Authority for Road Safety)
CNPD	Comissão Nacional de Proteção de Dados (National Commission of Data Protection)
DRG	Diagnosis Related Groups
F	Fatality
FEUP	Faculty of Engineering, University of Porto
ICD	International Classification of Diseases
INEM	Instituto Nacional de Emergência Médica (National Institute of Medical Emergency)
IRTAD	International Traffic Safety Data and Analysis Group
ISS	Injury Severity Scale
ITF	International Transport Forum
MAIS	Maximum Abbreviated Injury Scale
OECD	Organisation for Economic Co-operation and Development
SEI	Serious Injury
SLI	Slight injury

1. INTRODUCTION

1.1. Road Safety data: the need for more information on traffic injuries

International comparisons of road safety are limited to data on fatal accidents and fatalities. However, these data are not representative of all road safety problems and also, of the impact of the accidents on citizen's lives throughout serious injuries which in fact, can have a serious impact on their lives with considerable associated economic costs.

The OECD/ITF recommends that "a complete picture of casualty totals from road accidents is needed to fully assess the consequences of road accidents and monitor progress", as reported in the IRTAD report (*1*). To achieve this goal, accurate and complete data are essential. Nonetheless, it is well recognized that such information is rarely possible from a single data set, proved by a large work done in the field of record linkage, especially by public health researchers (*2-16*). Accident data collected by the police officers at the accident scene contain extensive information about the circumstances and location of the occurrence. These data allow safety researchers to analyze and investigate the accident characteristics to identify possible causes and posterior safety countermeasures. Despite that, limitations are found especially when people are injured in the accident (*4*). In fact, no single data set provides enough information to give a complete picture of road traffic injuries and to fully understand the underlying injury mechanisms.

Recently the European Union (EU) Member States identified a common medically-based definition of serious injury to be used in road safety statistics. The new EU common definition is based on the Maximum Abbreviated Injury Scale (MAIS). The MAIS derived from the Abbreviated Injury Scales (AIS), which is a severity score of a casualty with several injuries. On the other hand, the AIS can be derived from the commonly used International Classification of Diseases (ICD) codes. Therefore, the MAIS score allow assessing injury severity on the basis of a standardized medical indicator.

A consensus has emerged over adoption of the MAIS equal or greater than three as the EU definition of serious road-accident injuries. To produce comparable statistics in line with this EU common definition, authorities in EU countries are advised to do one of the following:

- bring together the relevant information from both police and hospital records
- use only hospital records
- use police records but make a correction for under-reporting

In most countries, the injury classification is based on on-site judgment, and the information on accident severity reported by the police is rarely checked with the medical records, except when the injured person dies at the hospital (*1, 4*). Hospital records have the potential to improve the injury accident data when combined with the police data. More detailed information on injury types and severity is reported by the hospitals, improving the data from the accident scene and providing a more comprehensive view of road accidents (*4*). Therefore, a linkage process to match records from different sources is needed in order to allow the injury classification based on MAIS as well as to provide a deep knowledge about road safety.

1.2. The LIVE project

The LIVE project is concerned with the method of record linkage of both police and medical data in order to obtain a complete picture of the victims of the accidents. The know-how provided by this process will allow to define measures embracing the road environment, particularly their influence on pedestrian accidents, but also to promote an efficient management system of the emergency services.

To develop the record linkage process in the Portuguese context, the project LIVE requires the collaboration of several entities such as the National Road Safety Authority (ANSR), which store the police records, and the hospitals. Thus far, in Portugal there is not a systematic process to obtain, link and analyze data from both ANSR and hospitals. This means that several data issues are expected to be found, leading to constraints in terms of linkage process. A report of these issues is needed in order to provide knowledge for future data improvements and therefore, making the linkage process simpler and more efficient. Considering the data issues, the linkage method to match records from both data sources is selected and evaluated. In addition, a methodology to assess the validity of the process should be considered, as the one presented in this report.

The project LIVE is engaged to provide a complete methodology to link police with hospital data in order to provide knowledge for future improvements of the data system and thus, developing the road safety research and practice.

1.3. This report

This report forms the Deliverable output of Task 1 “Data collection, analysis and recommendations” of the LIVE project. It sets out the results of the research team’s consideration of issues and methodologies relating to the record linkage process.

Since the LIVE project comprehends the first complete study related to the record linkage of different data sources to road safety analysis in Portugal, it was necessary to promote the topics and goals in particular in the health sector. A far as we know, the injury classification of the victims of an accident based on an injury scale, as the AIS, is not commonly known and applied at the health sector. Nevertheless, the diagnosis of the victims is described according to the ICD codification.

The initial stage of the project embraces the collection of the police and hospital data. In fact, a large quantity of data was collected and treated in order to standardized and prepare the record linkage process. Several issues on hospital data and police data were found.

Because clinical records include particularly sensitive information and need high protection, a written consent was required by the Ethic Commissions of the hospitals. Therefore, the National Commission of Data Protection (CNPD) has had to express its agreement and considering that any kind of directly personal identifiers such as name or ID card number were not provided. Therefore, the deterministic method for matching records was not an option.

Building on the previous facts, this report sets out an approach to perform a record linkage process. Also, considering the state-of-art in terms of linkage record methods as well as the IRTAD recommendations, the approach includes the following stages:

- Data preparation
- Selection of linkage variables
- Evaluation of process feasibility
- Computation of simple weights
- Restriction of comparison pairs (blocking)
- Comparison stage (matching)
- Simple weights assignments
- Computation of composite weights
- Decision stage (linking)
- Threshold determination
- Review of dubious pairs

In addition, a methodology for the validation of the linked records is described in this report and the results are statistically analyzed. Ongoing studies under Task 2 are also summarized. At the final, conclusions and recommendations are provided.

2. THE CHALLENGE OF POLICE AND HEALTH DATA LINKAGE

2.1. Introduction

The main positive driver for the LIVE project is the need to build for the first time in Portugal a record linkage process of both health and police data. From the linked records, several studies may be provided, some of them for the first time too. For instance, it will be possible to compare the number and type of injuries' severity recorded by the police, which are used to national statistic reports, with those recorded by the hospitals. Also, the injury severity will be for the first time analyzed in terms of using alternative injury scales.

This chapter explains the basis of the problem related to the linkage record process by identifying the possible issues that may be found in data sets and by reviewing some countries' experiences.

2.2. The basis of the problem

Several issues are expected to be found throughout the linkage process, as reported by IRTAD (*1*). Differences on the data collection system between entities (e.g. hospitals), misreporting and under reporting are issues that require a data treatment process to standardize all data sets and increase the potential of the linkage algorithm. Besides, a linkage process is more dependent on the quality of the data rather than the linkage algorithm itself.

The linkage process is developed using data from police and hospitals. In Portugal, police are obliged to go on the scene of an accident where there is at least an injured person. The police officer completes an accident form (on paper) on the scene of the accident, which they complete later when they return to their office (usually using a digital platform). Police data provide detailed information about accident circumstances, location, vehicles involved, and victim's characteristics such as age and gender.

The information on victims' severity reported by the police is rarely checked with the medical records, except when the injured person dies in the hospital (*1, 4*). In Portugal, only since 2010 the hospitals have to inform the Public Ministry which in turn informs the police when an accident victim dies within the 30 days following the accident, matching the international definition for fatality. Then, the police inform the ANSR which stores the victim information. Before 2010, the procedure to set the official number of fatalities was based on the fatalities declared at the accident site, adjusted by a coefficient of 1.14. Based on this recent process of information transmission, the ANSR (*17*) compared the number of fatalities recorded by the police before and after the update by medical information for the years 2010 and 2011, showing different adjustment factors: 1.26 for 2010 and 1.29 for 2011. In terms of injury severity, the diverse criteria and the non-medical background of the reporting police officer produce less than precise data to adequately study injury outcomes (*4*). Few reports on injuries and severity of accidents have been published, giving biased information on the actual safety performance.

2.2.1. *Police data issues*

Police records presents, in general, several issues such as, the police do not collect information on all non-fatal accidents or may never know of an accident in a rural area or in a private road causing small damages or slight injuries. Additionally, the accident records exhibit mistakes and omissions, sometimes related to important accident information. Despite police officers being aware of the official injury classification (see chapter 2.3 to the definition injury severity), decisions on the severity category of the victim based on on-site judgment might bias police official classification.

