Manuscript Number: JSR-D-15-00478R2

Title: Record Linkage for Road Traffic Injuries in Ireland using Police
Hospital and Injury Claims Data

Article Type: Full Length Feature Article

Keywords: Road safety
Accident records
Data linkage

Corresponding Author: Prof. Brian Caulfield, Ph.D

Corresponding Author's Institution: Trinity College Dublin

First Author: Jack Short

Order of Authors: Jack Short; Brian Caulfield, Ph.D

Abstract: The study of non-fatal road traffic injuries is growing in importance. Since there are rarely comprehensive injury datasets, it is necessary to combine different sources to obtain better estimates on the extent and nature of the problem. Record linkage is one such technique. In this study, anonymised datasets from three separate sources of injury data in Ireland -hospitals, police and injury claims are linked using probabilistic and deterministic linkage techniques. A method is proposed that creates a 'best' set of linked records for analysis, useful when clerical review of undecided cases is not feasible. The linkage of police and hospital datasets shows results that are not dissimilar to those found in other countries, with significant police understatement especially of cyclist and motorcyclist injuries. The addition of the third dataset identifies a large number of additional injuries and demonstrates the error of using only the two main sources for injury data. The study also underlines the risk in relying on the Lincoln-Petersen capture-recapture estimator to provide an estimate of the total population concerned. The data show that road traffic injuries are significantly more numerous than either police or hospital sources indicate. It is also argued that no single measure can fully capture the range of impacts that a serious injury entails.

Record Linkage for Road Traffic Injuries in Ireland using Police Hospital and Injury Claims Data

Jack Short and Brian Caulfield¹

Department of Civil, Structural and Environmental Engineering, Trinity College Dublin,
Dublin 2, Ireland

¹ Email: brian.caulfield@tcd.ie; Tel: +353 1 896 2534

Research Highlights:

- This paper uses data matching techniques to match three independent road safety databases.
- The results show an under reporting of several accident types as used by the official road safety statistics

Record Linkage for Road Traffic Injuries in Ireland using Police Hospital and Injury Claims Data

ABSTRACT

The study of non-fatal road traffic injuries is growing in importance. Since there are rarely comprehensive injury datasets, it is necessary to combine different sources to obtain better estimates on the extent and nature of the problem. Record linkage is one such technique. In this study, anonymised datasets from three separate sources of injury data in Ireland -hospitals, police and injury claims are linked using probabilistic and deterministic linkage techniques. A method is proposed that creates a 'best' set of linked records for analysis, useful when clerical review of undecided cases is not feasible. The linkage of police and hospital datasets shows results that are not dissimilar to those found in other countries, with significant police understatement especially of cyclist and motorcyclist injuries. The addition of the third dataset identifies a large number of additional injuries and demonstrates the error of using only the two main sources for injury data. The study also underlines the risk in relying on the Lincoln-Petersen capture-recapture estimator to provide an estimate of the total population concerned. The data show that road traffic injuries are significantly more numerous than either police or hospital sources indicate. It is also argued that no single measure can fully capture the range of impacts that a serious injury entails.

1 Introduction and Background

There are several reasons why traffic injuries are becoming more central to road safety policy. Firstly, significant progress has been made in reducing fatalities with many countries almost halving the number of fatalities in the last decade; however injuries have not declined as rapidly (Table 1). Secondly, the social costs of injuries are very significant and at least as large as the costs of fatalities (Department for Transport UK, 2012a; Ministry of Transport New Zealand 2012; SWOV, 2014). Thirdly, data on injuries are less reliable than on fatalities and are not comparable internationally; for example the number of injuries reported per fatality varies from 20 to 150 in OECD Countries (see Table 1 and Figure 1). Fourthly, the larger number of injury crashes can provide statistically significant results for policy analysis of crash factors. Finally, the policy focus on fatalities may mean that cost effective policies to reduce injuries are not being given adequate attention.

(Table 1 and Figure 1 around here).