2.2.2. *Hospital data issues*

The hospital data may include emergency medical data and/or hospital admissions data. The former is usually less developed, but may have potential to provide more information also about slight injuries. Note however, that health records are collected for medical and hospital administration (financial) purposes. This fact may contribute for several issues such as little requirement for field indicating road accidents, medical staff do not always treat data entry as high priority, data systems may differ from hospital to hospital (and even within hospital units), data may not be gathered into a national system, and hospital practices also may vary from place to place in the same country. In addition, ethical concerns about releasing confidential medical information may exist.

Additionally, changes in practice over time, and variations by country and within countries are usually applied to police and hospital data. Note that some of the previous issues will be less significant with more serious casualties.

As pointed out above, the commonly available records used in this type of data process contain many errors or omissions. Nevertheless, this data issue can be overcome by applying techniques to complete the linkage process, such as multiple imputation (*18*).

It should be stressed that independently of the quality of health data, it does not substitute the police data but supplement it.

Other sources besides police and hospitals may complement the injury information and/or may be used for a validation process. Other sources of information and data on fatal accidents are ambulance services, mortality registers, forensic reports, insurance records, fire services, etc.

2.3. The significance of linkage process

Linkage of accident details in police accident reporting systems with injury details in hospital records makes the best use of both data sources, allowing several studies such as a complete figure of the total number of casualties and type of injury severity, and an in-depth understanding of the medical consequences of particular types of accident.

Most common research and practice on road accidents rely on police reports, which, in most countries, classify accident victims into three categories of injury severity: Slight Injury (SLI); Serious Injury (SEI); Fatality (F). This is the case of Portugal, where the following definition of victim severity is considered:

- SLI if the person stays in hospital for no more than 24 hours;
- SEI if the person stays in hospital for more than 24 hours;
- F if the person dies within the 30 days following the accident as a result of the suffered injuries.

The basis for an injury classification is usually the on-site judgment of the investigation police officer and eventually possible subsequence information provided by hospitals; mainly inpatient time and deaths.

However, there are no commonly agreed definitions on injured road casualties; e.g., seriously injured may be classified by the type of injuries, the length of hospitalizations, etc; the KABCO scale is used in the Unites States of America and encompasses six severity levels (K = Killed, A = Incapacitating Injury, B = Non-Incapacitating Injury, C = Possible Injury, O = No Injury, and U = Injured, Severity Unknown). Therefore, the variability in criteria and the nonmedical background of the reporting agents can lead to data inaccuracy and misreporting.

An alternative to overcome this injury classification data inaccuracy is to use the ICD codes, an international codification derived from the medical diagnosis to describe the injuries.

Even though the fact that injury severity classification based on medical diagnosis may be obtained only by the hospital data, it is clear that using only this data important information related to the accident circumstances is neglected. Also, underreporting cases are not checked, holding the reality unknown. In fact, as it is concluded in this study, not always the inpatient is reported as traffic accident as the external cause.

Using linked records several studies may be conducted such as the analysis of the influence on the severity of the use of restraints and protections systems, driving and blood alcohol concentration, road and highway design. In addition, studies related to the emergency medical services may be developed in order to optimize the attendance service.

Overall, when using a standardized medical indicator, it is possible to classify the injuries of road accidents by an international criterion. By linking police and health sector data a reliable number of injuries can be identified by comparing the number of injured road users treated in hospitals to the number recorded by the police.

2.4. A summary of experiences on police and health data record linkage

As reported by IRTAD (*1*), many countries have carried out record linkage procedures of police and health data for a number of years. Some carry it out periodically, as is the case in some states in the United States or in regions of Australia.

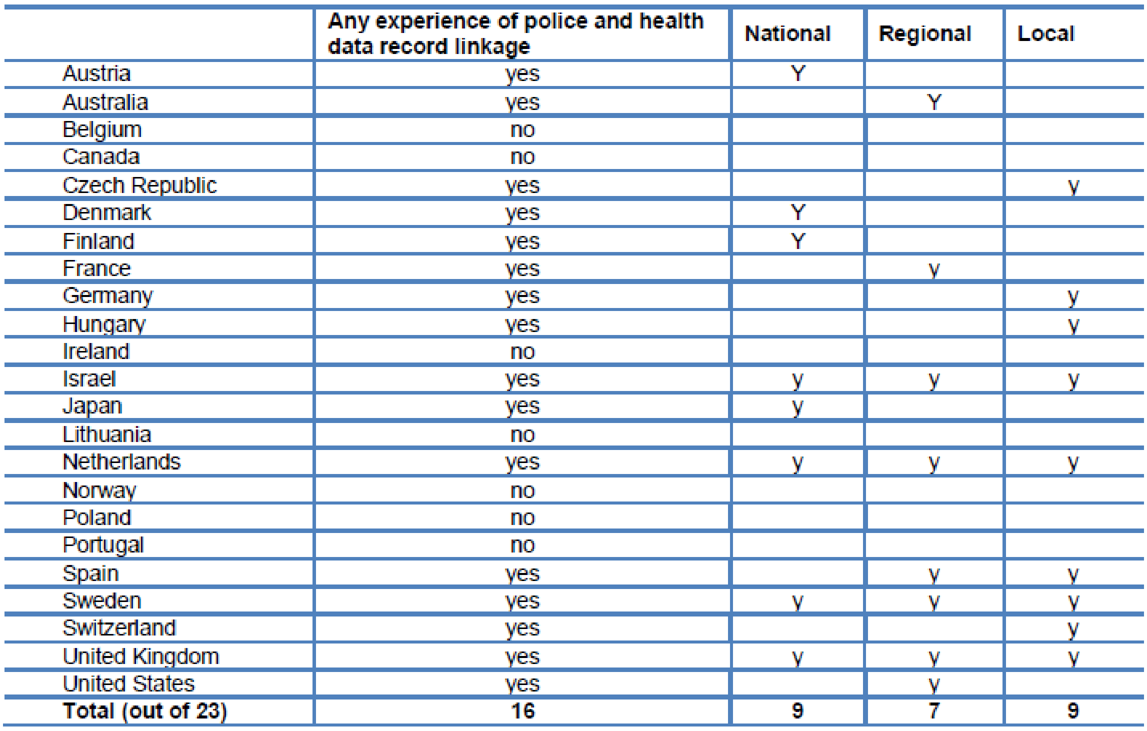
Record linkage is a process commonly used to link data sets from different sources from the public health area. According to Winkler (*19*), record linkage started with the work of Howard Newcombe in 1959 and Fellegi and Sunter in 1969, and since then has been widespread in public health and epidemiology. There are three main methods to link data sets: manual, deterministic and probabilistic. Manual linkage is the simplest and most valid method (*5*), but is usually impractical due to the large number of records. Nevertheless, even this method may not be considerably successful, as shown by a study conducted in Russia (*11*) applied only for fatalities, which results correspond to a linkage rate of 44.5%. The deterministic method links records with an exact match of the identification variables (also called data fields) thus it is subject to the quality of these variables. The method can be used with high quality unique identifiers such as personal identifiers. However, this kind of information is not frequently available because personal identifiers are considered to be confidential information and personal data protection rules are in force. Thus, most record linkage efforts are based on a probabilistic method such as the probabilistic weight-based or the probabilistic distance-based methods. Both methods require a previous data examination, including the selection of the linkage data fields. The number and type of linkage data fields influence the quality of the results because there are data fields likely to be a very accurate match value, yet with low weight power (e.g., the gender).

There are greater differences in the methodology used for record linkage. Those differences are mainly concerned with the number of variables used to link, the methods for record linkage (manual, deterministic, probabilistic) and the validation process.

Some countries have access to initials of the surnames and names, other do not have personal identifiers due to confidentiality and personal data protection. Depending on the number and type of variables in common between police data and hospital data, the record method is selected. Usually age, gender, date and hour of collision and hospitalization are variables used to link data. The most common methods are deterministic with some degree of tolerance and probabilistic based on weights or distance-based. Also, a combination of deterministic and probabilistic may be used.

In terms of validation process, clerical review, calculation of epidemiological criteria or simulation using a sample, is the methods reported only in a few studies. In fact, several limitations exist to apply a consistent validation process due to the lack of other data sources.

Figure 1 shows the experience of the IRTAD countries in terms of police and health data record linkage. Note that Portugal does not have any experience in this kind of process.



	Any experience of police and health data record linkage	National	Regional	Local
Austria	yes	Y		
Australia	yes		Y	
Belgium	no			
Canada	no			
Czech Republic	yes			y
Denmark	yes	Y		
Finland	yes	Y		
France	yes		y	
Germany	yes			y
Hungary	yes			y
Ireland	no			
Israel	yes	y	y	y
Japan	yes	y		
Lithuania	no			
Netherlands	yes	y	y	y
Norway	no			
Poland	no			
Portugal	no			
Spain	yes		y	y
Sweden	yes	y	y	y
Switzerland	yes			y
United Kingdom	yes	y	y	y
United States	yes		y	
Total (out of 23)	**16**	**9**	**7**	**9**

Figure 1 - Experience of police and health data record linkage IRTAD countries 2010 (Source: (*1*))

2.5. Conclusions

This chapter has provided a brief description of the issues of the police data and health sector data that may exist when using a record linkage process. These issues may be a constraint to the performance of the record linkage process, influencing the selection of the methodology to be applied. Therefore, a deep study of the data issues and its treatments are suggested in order to achieve better results.

Despite the effort that this kind of study implies, it is clear the significance of the record linkage process in road safety field. It is well recognized that linking records from multiple data sources is a promising method for injury surveillance evaluation. Analyses of this nature are necessary for identification of more effective countermeasures aimed at minimizing injuries that result from traffic accident.

Each country/region and each institution have their own data system and their own data protection rules, implying that a common methodology for record linkage is impossible to be defined. Nevertheless, data availability restrictions, as well as data issues, should be analyzed and reported in order to share a standardized method that, despite the data differences, may lead to a common data treatment strategy.

The methodology developed under the LIVE project to standardize the medical and police data in order to be possible to link each other may be of interest for similar cases. In fact, there are many other countries that until date have no experiences with record linkage.

In this context, we hope that the experience described in this report may be of interest to future users in Portugal or in other country with similar data characteristics.

3. THE PORTUGAL EXPERIENCE – PORTO REGION CASE

3.1. Introduction

Despite the number of road accidents in Portugal has decreased by 74%, the Portuguese average continues to be higher than the European average in terms of the number of fatalities per million inhabitants. In the field of injuries many actions remain to apply in addition to the efforts already done in the road safety field. However, it is not possible to define actions/strategies in an efficient manner if the reality is unknown.

Porto city is the second-largest city of Portugal in which the administrative limits (an area of 41.66 km²) include a population of 237,584 inhabitants distributed within 15 civil parishes (2011 data). The urban area of Porto, which extends beyond the administrative limits of the city, has a population of 1.3 million (2011) in an area of 389 km2, making it the second-largest urban area in Portugal. The Porto Metropolitan Area includes an estimated 2 million people. After Lisbon, the capital city of Portugal, Porto city exhibit the higher number of accidents with victims (945 number of accidents reported in 2011).

Until date, no experience in record linkage process exists in the Portuguese context. In effect, only few countries have experienced the linkage of police and hospital data. Therefore, from the results obtain for the Porto's region, it will be possible to assess an estimation of the real situation of road safety in Portugal. In addition, the methodology present in this report may be extrapolated to the whole country, thus enlarging the actual injury statistics.

In this chapter, a full linkage process is presented, considering the various steps usually associated to this process (see 1.3.5). The final results are analyzed in terms of bias after data linkage and validation, in order to investigate in which extent the results are consistent for policy-making and planning of injury prevention. Recommendations to improve road accident data collection and storage may be promoted from the outcomes of this research.

3.2. Data collection and assessment

The study was conducted by linking records of accidents occurred in the region of Porto between 2006 and 2011, recorded by the Portuguese Police of Public Security and sent to the ANSR. The studied region includes Porto, Maia, Gondomar, Vila Nova de Gaia and Matosinhos cities. Porto city is the second largest city in Portugal where two central public hospitals are located. Each hospital covers specific areas/cities of the region, and specific healthcare assistance by certain medical specialties.

Table 1 shows, besides other entities, the four hospitals that have collaborated with this study by sending the data sets. Afterward the Hospital Pedro Hispano was excluded from the study because the data set includes only 1.5 years, which correspond to 26 injuries at the emergency unit and 11 injuries at the admission unit, due to a data system changing. Also, in the case of Centro Hospitalar de Vila Nova de Gaia/Espinho, EPE (hereafter designated by Hospital VNG), it was found two significant limitations in the data set: (1) only about half of the cases identified in the emergency unit as requiring admissions was collected as inpatient; (2) no fatalities were recorded in the admission unit data set. The first issue results from the fact that the two units have independent data systems, with no data field in common (to match each other). Therefore, only the admission cases, which the external cause was identified by the ICD (E810 to E819), was sent to us. The second issue results from the fact that no data field exists in the data set to report the death of the inpatient. Also, it should be stressed that in the data set of the admission unit of the Hospital VNG might exists several missing inpatient records since this hospital do not have all medical specialties. These issues commit the under and the misreporting evaluation, however several studies can still be done if the sample obtained proves to be random.

In fact, each hospital has different data systems. The Centro Hospitalar São João, EPE (hereafter named Hospital SJ) separates the records of emergency and admission units in different data sets, while the Hospital Geral de Santo António – Centro Hospitalar do Porto, EPE (hereafter named Hospital SA) provides already linked unit records. The Hospital SJ is the largest one (general hospital covering a wide area) in Northern Portugal and thus, it is common to receive inpatient transfers from other hospitals.

The ANSR has provided the police records of all studied region. The police records in Portugal, as in most countries, present several issues as described before.

Portugal has a national data set of the ambulance emergency service gathered by the National Institute of Medical Emergency (INEM). Ambulance operators normally record their interventions whenever they are called to assist a person injured on the road, registering the information of the accident location and the hospital where the victim was taken to. However, this source is more likely to be accurate in the case of serious and slight injuries, but not for fatalities, because some of the victims die at the accident scene from where they are removed by other entities. This type of road accident information is not usually possible to obtain and therefore, the linkage process is in general focused on police and hospital data sets. Nevertheless, the ambulance service data may be used to validate the matching records resulting from the linkage process, as suggested by IRTAD (*1*). In this study, the INEM has provided the ambulance emergency data of Porto city covering the year 2010. Despite its own limitations, ambulance data may be a good alternative to validate the linkage process by overcoming the problem of personal identifiers. According to the IRTAD report (*1*), this kind of data set was never used before, which reinforces the interest of this study.

In sum, at the end of the data collection process, five institutions were included in the study, representing a total of seven data sets; Hospital SJ and Hospital VNG provided two data sets each. The involved institutions have different interests and perspectives in the accident injury occurrence, reflected by the differences in their data sets, resulting in a complex data system that may produce a low linkage rate. To reduce this risk, data preparation is needed.

Table 1 - Description of the sources of road victims' information

Source of information	Entity	Area covered by entity	Stage of data analysis process
Police data	ANSR	All	Record linkage
Hospital data (emergency and admission)	Hospital Geral de Santo António – Centro Hospitalar do Porto, EPE	Porto and Gondomar	Record linkage
Hospital data (emergency and admission)	Centro Hospitalar de Vila Nova de Gaia/Espinho, EPE	Vila Nova de Gaia and Espinho	Record linkage
Hospital data (emergency and admission)	Centro Hospitalar São João, EPE	Porto and Maia	Record linkage
Hospital data (emergency and admission)	Hospital Pedro Hispano	Matosinhos	Excluded
Ambulance emergency data	INEM	Porto	Validation and calibration

This study is focused on the region of Porto. The police accident reports were limited to the boundaries of the cities described above, and the hospital records were obtained from the three hospitals located in Porto and Vila Nova de Gaia cities.

Personal identifiers and victims' names were not provided in order to comply with data protection rules. This handicap on the data requires the establishment of a methodology that may be used worldwide, thus the matching variables are those common in this kind of data sets.

3.2.1. *Police data description*

The police records have the following information: accident ID, time and date of the accident, location (street and parish), victims' age, gender and role (passenger, driver or pedestrian), and the number and severity of victims per accident. In a second stage of deliverable data, the ANSR sent more information related to the accident and its victim(s) such as weather and lighting conditions and road maintenance, alcohol-blood test, safety equipment, type of manoeuvre/action and driver license.

The data set from police provided by the ANSR contains a total of 18,529 victim's records in a total of 14,155 accidents in the Porto's region. The records are shown in Table 2 divided by year (from 2006 to 2011) and by victims' severity. This table shows an increase of the number of accidents in 2009 and 2010 however, the injury severity tends to diminish.

According to the police data, as shown in Table 3, the severity of pedestrian injuries is higher than the severity of the injuries of the drivers and passengers. Considering that 1 represents male victims, 3 represents female victims and 2 represents missing gender information, Table 3 shows that driver victims are usually male while passenger and pedestrian victims are usually females. Also, the average age of pedestrian victims is higher than the average age of the passenger or driver victims; this may suggest that the accidents involving pedestrians have higher impact on elderly people.