Defining and accurately counting road traffic injuries are well known problems (Haagsma et al., 2012; Cryer & Langley, 2006; Fingerhut, 2004; Langley & Brenner, 2004). Difficulties with police injury assessments are also amply documented and the biases and understatement well evidenced (Jeffrey et al., 2009; Department for Transport UK, 2006; Alsop & Langley, 2001; Lopez et al., 2000). A particular issue is the definition of serious injury which varies widely (International Transport Forum, 2012). While clinical assessments, using the Abbreviated Injury Scale (Association for the Advancement of Automotive Medicine, 2008; Gennarelli & Wodzin, 2006) or its derivatives provide more comparable results, these too are not without difficulties. Following research and reports in the European Union (European Commission, 2008) and the International Transport Forum at OECD (International Transport Forum, 2012), a decision to adopt a definition of serious injury as those injuries with a maximum abbreviated injury score (AIS) of 3 or more (MAIS3+) as a European standard has been agreed. This will bring improved comparability to international data on serious injuries though issues of completeness and relevance will remain. Police and hospital data are generally the main sources for traffic injury data but there are also other sources including accident and emergency cases, insurance data and household surveys. Combining different sources of data and assessing the significance of missed data are key challenges. The tools available include record

linkage to combine sources and capture-recapture to make estimates of the unknown missed populations.

This paper explores record linkage techniques and their application using three data sources in Ireland. The data are from police, hospitals and the Injuries Board, an additional source that deals with claims for injury compensation. Section 2 describes the method of record linkage and its application to road crash data. A problem in record linkage is to determine thresholds for deciding on matches and non-matches. A method is suggested which has some advantages over usual methods. Section 3 describes the data available in Ireland and their limitations. In Section 4, the results are presented from the linkages carried out. This section also contains a brief discussion on capture-recapture and in particular the use of the Lincoln-Petersen estimator. Finally, Section 5 discusses the limitations of this work, summarises the main conclusions and examines some policy issues that arise from the analysis.

2 Methods in Record Linkage

2.1. Background and Applications in Road Safety

Record or data linkage involves bringing together corresponding records from two or more files (Winkler, 1999). According to Felligi (1997), it began in the 1960s, with the production of large files about individuals in different domains, as well as the increased role of government in data collection and analysis and the rapid development of computer technology. The use of record linkage in road safety research and practice is relatively recent and has served different objectives. It has been used (usually together with capture-recapture) to make estimates of police underreporting of fatalities and to make estimates of the number of injuries. The International Transport Forum (2012) cites 16 Countries where the technique has been used in road safety. Papers from France (Amoros et al., 2007), and the Netherlands (Reurings & Stipdonk, 2011) are examples. The UK Government uses this method to calculate the social costs of crashes (Department for Transport UK, 2012a) and New Zealand uses it as a benchmark for the police as well as in the calculation of social costs (Ministry of Transport New Zealand, 2012). The method can contribute to a better understanding of the crash problem; specifically the combination of information from different sources can be a valuable research resource on crashes and their consequences.

2.2 Classical and Bayesian Approaches

The original insights of Newcombe et al. (1959) were given a solid mathematical foundation by Felligi and Sunter (1969). The starting point for their paper is the division of the set of pairs, one from each of the two sets to be linked, into sets M and U of matches and non-matches. M and U are not known and the task is to find decision rules to decide if pairs can be deemed to be matches or non-matches. A comparison vector $\mathbf{Y}$ is formed for each pair of records, one from each set. The comparison vector consists of 1s for a link and 0s for a non-link on each of the variables being compared. A linkage rule divides the space of comparison vectors into three categories- A_1 , deemed matches, A_2 , possible matches and A_3 non-matches. Felligi and Sunter (1969) set out the circumstances under which a linkage rule can meet the conditions of satisfying the two predefined levels of statistical error

$$\mu = P(A_1 | U) \text{ and } \lambda = P(A_3 | M).$$

To do this, all possible pairs from the two sets are ranked in descending order of the ratios (called likelihood ratios) of the two conditional probabilities, the probability that a pair linked on a variable is a true match divided by the probability that a pair linked on that variable is not a true match. These conditional probabilities are conventionally known as m and u probabilities (Clark, 2004).

This ranking of likelihood ratios and predetermined statistical error levels lead to two thresholds, a higher and lower, which determine membership of the sets A_1 , A_2 and A_3 . Pairs whose likelihood ratios exceed the higher threshold are deemed matches and those below the lower threshold are considered non-matches. Likelihoods between the thresholds are considered as possible matches and are reviewed clerically.

The assumption that the variables being compared can be treated independently allows the likelihood ratios to be calculated for each variable separately and multiplied. Using logarithms to base 2, Newcombe et al. (1959) calculated weights for each pair, as the sum of the logs of the likelihood ratios. This practice has been retained, for example in the computer programme used here (Linkage-Wiz, 2013) even though it is not mathematically necessary.