Table 2 - Number of injuries and accidents per year from the police accidents' records

Year	Fatalities	Serious Injuries	Slightly Injuries	Number of Accidents with victims
2006	24	119	2790	2184
2007	36	98	2750	2191
2008	25	104	2874	2248
2009	17	85	3203	2552
2010	28	60	3199	2548
2011	21	86	3010	2432
Total	151	552	17826	14155

Table 3 - Number of injuries per type of victim from the police victims' records

Road User Type	Average Age	Average Gender	Average Severity	Fatalities	Serious Injurie	Slightly Injuries
Driver	38	1.66	0.045	83	263	9247
Passanger	34	2.26	0.023	13	90	4917
Pedestrian	45	2.12	0.079	55	199	3662
Total	38	1.92	0.046	151	552	17826

3.2.2. *Hospital data description*

The hospital emergency service records have information on the victims' age and gender, date and time of arrival at the hospital and final destination. The inpatient records describe the victims' gender and age (birthdate in the case of Hospital SA), date of admission, date of release, destination after discharge, diagnosis-related group code (DRG), several diagnosis and treatment codes, and description of the victims' condition. Table 4 shows the number of victims that were reported in the studied hospitals, separated by emergency and admission units.

As can be concluded by Table 4, Hospital VNG recorded the higher number of victims at the emergency unit, and Hospital SJ recorded the higher number of inpatients. The former may be the consequence of the Hospital VNG covers an area that includes the city with more inhabitants in the North region, and the latter may be because the Hospital SJ is a general hospital.

Table 4 – Number of victims' records by unit of each hospital

	Number of victims' records	
Hospital	Emergency Unit	Admission Unit
Hospital Geral de Santo António – Centro Hospitalar do Porto, EPE (Hospital SA)	5652	904
Centro Hospitalar de Vila Nova de Gaia/Espinho, EPE (Hospital VNG)	10963	590
Centro Hospitalar São João, EPE (Hospital SJ)	9370	2541

3.2.3. *Ambulance data description*

The provided ambulance records have information on the gender and age of one victim of the traffic accident, date and location of the accident, destination hospital, and time of arrival to the accident location. Until date, it was not possible to obtain information on the time of arrival to the hospital, which would have been helpful to assess the tolerance of the time matching variable used in the linkage process (see chapter 3.3.3 and Table 5). Unfortunately, the data sent by the INEM is related only to the year 2010, and in some records with missing information. Despite the INEM has a data bank, there are several limitations when using it, which constrain the data treatment and analysis.

Overall, as it can be concluded by the description of the several data sets used in the study, data treatment and standardization is necessary to increase adaptability and compatibility. In order to provide reliable data for the linkage algorithm, specific data issues are described in the next section.

3.3. Linkage process

3.3.1. *Standardizing*

The first step to standardize the data fields from different sources was to create an internal identification number composed by 4 groups for each record, as shown by the following example: PLC12N00018Y06. The 3 first letters identify the record source (record source code: PLC – Police; EMR – Emergence service; INP – Inpatient service; INM – INEM ambulance service). The second group consists of a number with 1 or 2 characters identifying the origin of the record. For police records, the number refers to the city where the accident occurred; in this case "12" stands for Porto. For hospital records, the number indicates the hospital (e.g. hospital code: 2 – SAH; 3 – SJH). The following group consists of a number with 5 characters preceded by the letter "N", being a time ordered number that identifies each record in its data set source. Finally, the last group represents the record year (character "Y" plus the two last digits of the year). This code allows calling back the original data when standardizing and compressing every variable field needed for the linkage process.

Each variable name is followed by the data source code, e.g., "AGE_PLCCODE*", in order to identify its source. Since birthdates are not available for every record, age has been used instead, being rounded to the closest lower integer value in the case of birthdates records (variable AGE). For the gender (variable GENDER), a dummy variable was used (1 – male; 2 – female). The use of a numerical value to identify the victims' gender allows a fast and direct way of calculating stats regarding this variable (e.g., a gender mean value of 1.5 indicates the same quantity of each gender). However as explained before, during the data treatment, and in order to deal with the gender missing data, the gender field was redefined by using 3 values: 1 to identify males; 2 to identify missing gender information; 3 to identify females. In the matching process, a tolerance of +/- 1 was allowed, accepting a link between known genders and missing gender information.

All the relevant dates (INDATE – date of entrance of the victims in the source records) were converted to an integer scale, the value 1 representing the first day in the period under analysis (January 1st, 2006). This option simplifies future calculations and eliminates possible errors due to the irregularities of the calendar (month size, year size). For the "time" variable (INTIME), a similar procedure was applied. Every hour is presented as a decimal part of the day multiplied by 100 and rounded to an integer value. However, an accident can occur close to the end of the day and the victim can arrive at the hospital only the day after. Thus, time must be continuous to allow date comparison. To give continuity to the time variable, the INDATE value multiplied by 100 was added to the previous integer value, as described in Equation 1. This equation creates a time ordered sequence, where each 100 units represent one day and each unit represents approximately 14 minutes.

$$\text{INTIME}=\text{time(minutes)}/1440\times100+\text{INDATE}\times100 \quad (1)$$

An extra standardization of the victims' classification (SEVERITY) was performed, but it was not used in the matching process. A simple discrete ordered scale from 0 to 2 has been built (0 for SLI, 1 for SEI and 2 for F), allowing a fast statistic assessment of this variable. It should be noted that such scale does not intend to be a measure of severity itself.

Furthermore, the data were prepared by searching for missing fields and duplicated entries. Missing fields can be easily found using a filter for blank fields or uncommon values (e.g., 01-01-1900 usually appears when a date field is unknown or not checked). A linkage process matching the data set with itself reveals possible duplicated entries. Duplications were cleaned and missing fields were either turned into unknown variables or deleted in the cases of insufficient data for linkage. Both hospital and police records have few of these problems. Furthermore, the ambulance service presents several records without information on age, gender, or hospital destination, thus being eliminated from the set. Because this data set is used merely for validation and calibration, the elimination of some records does not produce any bias.

3.3.2. *Feasibility*

Before proceed with the linkage methods, some statistical tests were run to assess the data set feasibility for matching. Roos and Wadja (*2*) presented a simple first statistical approach for this evaluation. The method consists in multiplying the number of categories by each variable and comparing it with the sum of the records from both data sets. Both police and hospital data sets share information on the victims' age and gender, and the date and time of the accidents. Thus, to calculate linkage feasibility, the number of categories within each variable must be assessed according to Equation 2:

$$2(\mathrm{GENDER})\times 365\times 5(\mathrm{INDATE})\times 12(\mathrm{INTIME})\geq \mathrm{recordsPLC}+\mathrm{recordsEMR} \qquad (2)$$

Cook et al. (*7*) proposes a method that improves the feasibility analysis by looking into the information contained in each variable. This method calculates the weight needed to achieve a specific probability of two records being true matches. The desirable probability *p* of selecting a true match corresponds to a weight *w* given by Equation 3.

$$w = log_2\left(\frac{p}{1-p}\right) - log_2\left(\frac{Em}{PLCrec\times EMRrec - Em}\right) \qquad (3)$$

where

p = probability of a match being a true match,
Em = expected matches,
PLCrec = number of police records, and
EMRrec = number of hospital records.

Statistical assessment on the INEM (ambulance) reports shows that for Porto's city 98% of the victims were transported either to Hospital SJ or Hospital SA, for Maia city, 97% of the victims were transported to Hospital SJ, and for Gondomar city, 88% of the victims were transported to Hospital SA. These values show that despite each hospital has a specific coverage area, in some cases the victims might be transported to another hospital. The numbers of victims recorded by the INEM were 6,604, 2,795 and 2,935 for Porto, Maia and Gondomar cities, respectively, which sum a total of 12,334 victims. However, the expected matches are only those victims that were transported to the respective hospital that covers the area of the accident site. This leads to a total of 11,766 expected match records.

The police data set had a total of 12,334 records within the cities of Porto, Gondomar and Maia and the two hospitals have a total of 15,022 emergency entries. The use of Equation 2 shows that the referred variables are able to accommodate the existing records, evidencing the feasibility of the linkage process. A total of 11,766 matches are expected (*Em*). The application of Equation 3 leads to a weight (*w*) of 18.19.