The decision thresholds emerge from the predefined acceptable levels of statistical error. However, once these are defined, the size of the set of possible matches, the indeterminate set A_2 , can be of impractical size. Then, judgements about the data as well as the possibility to verify the true status of links become important. In the linkages undertaken in the present work, and in much research work, there is no information additional to that in the records. Clerical review of uncertain cases cannot provide further indications on whether a true match is more or less likely than indicated by the probabilistic calculations. For practical purposes then, the ineffectiveness of clerical review often requires a single threshold for scores, above which a pair is deemed a match and below which it is deemed a non-match.

2.3 Decision Rules for Matches and Non-matches

In the road accident literature, a way to deal with the problem of a decision rule has been to use “matching standards” Examples include Razzak & Luby (1998), Morrison & Stone (2000) and Department for Transport, UK (2012). This involves using a range of possibilities for the number of true matches on the basis of the number of variables that are linked. So different matching standards can be applied and tested, for example, for no variable unlinked, one variable unlinked, two variables unlinked and so on. The advantage of this method is that it can help provide orders of magnitude for matches and can show approximatively how the matching results would vary depending on the different standards applied. However, there are disadvantages too. The main one is that a non-link on one particular variable is not the same as a non-link on another. Some of the variables are strong discriminators and a non-link on such variables is different in probabilistic and practical terms. Treating them the same inevitably loses information. However, the method can be completely consistent with theory if the matching standard variables are selected in order of declining likelihood.

Here a further method is proposed. It is based on a Bayesian approach, not explicitly mentioned by Newcombe et al. (1959) or Felligi and Sunter (1969). It is similar to the classical method but has the advantage that, on using a prior probability of a match, it provides a simple method to estimate the posterior, that is, actual probability of a match.

Using the odds form of Bayes’ theorem and following Clark (2004), the posterior odds of a true match is the prior odds times the product of the likelihood ratios. This allows calculation of a direct relationship between the weights as defined above and the posterior probability (P) as:

$$\text{Posterior odds} = \text{Prior odds} * 2^S, \text{ where } S = \sum_1^k \log_2 \frac{m_i}{u_i} \quad \text{Equation (1)}$$

with the m_i and u_i values being the conditional probabilities of links given a match and non-match respectively over the k linkage variables, and the logs are taken to base 2.

Since the posterior probability $P = \text{Posterior odds} / (1 + \text{posterior odds})$, the scores from the linkage model output can easily be converted into probabilities. Consider now the ranking of pairs by descending values of the posterior probability instead of descending values of the likelihood ratios. This ordering is identical to that of the likelihood ratios (because the ranking of $cL / (1 + cL)$ is the same as that of L where L is the likelihood and c is the prior odds).

In any linkage exercise, the maximum number of true matches is the number in the smaller dataset. Suppose this number is M . Let the linking process be undertaken so that linked pairs are the best possible, in the sense of having the highest posterior probabilities. Once pairs are linked, they are not considered for further links. The process continues until M pairs have been listed. Now consider the posterior probabilities p_i ($i=1\dots M$) ranked from the highest to the lowest over the maximum total number of possible true matches, M , as illustrated in Figure 2 for the police and hospital linkage. The challenge is to decide on how many of these M pairs are likely to be true matches. The answer depends on the probabilities—the higher the probability the more likely that a pair is a match. If the first pair (or the first group of pairs) is a true match with probability p_1 then the expected number of matches is p_1 (or kp_1 if there are k values of p_1). So for example, if there are say 1000 pairs with $p=0.999$ then the expected number of matches is 999. Similarly with p_2 and so on down the list. The total expected number of matches will therefore be $\sum_1^M p_i$, the sum of the posterior probabilities.

This number is clearly less than M , say n to the nearest integer. Now select the first n pairs from the list, ranked in order of descending probability. This is a “best” set of matches in the sense that any other set of the same size would have at least one pair with lower probabilities. Moreover, any smaller or larger set would not have the correct number of expected matches.

This method has the advantage of providing an estimate, n , of the number of matches as well as a dataset for analysis. Such a dataset is a practical solution, which allows analysis of specific overlap groups, improved understanding of the structures of the injured population and has additional information when variables from one set are combined with variables from another for the same individual. If the aim is to avoid false positives, for example to carry out an analysis of perfect matches, then the cut off can be made earlier. If the aim is to avoid false negatives, a cut-off point lower down the list can be taken. This can be equivalently stated in terms of statistical errors. For a lower type 1 error (i.e. fewer false positives) select a higher probability breakpoint; for lower type 2 error (fewer false negatives) select a lower probability.

A closing comment on the prior probability; the prior probability is the probability of a match by chance between the two given datasets. For example, if there is a belief that about 10K records match from datasets of 50K and 40K, then the prior probability is the estimated number of matches divided by the product of the two file sizes, that is $10K / (50K * 40K) = 0.000005$. When the linkage programme is run, the number of matches can be better estimated and the prior probability refined, a valuable feature of the Bayesian approach.

2.4 Statistical Errors

The statistical errors are determined by the method of dividing the pairs into matches and non-matches with no intermediate set of possible matches and by the sizes of the m and u factors, the conditional probabilities of matches or non-matches. Broadly, the greater the m values (the nearer to 1 they are) the more likely it becomes that linked pairs are true matches. So high m values reduce the number of false negatives. Similarly, for the u values, the closer to zero they are, the less likely that matches occur purely by chance. And small u values reduce the probability and the number of

false positives. For example, date of incident and age with u values of 0.001 and 0.01 respectively, have strong discriminatory power in the linkage.

The method above to create a single dataset allows also an intuitive and simple method to calculate statistical errors. The set A_1 , the links defined as matches, is the set of the n largest values of p_i , which are the first n pairs from the monotonically decreasing sequence of probabilities. The set A_3 , the set of non matches, consists of all other pairs, but specifically the pairs in the monotonically decreasing sequence of p_i after the n th pair on to the M^{th} . Using this notation, a credible interval for the number of matched pairs is calculated and then estimates are made of the number of false positives and negatives.

Credible (confidence) interval for the number of linked pairs:

The p_i the posterior probabilities of a match, are independent binomial probabilities and the variance of each p_i is $p_i q_i$. Therefore, the variance of the number of matches, $V(n)$ is given by

$$V(n) = V(\sum_1^M p_i) = \sum_1^M V(p_i) = \sum_1^M p_i q_i = t \quad \text{Equation (2)}$$

and a 95% credible interval is given by: $(n - 1.96\sqrt{t}, n + 1.96\sqrt{t})$.

Expected number of false positives and false negatives:

Let $E(FP)$ be the expected number of false positives and $E(FN)$ be the expected number of false negatives. A false positive occurs when a pair is in A_1 when it is unmatched. The probability for any pair i in A_1 to be unmatched is $1 - p_i$. The expected number therefore is the sum of these probabilities over the set A_1 .

$$\text{So } E(FP) = \sum_1^n (1 - p_i). \quad \text{Equation (3)}$$

A false negative occurs when a pair is in A_3 when it is matched. For each pair i in A_3 , p_i is the probability that it is matched. So the expected number of false negatives,

$$E(FN) = \sum_{n+1}^M p_i = \sum_1^M p_i - \sum_1^n p_i = \sum_1^n (1 - p_i) = E(FP) \text{ as } \sum_1^M p_i = n. \quad \text{Equation (4)}$$

It is seen that in this method the number of false positives and negatives are equal and equal to the sum of the probabilities of not being a match over the set of n pairs in A_1 .

2.5 Record Linkage of Three Datasets

Linking three datasets is less common and different methods to do so are obviously possible. Here, the following procedure was followed. The datasets were first cleaned and then linked in pairs to find and identify matches using the software available (Linkage-Wiz, 2013). Then, for each pair of datasets a “best” set of links was defined in the manner described in Section 2.3 above. This gave three sets of pairs of likely matches. In the second step, these three sets of matched pairs were combined and filtered for triples. Triples occur when, for example, a police case is matched with a hospital case and separately with an Injuries Board case. Putting these triples together gives the cases that are in all three datasets. In consequence, the three sets of pairs are divided into either matched pairs or triples. In the final step, the unmatched records from the three original sources are added to the file of unique doubles and triples, and needed variables are transferred to the retained case record. In the resulting dataset (which will be called the Combined Injuries Dataset) each case refers to a single individual and the data for this individual contains the relevant information from

each of the original data records. Note that the second stage of the matching was carried out using unique record numbers only and did not use the variables on the individual records. It is thus a deterministic method and was carried out using SPSS software.