Because the total weight of the variables 18.62 (Gender = 0.69, date = 7.69, age = 3.91 and time = 6.33) (*7*) is greater than the required weight for the desirable probability, it is possible to conclude that the linkage process is feasible and true matches occur when all variables match.

3.3.3. *Model*

This methodology aims to be reproduced in other countries, namely in Europe, with a maximum autonomy and a minimum human intervention. Special care was taken in using as common and few variables as possible. Validation of the results and calibration of some parameters are required. As previously mentioned, the matching variables rely on the time and date of occurrence, and on the victims' gender and age.

Because the data set has no common identifiers, the correspondence in GENDER, AGE, INDATE, and closeness in INTIME is required to accept a match as true, leading to a mixed method where GENDER and AGE enters as a deterministic variable, INTIME as a probabilistic variable, and INDATE as a blocking variable. Blocking exists to reduce the number of comparisons and increase the efficiency of the linkage process. The data set was subdivided into a set of mutually exclusive subsets (blocks), assuming that no match can occur across different blocks (*20*). The blocking variable defines the category of the block; in this study, the accident date was chosen, grouping records of the same day (a 1 day tolerance was considered).

The matching model was formulated as follows:

```
SCORE=0
For all $0 \leq \text{INDATE\_EMR*} - \text{INDATE\_PLC*} \leq \square_{\text{indate}}^{I}$
SCORE = SCORE+ $\beta_{\text{indate}} \times \alpha_{\text{indate}}^{I}$

    For deterministic variables
        if $i\_PLC - i\_EMR \leq |\square_i^{I}|$
        SCORE = SCORE+ $\beta_i \times \alpha_i^{I}$
        If not SCORE = SCORE− $\beta_i \times \alpha_i^{I}$

    For probabilistic variables
        (option 1)for all $i\_EMR \geq i\_PLC$ or
        (option 2)for all $i\_EMR \leq i\_PLC$ or
        (option 3)all $i\_EMR \wedge i\_PLC$
            if $i^{C}\_|EMR - i^{C}\_PLC| \leq \square_{\text{intime}}^{C}$
            SCORE =SCORE + $\beta_i \times \alpha_i^{C}$
            From C+1 till $C^{\text{end}}$
            Else if $i\_|EMR - i\_PLC| \leq \square_i^{C+1}$
            SCORE =SCORE + $\beta_i \times \alpha_i^{C+1}$
            If not SCORE=SCORE
        End cycle
End cycle
Order SCORE (greater to lower)                                    (4)
```

where

PLC = set of source records,
EMR = set of reference records,

SCORE = score between PLC and EMR record comparison,
C = class that defines the level of agreement (e.g., I or II),
C^{end} = last class C
$\square_i^C$ = error term representing the tolerance of variable *i* of class C,
β_i = score increment representing the agreement weight of variable *i*, and
α_i^C = compensation on the weight regarding the class C.

Notes: For the probabilistic variables it is possible to restrict values from emergency data set to be higher than the police data set, lower than the police data set or with no restriction, respectively option 1, 2 and 3. In addition, in the case of accidents occurred in Maia and Gondomar cities, the victim's transportation will be longer than the transportation of the victims of accidents occurred in Porto city, because both hospitals that cover these cities are located in Porto city. Thus, class II having a penalty for Maia and Gondomar cases.

Table 5 summarizes the model parameters used in the model (see top of the table), and the possible scores to be reached for the two scenarios: Porto city, and Maia and Gondomar cities.

Because Hospital VNG presents some major problem related to the issues previously described, it was decided work separately in the linkage process. Nevertheless, it was used a linkage method between police and emergency records quite similar to the one used for the others hospitals. The only difference was that Hospital VNG has a field that indicates the type of victim within a road accident: driver, passenger and pedestrian. Because police records also have this information, we could set a new variable TYPE that can assume 4 values; 1 if driver, 2 if passenger, 3 if unknown and 5 if pedestrian. This coding allows that with the use of a tolerance of 2, an unknown field can match either pedestrian or driver/passenger value. Three scenarios are possible: 1 records from police and hospital have a exact match of variable TYPE – score +1; one of the records (either form police or hospital) has variable TYPE as unknown or one is passenger and other is driver, match with maximum tolerance of 2 – score +0; variable TYPE from police to hospital has a difference higher than 2 – score -1.

Table 5 - Model parameters and possible outcomes

Parameters	β_i	α_i^{I}	α_i^{II}	□$_{indate}^{I}$	□$_{intime}^{II}$	□$_{age}^{I}$
Porto	*k* (constant)*	1	0.5	8 (115.2 minutes)	1 (14.4 minutes)	0
Gondomar	*k* (constant)*	1	-0.5	10 (144.0 minutes)	>2 (28.8 minutes)	0
Maia	*k* (constant)*	1	-0.5	10 (144.0 minutes)	>2 (28.8 minutes)	0

*the simplest case considers βi constant thus any value will lead to the same results

Scores	**Scenario Porto**
4 *k*	All agree
3.5 *k*	All agree within tolerance class I or II
3 *k*	INTIME don’t match within tolerance class I and II
2 *k*	AGE or GENDER don't agree with tolerance class I
1.5 *k*	AGE or GENDER and INTIME don't agree within tolerance class I
1 *k*	AGE or GENDER and INTIME don't agree within tolerance class I and II
0 *k*	Nothing agrees
Scores	**Scenario Maia and Gondomar**
4 *k*	All agree with tolerance class I
3 *k*	AGE and GENDER agree but INTIME don’t agree (arrive at hospital in more than 144 minutes)
2.5 *k*	All agree within tolerance class I or II
2 *k*	AGE or GENDER don't agree with tolerance class I
1 *k*	AGE or GENDER and INTIME don't agree within tolerance class I and II
0.5 *k*	AGE or GENDER and INTIME don't agree within tolerance class I
0 *k*	Nothing agrees

3.4. Model Results

The model was run using the software *LinkageWiz*, with the due adaptations and using a 1:1 matching relation, meaning that to one record source corresponds to a single reference record. This software uses probabilistic data matching algorithms and allows users to define their own variables and weights, being a good tool for implementing the presented model after some adaptations. In this study, the software was mainly used to create an automatic procedure for comparing and rating pairs of records. The results were then ranked and analysed.

The linkage process was applied separately by city, because of the differences that may occur on the time between the accident and the arrival to the hospital. The matching process starts with matching both hospitals located in Porto city, Hospital SJ and Hospital SA, with accidents occurred in Porto city. Then, the remnants of each hospital are matched with the two other cities – Hospital SJ remnants are linked to Maia city and Hospital SA to Gondomar city.

The linkage process between the police and hospital emergency data sets resulted in a total of 4,210 linked records (see Figure 1, Diagram 1). This result leads to a matching success of 34% (45% success rate for Porto, 26% for Maia and 17% for Gondomar). For the Hospital SJ, the linkage between the emergency and inpatient entries resulted in 1,103 matches out of 9,370 emergency records from which 1,182 were clearly stated as requiring inpatient. In addition, the losses resulted from the linkage process between the police and the hospital brought down the total number of requiring inpatient to 185. From this number, 27 were linked to emergency victims without information mentioning the need of hospital admission (false positives), and from the ones that explicitly required hospital admission, 20 were not connect (true negatives). Thus, for the case of Hospital SJ, we have a success of emergency to admission linkage rate of 89% within the records linked to the police.

In sum, the three cities' data sets, together, reach a success linkage rate of 34%. The ambulance service data set of 2010 was used for validation purposes (only for Porto city). After the data treatment, it was possible to verify 98% of matched records (see Figure 1, Diagram 3), denoting that the linkage process has a potential of 98% of true matches.

The linkage methodology used in Porto's hospitals (Hospital SA and Hospital SJ) was successfully applied to the Hospital VNG case. In fact, adding the variable TYPE (only possible to be used in this case), the linkage success rate increased up to 56%. While this variable is not relevant for the matching success it brings more certainty to the linked record results.

Despite this increase in the linkage success rate, Hospital VNG separated the records of the emergency units and of the admission units. When analyzing both data sets, we concluded that emergency records reported around 987 victims as requiring hospital admission, while the admission unit records only have around 590 inpatients as a result of road accident. Therefore, almost half of the traffic injuries requiring hospital admission were lost. The explanation provided by the hospital was that some entrances in the inpatient registries were not reported as victims of accidents (by using the ICD code of external cause) and the data sent to us was filtered by the ICD code describing the cause as a traffic accident.