3. Data Sources and Linkage Variables

This paper uses three data sources—police, hospitals and the Injuries Board. For each source, individual anonymised data were available. Police and hospital data were obtained for the years 2005-2011 and for the Injuries Board data were available only for the years 2010-2011.

The three data sources separately provide valuable comparative information on aspects of the injury problem. The data show different structures and composition by age, gender, county and severity. The large size of the Injuries Board dataset indicates immediately the limitations of relying only on police and hospital data. However, the main aim in this paper is not to analyse these data separately but to analyse them together in a single dataset and this is done for police and hospital data over the period 2005-2011 and for the three sources for the years 2010-2011.

3.1 Police Data

As in many countries, police data in Ireland are the principal and official source of crash and injury data. The difficulties of collecting data at accident scenes are obvious and the risks of inaccuracies are compounded by the fact that the data are later telephoned to a call centre where they are entered into the police database. The process has suffered from delays in transferring the data to the Road Safety Authority who are responsible for publishing the results. Limited editing and consistency checks are undertaken with the result that there are several data fields with incomplete or inconsistent answers. The data cover all police-reported accidents over the period 2005 to 2011. In relation to injuries, it is the police assessment of minor or serious injury on the scene that is used. The definition of serious injury in Ireland is “...an injury for which the person is detained in hospital as an ‘in-patient’, or any of the following injuries whether or not detained in hospital: fractures, concussion, internal injuries, crushings, severe cuts and lacerations, severe general shock requiring medical treatment.” (RSA , 2013). Police data are collision based so that each data line contains all the information on a crash and the individuals involved; for this analysis, the collision-based data were converted to individual data to allow the matching with the other datasets to be carried out. In 2014, improvements to the data collection and transfer process were introduced and it is to be hoped that this will result in improvements in data quality in future.

3.2 Hospital data

Hospital discharge data in Ireland covers around 90% of hospital admissions (Economic and Social Research Institute, 2013). Some private hospitals and for example psychiatric hospitals are not included, though very few road accident cases would go to these hospitals. Accident and Emergency (A & E) cases are not included unless they lead to admission to hospital. The hospital data, known as the Hospital In-Patient Enquiry (HIPE) were managed until 2014 by the Economic and Social Research Institute (ESRI) for the Health Service Executive (HSE).

The data contain no unique personal identifiers, like names, addresses or social security numbers. Hospitals have specialised staff to undertake the clinical and other coding. There is a set of coding manuals, instruction booklets and regularly updated information leaflets on the coding (Healthcare Pricing Office, 2015). The hospitals forwarded the data electronically to the ESRI, where a unit undertook the compilation and verification of the data.

The statistical unit in hospital discharge data is an episode of care. An episode begins on admission and ends at discharge or death. The same person may have more than one episode for the same incident and therefore the number of episodes of care overestimates the number of people involved.

Transport accidents, consisting of emergency cases coded with a “V” code in the World Health Organisations International Classification of Diseases, 10th edition (ICD-10),(WHO, 1990) are selected from the hospital discharge data. The V code is one of the codes used for injuries due to external causes and gives details on the mode of transport involved and on the type of accident. The codes are detailed (for example, codes V00 to V09 are for pedestrians, V10 to V19 for cyclists and so on) and can have up to five characters, giving the mode, the type of crash and with an important categorisation being whether the incident was a traffic or a non-traffic accident. External injury cause appears to be reasonably well coded in Ireland and there are about 5% of hospitalised persons with injury codes that have an external cause that is not stated or is in the residual category “exposure to unspecified factors” (code X59). It is thus believed that the dataset used contains most of those involved in transport accidents. A method to estimate the numbers in the different external cause categories from the injury descriptions has been proposed by Bhalla et al. (2008). But given the relatively small number of such cases this was not deemed necessary in the present study. The hospital data has about 30 variables with administrative, demographic and clinical information for discrete episodes of care. Preparing the hospital dataset for analysis included the following steps; restricting the dataset to injuries by excluding fatalities that occurred up to 30 days after admission; restricting the data to emergency in-patient episodes; limiting it to admissions in the period-not discharges -as is used by hospitals and finally eliminating additional episodes of care for the same injury. These additional episodes were found (after a series of tests on different options) by matching on key variables, age, gender, county of residence and principal diagnosis to one digit (e.g. S3), and a gap of no more than two days between the discharge and hospital re-entry. These procedures led to the reduction in the file from 45,698 cases to 40,822. Of these, almost 1,000 (2.2% of the original dataset) were estimated to be additional episodes of care for the same incident. The additional episodes of care were eliminated and the length of hospital stay accumulated on the retained data record. This elimination of duplicate episodes of care is a separate exercise from the record linkage and was carried out using an SPSS programme.