For that reason, the data set of the Hospital VNG will be used separately in future studies. The data will not be discarded as it can present a very good resource to validate conclusions and result models that can arise from the deeper study of Porto, Gondomar and Maia victims.

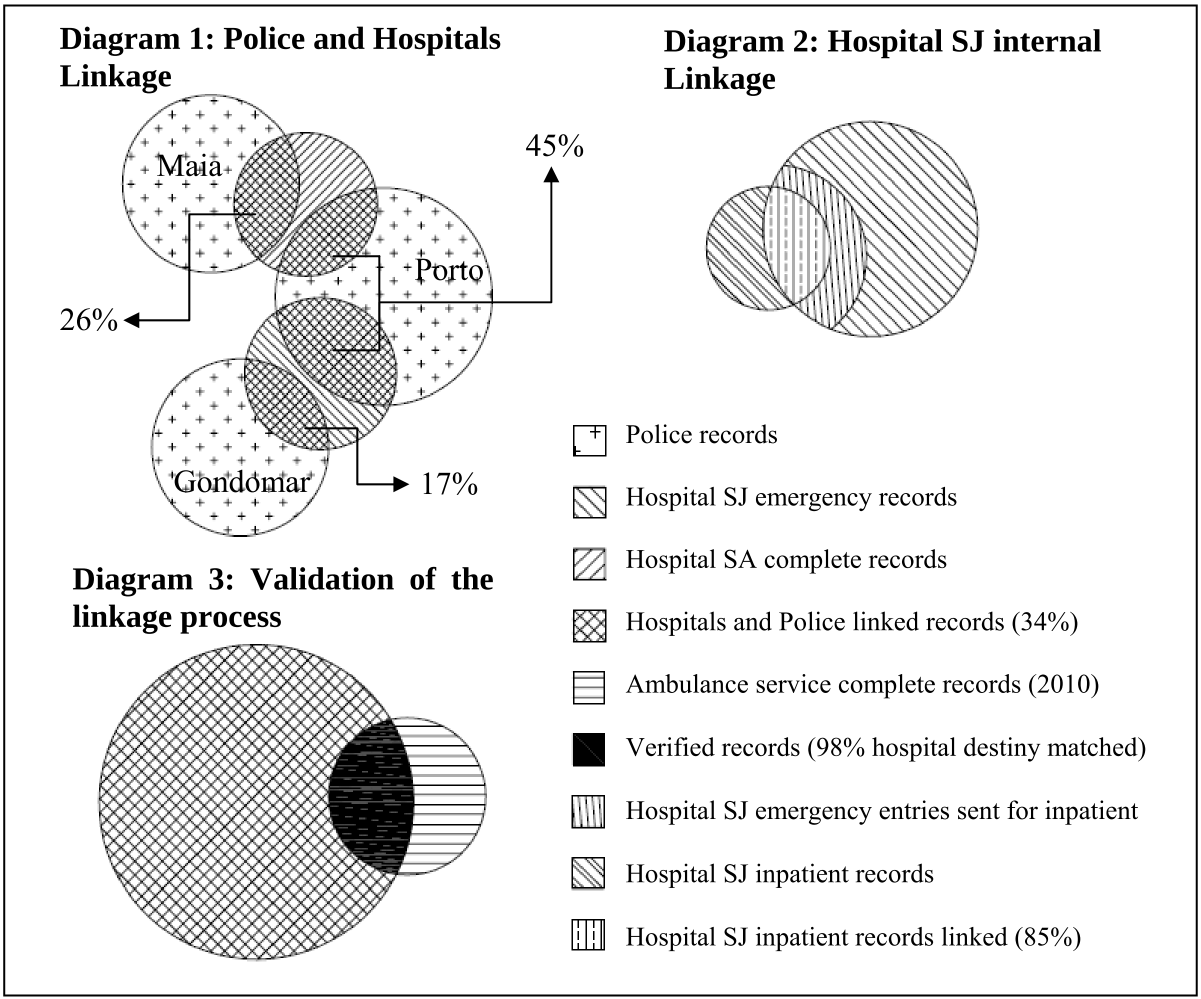


Diagram 1: Police and Hospitals Linkage
Maia
Porto
Gondomar
45%
26%
17%
Diagram 2: Hospital SJ internal Linkage
Diagram 3: Validation of the linkage process
Police records
Hospital SJ emergency records
Hospital SA complete records
Hospitals and Police linked records (34%)
Ambulance service complete records (2010)
Verified records (98% hospital destiny matched)
Hospital SJ emergency entries sent for inpatient
Hospital SJ inpatient records
Hospital SJ inpatient records linked (85%)

Figure 2 - Linkage Venn-diagrams results

3.5. Investigation of Bias

A loss of cases resulting by non-matching records is inherent to this type of linkage process. Thus, this section presents a first approach to investigate a possible bias resulting from the linkage process based on the statistical analysis of the variables age, gender, injury severity, and diagnosis-related group code (DRG) by comparing them before and after the linkage. Note that each individual data set is likely to be inaccurate and biased in reporting as a consequence of the various problems described in the section 3.2 and widely reported in the literature (*9, 21-24*).

The age, gender, and injury severity are analyzed using the police data set as a source reference, because the hospital data sets may include accidents occurred outside the studied area. The DRG is analyzed using the hospital data set as a reference, because this information is only recorded at the hospitals. The DRG is based on a codification process to assess the inpatient cost. Table 6 presents the average and standard deviation of these variables before (full data) and after (linked data) the linkage process.

Table 6 - Average (standard deviation) of the age, gender, severity and DRG

Source Reference	Variable	Full Data	Linked Data
Police	**Age**	37.5 (20.2)	40.4 (19.7)
	Gender	1.90 (1.00)	1.98 (1.00)
	Severity	0.03 (0.17)	0.07 (0.34)
Hospital Inpatient Time	**DRG**	438.5 (268.1)	396.5 (257.3)

The statistical values shown in Table 6 are very similar between the full and linked data. Although these results should be interpreted with caution, it is possible to conclude from this preliminary analysis that no significant bias is presented by the sample resulting from the linkage process. Nevertheless, future analysis should be undertaken using, for example, the Pearson´s χ^2 test (*8*).

The main differences between the full data and the linked data are in the severity variable. This may be because some SLI did not attend at the hospital and consequently do not appear on the linked data.

3.6. Statistical analysis of the linked data: a first overview

Considering that there was not a significant bias between the linked records and all records (before linkage process), in this section a previous statistical analysis of the linked records is presented (including the Hospital VNG linked records). Note that victims who died at the scene or were admitted/transferred to other hospitals are not included in this analysis.

3.6.1. *Victim's age and gender*

Table 7 shows the distribution of the linked records by gender and age. Although with a slight difference, accidents affected more males than females as victims (52% and 48%, respectively) probably because men are more involved in road accidents since more men own, drive and use motorized transport. These statistical values are in line with overall results presented elsewhere (*25*). The age group with more number of victims is 20-24. The victims between the age of 20 and 34 represent

34% of the total number of victims, and again with more victims that are men than women (54% and 46%, respectively). In sum, the highest percentage of the victims is young men. Note that these results are in line with the previous analysis provided by the police data (see Table 3).

Table 7 – Gender and age of the victims

Number of victims (linked records)		
Age	**Male**	**Female**
0-4	46	53
5-9	73	69
10-14	103	97
15-19	280	220
20-24	501	391
25-29	428	385
30-34	451	385
35-39	366	355
40-44	310	269
45-49	280	279
50-54	243	255
55-59	193	184
60-64	171	148
65-69	142	107
70-74	123	123
75-79	90	130
80-84	57	63
85+	25	28
Total	3882	3541

3.6.2. *Type of road users' victims*

As shown in Table 8, drivers are the highest share of victims of road accidents (57%) in which males represent the highest percentage (65%). In contrast, females are the highest share of passenger victims (65%), specially seating on the front seat of the vehicle. Pedestrian victims correspond to 14% of the total of the victims in which women are the highest percentage (61%). Again, this analysis is in line with the previous one presented in Table 3.

Table 8 – Victim's road user type and seat position in the vehicle of the victims

Road user type	Male	Female	Total
Driver seat	2708	1488	4196
Front seat passenger	299	679	978
Back seat passenger	292	449	741
Passenger (seat position unknown)	67	118	185
Pedestrian	516	807	1323
Total	3882	3541	7423

3.6.3. *Driver victim's alcohol*

Table 9 shows that 4% of the driver victims presented values of blood-alcohol test higher than 0.50 (50 mg/100 ml), which is maximum Portuguese legal value. Males represent the highest percentage of drivers with blood-alcohol test above the legal value (92% of those tested positive were men). Finally, it is impressing to note that, from those drivers under illegal condition, 50% of them were under influence of alcohol with values higher than 1.50.