A further data issue concerns the nature of the transport accident. As mentioned, the ICD-10 V codes make a distinction between a traffic accident and a non-traffic accident. A traffic accident is one that occurs on the public highway and that therefore could appear in the police reported data as it would correspond to international norms for transport statistics (UNECE, 2009). The distinction between traffic and non-traffic accident occurs at the fourth or even fifth character of the ICD-10 V coding. The descriptions are complicated (for example, code V192 is for “ Unspecified pedal cyclist injured in collision with other and unspecified motor vehicles in non-traffic accident”) or the appropriate category is not obvious(for example, code V198 is for “Pedal cyclist [any] injured in other specified transport accident”) and evidence of miscoding was found. As an example, in Ireland almost two thirds of cyclist hospitalisations are categorised as non-traffic accidents. This is far higher than in the United Kingdom (Department for Transport UK, 2012a) and is illustrated by the following: the ICD-10 codes V180 to V182 concern cyclists in non-traffic accidents. V180 is the code for cycle riders in non-traffic accidents; V181 for cycle passengers in non-traffic accidents and V182 is for unspecified cyclists in non-traffic accidents. Code V182 has over 1,500 cases and it seems strange that coders are unsure whether the person was a cyclist or a passenger but were sure that it was not a traffic accident. (Short and Caulfield, 2014). Similar problems with coding traffic and non-traffic accidents were found in Kudryavtsev et al. (2013)

Moreover, in the record linkage results described in Section 4, some of these supposedly non-traffic accidents are matched with police recorded crashes. Since the ICD-10 coding instructions indicate

that the default option is a traffic accident, this issue merits a review at hospital level. The decision here was to retain for initial analysis all transport accidents and make distinctions between traffic and transport accidents where appropriate. This decision would appear to be least justified for some of the V categories like V80-V89 which concern vehicles generally used off-road and for V90-V99 which are residual categories. However, the doubt about coding accuracy persisted with these categories and in particular with the V98 and V99 codes which are often used in the hospital coding. Matches with police cases were found in many of these categories, which is a justification for the decision to retain all V cases for analysis. Clearly, in future, once hospital coding has been reviewed, more specific decisions on exclusions can be taken.

3.3 Injuries Board data

The Injuries Board (IB) was set up in 2004 in response to concern about the high costs of injury claims and insurance in Ireland. While the concern was mainly about road crash claims, the Injuries Board was given a broader mandate covering also work related and public liability injuries. This study focuses on the road crash claims only.

The system is a “fault” one where a claim is made by a “claimant” against a “respondent”, - usually represented by an insurance company. Each claim has to be submitted with a payment (€45 refundable) and a medical report from a doctor confirming the injuries. The assessment by the Injuries Board of a claim leads to a number of possible outcomes, such as the payment of an agreed sum or the rejection of the offer by one of the parties, when litigation can be pursued outside the responsibility of the Injuries Board. A person injured in a road accident who is not at fault and wishes to obtain compensation must apply to the Injuries Board if they are unable to agree settlement with the respondent. The respondent can accept or refuse Injuries Board arbitration. The system is separate from the motor insurance business but many injury insurance claims are settled through the Injuries Board. The number of injury claims that do not go to the Injuries Board is not available as the insurance industry do not publish it.

For this study, the Injuries Board supplied anonymised individual data on motor claims for the years 2010 and 2011. There is no formal definition but according to the Injuries Board motor claims typically involve a road traffic accident involving one or more vehicles (Injuries Board, 2015). The work related and public liability claims were not included in the data file provided by the Injuries Board. This exclusion may mean that a small number of traffic related incidents are not included, for example if a claimant takes a case against a local authority, for example for defective infrastructure, rather than another road user. These are not distinguished separately by the Injuries Board and could therefore not be included. Finally, accidents involving professional drivers or those driving for work purposes would be included as motor cases.

The data fields are; the date of the incident, the date of closure of the file, the size of the award, the date of birth, gender and county or city of residence of the claimant. No information was available on the nature of the incident.

The two more significant problems with the IB data concern the limited amount of information that is collected and the specific data field on the size of the award. In relation to the first issue, to improve the usefulness of the data in future, coding of the incident and the modes involved (for example using the ICD-10 V codes) would provide very useful additional information. At present the dataset contains no clinical or other medical indication and since there is a medical assessment of each case there is the potential to make a direct link to clinical severity measures through coding at least a sample of the injuries to ICD-10. On the second issue, the category “no award” does not necessarily mean that no compensation was paid, only that it was not paid by the Injuries Board.

This means that using the award size as a measure of injury severity is imperfect and may be potentially biased.

The Injury Board dataset brings new insights to the traffic injury problem, and specifically the identification of a set of additional traffic injuries not seen in police or hospital data. The number of injury claims is far larger than the number of injuries in these other sources and therefore gives new indications on the real size of the accident problem. While many claims may be for minor injuries, a significant number are serious and have considerable financial and other costs for the individuals and for society. The compensation awarded by the Injuries Board amounts to around €150 million per annum (Injuries Board Ireland, 2013) with an average payment of over €20,000. However, the numbers are affected by administrative and other factors affecting claims like public knowledge and opinions on the scheme and the influence of legal professionals in encouraging or supporting claims. For example, the number of injury claims has been rising (Injuries Board Ireland, 2013) while both police and hospital numbers have been falling.

3.4 Summary of Data Sources

The datasets from the police, hospitals and Injuries Board are each subsets of the set of all people injured in road accidents. These datasets are obtained in quite different ways. Firstly, police-reported injuries depend on the injuries being known to and compiled by the police. The assessment of injury in police-reported data is subjective, based on observation and experience, statements from those involved and the formal guidance on what constitutes a serious injury. Secondly, hospital admissions are influenced by supply side factors as well as personal characteristics like age and gender, the circumstances of the patient and the nature of the injury. Finally, the Injuries Board data are self-reported injuries and are also influenced by external factors. In addition, Injuries Board claimants require a respondent—someone alleged to be at fault for the injury. This results for example in the exclusion of single vehicle accidents where fault is entirely with the driver, a category that can appear in both police and hospital data. In summary, the definition of an injury in each dataset is different and moreover, it is entirely possible that clinically identical injuries could be included or excluded from each dataset. The conclusion is that the basic statistical unit cannot be rigorously defined for any of the groups. Presence in the dataset is the defining characteristic. This is a feature of injuries data which originate from different sources and is one of the challenges in analysing them.

3.5 Linkage Variables

The record linkage variables are the date (the date of the incident for police and IB data and the date of admission for hospital data), gender, age, county and mode. The linking variables differ slightly between the datasets. Age was available in years only for the police data; year and month of birth was available for the Hospital data and IB had full date of birth. Date of birth for hospital patients is assigned to the 15th day of the month of birth. This is for data protection reasons, where Ireland has strong legislation. This means that age for truly matched pairs can differ by one in some cases. In linkages involving police data a one-year age difference was also accepted. The location variable is county and this was county of residence for hospital and Injuries Board data but county of crash for police data. These are the same in over 85% of cases but this problem reduces the power of this variable to discriminate. It was not possible to link crash locations to hospitals as the hospital data is not available at hospital level and no more refined location variable is available. In the linkages, hospital admission was accepted one or two days after the crash. Table 2 summarises the variables used for the three linkages.

(Table 2 around here)

The large number of records means that a linkage is feasible using Jaro's criteria (Jaro, 1995). Essentially, this requires that the number of different possible combinations over the variables being linked is greater than the total number of records; this is clearly the case as there are about 65 million possible combinations (2 for gender, 100 for age, 26 for county, 2,500 for date and 5 for mode) and around 100,000 records. The linkages followed the procedural recommendations of Hook & Regal (1999). A refinement to the matching programme adapted some u probabilities for specific age groups, lesser used modes and smaller counties with lower probabilities of being linked by chance.

Two sets of linkages were undertaken. Firstly, police and hospital data were linked for the period 2005-2011 and secondly, the three datasets were linked for the years 2010-2011. Police data for this study consists of 58,784, injury cases from 2005 to 2011. Of these 15,131 cases related to 2010 and 2011. Hospital data had 40,822 cases (9,559 for 2010 and 2011) and the Injuries Board consisted of 32,492 cases for 2010 and 2011.