Table 9 - Blood-alcohol test of driver's victims

Blood-alcohol Test (mg/ml)	Male	Female	Total
0.00	2509	1474	3983
]0.00 – 0.20[	5	0	5
[0.20 – 0.50[	34	1	35
[0.50 – 1.50[	78	8	86
≥ 1.50	82	5	87
Total	2708	1488	4196

3.6.4. *Length of stay in hospital and injury severity*

An analysis of the length of stay (in days) in hospital allows a first glance in terms of victims' severity or at least in terms of hospital costs. As reported by IRTAD (*1*), a number of agencies have used length of stay in hospital as a proxy for severity in their national non-fatal indicators of injury incidence. However, studies have showed that this indicator is not a stable proxy measure for severity (*26*). Clark and Rosen (*26*) state that "In many health systems, there are managerial and financial incentives to reduce length of stay". Therefore, we intend to analyse, in the next LIVE project task, the suitable of this indicator by comparing with other injury scales, specifically the MAIS.

Table 10 shows that a higher number of victims did not need to be admitted as inpatient (92%). Besides those victim records, a high percentage of the victims stayed in hospital more than seven days (43%), and 16% stayed one day or less in the hospital.

Table 10 – Length of stay in hospital

Length of stay (in days)	Number of linked records
≤ 1	83
2	55
3	47
4	48
5	36
6	31
7	8
>7	228
Not need hospital admission	6867
Total	7403

Table 11 compares the length of stay with the severity recorded by the police according the definition previously referred (see chapter 2.3). 94% of the victims reported by the police as SLI were reported only at emergency unit. From those SLI police records that needed to be admitted as inpatient, 36% stayed in the hospital more than 7 days. In addition, 36% of the victims reported as SEI were not admitted in the hospital. These percentages may be interpreted as incorrect severity record by the police. Note that in this comparison the fatalities were excluded because there are even more errors associated to these values since the victims that died in the accident scene may be not reported by the hospital and also, the Hospital VNG did not report the fatalities. Severity indicators will be one of the focuses of the next LIVE project task.

Table 11 – Comparison between length of stay in hospital and the police severity information

Length of stay (in days)	SLI	SEI
≤ 1	70	12
2	45	10
3	40	6
4	46	2
5	33	3
6	25	6
7	22	6
>7	161	66
Without hospital admission	6703	63
Total	7145	174

From this brief statistical analysis, we may conclude for instance that high differences in terms of victims' demographics exist. In fact, these differences may have influence in the linkage process based on probabilistic methods when personal information such as name and personal ID are impossible to get due to confidential information protection. A numerical methodology for an algorithm based on the data set's demographics, mainly age and gender is being undertaken. By analyzing the various demographic fields it is possible for the algorithm to calculate individual weights that depend on the occurrence of each fields' values among a desired data set for a specific period of time. The method resembles a human linkage process thus, has a higher potential to avoid false matches and false non-matches therefore clarifying and improving the clerical review process.

4. TOWARDS AN INJURY CLASSIFICATION BASED ON MEDICAL SCALE

One of the main objectives of the LIVE project is to study injury scales to describe traffic victims' severity. Until date, in Portugal, traffic victims are classified as SLI, SEI and F by the police that collected the accident data. Only since 2010 the hospitals have to inform the police and the ANSR when an accident victim dies within the 30 days following the accident.

As it can be concluded by the previous analysis, in particular by Table 11, differences can be seen between the police classification of the victims and the actual observations of the length of stay in hospital of the victims. This analysis suggests the existence of misreporting in terms of the classification of the casualties. To overcome the misreporting, an injury classification alternative is required.

The importance of counting and classifying road traffic injuries is self-evident to anyone who wishes to understand the nature and scale of such trauma, and to develop appropriate policies to diminish road traffic deaths and morbidity (*27*). Recognizing that severity classification is critical for surveillance, epidemiological investigations and evaluations of programs and policies aimed at mitigating the impact of injury, , the number and variety of injury severity scales have increased over the past two decades.

Several studies were undertaken to analyze the best measure injury severity based on information typically available in administrative hospital data sets. According to the PENDANT (Pan-European Co-ordinated Accident and Injury Database) European project, only two scales allow the degree of required information to be handled in an appropriate and simplistic manner, namely the AIS and the ICD scale. These two scales are a consensual choice that has been used wherever is possible in the road accident injury studies.

The ICD is used for coding all inpatient hospital data and therefore not exclusively used for coding injury/trauma. The ICD has been used at an international level since 1893 and is now under the responsibility of the World Health Organization. Further revisions were undertaken and the most recent revision is the ICD10 (the 10th revision). According to IRTAD (*1*), in most countries, hospitals do not define levels of injury as such, but use the ICD (ICD9 or ICD10), which is derived from the medical diagnosis to describe the injuries. This classification of diagnosis for all health conditions includes diagnostic codes for both, nature of injury and external causes of injury. Note that ICD code does not incorporate an explicit severity dimension. The ICD is widely used to classify health conditions in the clinical, administrative, public health promotion and research settings.

The AIS is a specialized trauma classification of injuries based mainly on anatomical descriptors of the tissue damage caused by the injury and was introduced in 1971 by the Association for the Advancement of Automotive Medicine (the last update was in 2008). The AIS is an anatomical-based coding system created to classify and describe the severity of specific individual injuries. The severity of an injury based on "threat to life" on a 6-point scale range from minor to untreatable injuries. The severity score ranges from 1 (relatively minor) to 6 (currently untreatable), and is assigned to each injury descriptor. Injuries should be coded to the AIS by trained staff of trauma services or by specialists in injury data collection. Assigning AIS scores based on medical records is time consuming; it is not routinely done outside of trauma centres.

The AIS forms the basis of some severity scores such as the MAIS and Injury Severity Score (ISS). The MAIS is the maximum AIS severity score of a casualty with several injuries. MAIS is an internationally accepted summary measure of injury severity, and was recently assumed as the common international definition in EU.

An approach for using the ICD for severity assessment had been developed by using an algorithm that translates ICD into AIS codes. Resulting severity scores referred to as ICD/AIS scores are considered to be conservative measures of injury severity. Despite there are some tools such as ICDMAP developed to provide this conversion, it is not straightforward to obtain because usually these tools are proprietary as is the AIS. Furthermore, the ICD and the AIS are frequently updated to new versions, which imply that a stable system is needed (human, equipment and financial resources) in order to ensure the feasibility of the results.

In Portugal, the ICD codes are applied, however the AIS scale is not commonly known by the medical community, which leads to a lack in the AIS practice.

Consequently, at the moment, there is no tool (medical expertise or software) in Portugal that may support the common definition of serious injury based on the MAIS, already established by the EU. In this context, in order to be possible to apply the AIS scale, international contacts and knowledge are needed to overcome this gap. The APOLLO European project has been our source, hoping to provide compatible information to the Portuguese context, not only to the purpose of the LIVE project but most of all to guarantee the introduction and future systematic practice in Portugal of an injury scale based on medical diagnosis such as the MAIS.

5. CONCLUSIONS AND RECOMMENDATIONS

Hospital records have the potential to improve the injury accident data when combined with the police data. More detailed information on injury types and severity is reported by the hospitals, improving the data from the accident scene and providing a more comprehensive view of road accidents.

Furthermore, casualty figures allow international comparisons, highlighting road safety as a priority for action and promoting robust arguments for intervention.

In Portugal, until date, there is not a database connecting the accident police data set and the injuries medical hospital information that may support several studies of the relationship between the accident sites and the injury severities.

Under the project LIVE, a first approach was developed to link police and hospitals data sets covering the Porto region.

From the work developed under this project, several issues related to the data sets were found that have impact on the results of the record linkage process, which in time reduces the possibility of having an actual knowledge of the number of victims and their severity. Nevertheless, the validation process and the bias investigation after the data linkage may ensure the quality of the results in terms of sample representativeness of the data sets.

From this work, the following conclusions and recommendations are presented:

- The combination of the police and hospital data is not straightforward. Differences in definitions, changes in data collection and recording practices are issues that may affect the comparisons between data sets, and also the analysis of trends over time. In this point, national recommendations in terms of best practices to gather the data in order to a standardized process, particularly in the hospitals, should be established and spread.
- In order to increase the number of records matched between emergency and admission units, we suggest to work with all inpatient records which have ICD codes that may be related to a road accident (for example, using the Barell Matrix or another dictionary of conversion), since, in some cases, the identification of the external cause of the inpatient, using the respective ICD code to identify traffic accident, is not reported.
- Each studied hospital has a distinct data system, each one with specific issues. To overcome those several and specific issues, we suggest the use of the personal ID number of the national health system. In Portugal, each person has a personal ID number as a national health user, which is collected by the healthcare system. This number allows linking the information of a victim between any hospital unit and between hospitals (when a transfer occurs). With this number, it is possible to follow the progress of the health condition of the victim until the hospital discharge, in any healthcare institution as well as in the mortality register. However, at the present, the CNPD considered that data protection rules are applied to this personal data. Thus, legal conditions should be provided in the future in order to support the use of the ID number.
- The data fields available to the linkage process determine the methodology to be used. If a personal ID number is used, a simple deterministic methodology is possible to be applied.

- Specific data field information might improve the linkage results. For instance, one possible improvement allowing the increase in the rate of True Matches and decrease in the rate of False non-matches would be to use the birthday instead of the age. This would increase the probability of uniqueness in 365 times. Therefore it would be a good practice for each entity to include the birthday in its records.
- In this study, the mixed deterministic and weight-based probabilistic method to link the police and hospital records proves to be efficient in gathering the various field characteristics. Fields such as age, gender and date are used as deterministic variables and thus, linkage may only exist if they all match. When those fields match, time can be used as a probabilistic variable to access the afterwards likelihood of two records being a true match. In this field, future developments are planned in order to improve the record linkage results, using an algorithm to calculate individual weights based on the data set's demographics, mainly age and gender.
- The commonly available records used in this type of data process contain many errors or omissions. In this study, missing values were detected using separate analyses of the data sets. This data issue may be overcome by applying techniques to complete the linkage process, such as multiple imputation and capture/recapture method. However, in general, to be possible to apply a mathematical technique such those, several conditions need to be meet, which are in many cases not possible (*11*).
- The linkage rate found in this study lies within the range of rates found in other studies, which suggest that the linkage process is feasible to apply in the Portugal data system.
- The ambulance service data is an alternative to personal identifiers to validate matched records (98% of the matched records were validated by these data). One step forward would be to get information on the time of arrival of the ambulance to the hospital, providing a better validation and calibration. In this sense, recommendation in terms of data collection and storage should be established and spread by the INEM, which is a national institution with a unique data system, in order to have access to this important information.

REFERENCES

1. IRTAD. 2012. Reporting on Serious Road Traffic Casualties. Combining and using different data sources to improve understanding of non-fatal road traffic crashes. OECD/ITF.
2. Roos, L. L. and A. Wajda. Record linkage strategies. Part I: Estimating information and evaluating approaches. *Methods Inf Med,* Vol. 30**,** 1991, pp. 117-23.
3. Van, H. T., P. Singhasivanon, J. Kaewkungwal, P. Suriyawongpaisal and L. H. Khai. Estimation of non-fatal road traffic injuries in Thai Nguyen, Vietnam using capture-recapture method. *Southeast Asian J Trop Med Public Health,* Vol. 37**,** 2006, pp. 405-11.

4. Benavente, M., M. Knodler and H. Rothenberg. Case Study Assessment of Crash Data Challenges. *Transportation Research Record: Journal of the Transportation Research Board,* Vol. 1953**,** 2006, pp. 180-186.

5. Cirera, E. V. A., A. PlasÉNcia, J. Ferrando and P. Arribas. Probabilistic Linkage of Police and Emergency Department Sources of Information on Motor-Vehicle Injury Cases: a Proposal for Improvement. *Journal of Crash Prevention and Injury Control,* Vol. 2**,** 2001, pp. 229-237.

6. Clark, D. E. Practical introduction to record linkage for injury research. *Injury Prevention,* Vol. 10**,** 2004, pp. 186-191.

7. Cook, L. J., L. M. Olson and J. M. Dean. Probabilistic Record Linkage: Relationships between File Sizes, Identifiers, and Match Weights. *Methods of Information in Medicine,* Vol. 40**,** 2001, pp. 196-203.

8. Cryer, P. C., S. Westrup, A. C. Cook, V. Ashwell, P. Bridger and C. Clarke. Investigation of bias after data linkage of hospital admissions data to police road traffic crash reports. *Injury Prevention,* Vol. 7**,** 2001, pp. 234-241.

9. Dandona, R., G. A. Kumar, M. A. Ameer, G. B. Reddy and L. Dandona. Under-reporting of road traffic injuries to the police: results from two data sources in urban India. *Injury Prevention,* Vol. 14**,** 2008, pp. 360-365.

10. Jones, A. P. and G. Bentham. Emergency medical service accessibility and outcomefrom road traffic accidents. *Public Health,* Vol. 109**,** 1995, pp. 169-177.

11. Kudryavtsev, A. V., N. Kleshchinov, M. Ermolina, J. Lund, A. M. Grjibovski, O. Nilssen and B. Ytterstad. Road traffic fatalities in Arkhangelsk, Russia in 2005–2010: Reliability of police and healthcare data. *Accident Analysis & Prevention,* Vol. 53**,** 2013, pp. 46-54.

12. Gennarelli, T. A. and E. Wodzin. AIS 2005: A contemporary injury scale. *Injury,* Vol. 37**,** 2006, pp. 1083-1091.

13. HB, N. Handbook of record linkage: methods for health an statistical studies. *Oxford University Press,* Vol. Administration and business**,** 1988, pp.

14. Lujic, S., C. Finch, S. Boufous, A. Hayen and W. Dunsmuir. How comparable are road traffic crash cases in hospital admissions data and police records? An examination of data linkage rates. *Australian and New Zealand Journal of Public Health,* Vol. 32**,** 2008, pp. 28-33.

15. Petridou, E., G. Yannis, A. Terzidis, N. Dessypris, E. Germeni, P. Evgenikos, N. Tselenti, A. Chaziris and I. Skalkidis. Linking Emergency Medical Department and Road Traffic Police Casualty Data: A Tool in Assessing the Burden of Injuries in Less Resourced Countries. *Traffic Injury Prevention,* Vol. 10**,** 2009, pp. 37-43.

16. Soufiane Boufous, Caroline Finch, Andrew Hayen and A. Williamson. 2008. Data Linkage of Hospital and Police Crash Datasets in NSW. Sydney: NSW Injury Risk Management Research Centre University of New South Wales.

17. ANSR. 2012. Estratégia Nacional de Segurança Rodoviária ENSR - Documento de Apoio à Revisão Intercalar 2012-2015. Lisboa, Portugal: Autoridade Nacional de Segurança Rodoviária.

18. McGlincy, M. H. A Bayesian record linkage methodology for multiple imputation of missing links. In *ASA Proceedings of the Joint Statistical Meetings*. 2004, 4001-4008.

19. Winkler, W. E. 2010. The State of Record Linkage and Current Research Problems. U. S. Bureau of the Census, Washington DC.

20. Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association,* Vol. 84**,** 1989, pp. 414-420.

21. Alsop, J. and J. Langley. Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis & Prevention,* Vol. 33**,** 2001, pp. 353-359.

22. Amoros, E., J.-L. Martin and B. Laumon. Under-reporting of road crash casualties in France. *Accident Analysis & Prevention,* Vol. 38**,** 2006, pp. 627-635.

23. Jaro, M. A. UNIMATCH: a computer system for generalized record linkage under conditions of uncertainty. In *Proceedings of the May 16-18, 1972, spring joint computer conference*, No. ACM, Atlantic City, New Jersey, 1972, pp. 523-530.

24. Yamamoto, T., J. Hashiji and V. N. Shankar. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accident Analysis and Prevention,* Vol. 40**,** 2008, pp. 1320-1329.

25. Lopez, D. G., D. L. Rosman, G. A. Jelinek, G. J. Wilkes and P. C. Sprivulis. Complementing police road-crash records with trauma registry data — an initial evaluation. *Accident Analysis & Prevention,* Vol. 32**,** 2000, pp. 771-777.

26. Clarke, A. and R. Rosen. Length of stay: How short should hospital care be? *The European Journal of Public Health,* Vol. 11**,** 2001, pp. 166-170.

27. Morris, A., M. Mackay, E. Wodzin and J. Barnes. Some Injury Scaling Issues in UK Crash Research. In *2003 International IRCOBI Conference on the Biomechanics of Impact*, No. IRCOBI, Lisbon, Portugal, 2003, pp. 283-291.