Table 3 sets out the variables used and the m and u values for the police-hospital linkage for the period 2005-2011. The m estimates are derived from the characteristics of the data and from an ex-post examination of the likely matches using the linkage programme. Working estimates for the m values can be obtained, for example by taking the count of the estimated number of matches on each variable at different points on the file of potential matches. This file of potential matches is arranged in descending order of the posterior probabilities as in Figure 2. The number of links for a specific variable can be calculated, for example, at record numbers 5,000, 10,000, 15,000 and 20,000 on the file of potential matches between the police and hospital data. The m values used were those at record 11,509. As an example, the m value for age in the police-hospital linkage is .917 as 91.7% of the records matched on age over the first 11,509 values on the probability list. The matching process used added a refinement to the u values (the probability of a link by chance) for specific categories. For example a match on one of Ireland's smaller counties is less likely by chance than for one on Dublin, the largest and the u probabilities were amended to reflect this. This gave more precise posterior probabilities for specific categories of cases.

(Table 3 and Figure 2 around here)

4 Linkage results

4.1 Police-Hospital Linkage, 2005-2011

Table 4 presents summary information on the results of the record linkage between police and hospital data over the period 2005-2011. The table shows, in the first three data columns, for the different road user categories, the number of cases in the police and hospital datasets as well as the number of matches as calculated by the method set out in Section 2.3 above. Then the aggregate number of distinct cases, which is the sum of the number of police and hospital cases minus the number of matches, is shown. In the final four columns the match rates and ascertainment rates are shown. The match rates give for each dataset the share of the dataset that is matched and the ascertainment rate shows the share of the total number of identified cases accounted for by each of the datasets. It is an indicator of the degree of understatement in each dataset.

The aggregate number of distinct cases is 88,097 made up of 58,784 from the police dataset 40,822 from the hospitals minus the 11,509 matches. The number of matches is equivalent to 19.6% of the police file and 28.2% of the hospital file. Though precise comparisons are not easily made between countries, international experience confirms matching rates of these orders of magnitude

(Department for Transport UK, 2012b; European Commission, 2008; International Transport Forum, 2012).

(Table 4 around here)

For several categories of road user, the differences in the matching rate from the average are statistically significant (with the large sample sizes, differences of over 2 percentage points from the average are statistically significant at the 95% level). Looking first at the police matching rates, (that is the proportion of police-recorded cases that are matched with a hospital case) the highest is for serious injuries where over 50% are linked. Males, cyclists, motor-cyclists and pedestrians, those aged 0-14 or over 65, cases in Cork, Wexford and Donegal all have higher matching rates than average. And those in Dublin, females, other modes and those aged 25-34 have lower matching rates.

For hospitals, the highest matching rates are seen for pedestrians, car occupants, those in traffic (as distinct from non-traffic) accidents, those with clinically serious injuries, people from 35-44 and cases in Cork. Very low rates are seen for cyclists; other modes and those aged from 0-14 also have low matching rates.

Cyclists illustrate the differentials in matching rates most strikingly, with only 8% of the hospitalised cyclists being matched with a police recorded injury. As argued by Short and Caulfield (2014), this is evidence of a weakness in cycling policy because the official data seriously understate the risks involved. Car occupants provide a different example. Hospitalised car occupants are more likely to be matched with police-recorded injuries than other modes except for pedestrians. In contrast, police-recorded car occupants with injuries are less likely to be matched to hospital patients than other categories of road user. This shows that there are many car occupant injuries which do not result in hospitalisation and also that a higher proportion of hospitalised car occupants have been reported to the police. This illustrates how police are more likely to attend, and report on accidents involving cars than cycles or motorcycles.

The different match rates for the traffic injury variable (as distinct from the non-traffic variable as defined by the V codes discussed in Section 2 above) do show that there is a significantly higher match rate for traffic accidents. However, if the coding of traffic and non-traffic injuries were correct there would be no matches, as non-traffic accidents would not appear in the police data. This is therefore a confirmation of the V coding problem identified above.

The final columns in Table 4 show the ascertainment rates, the share of the aggregate of the two sets accounted for by each dataset. This rate pertains only to the two datasets and does not make any estimates for injuries outside the sets. For police, the highest rates are for car occupants and pedestrians and lowest for cyclists, motorcyclists, those under 14 and over 65. For hospitals highest ascertainment rates are for cyclists, motorcyclists, those 0-14 and over 65 while the lowest are for car occupants, those from Dublin, persons from 35-44 and females. The high hospital count for the 0-14 age group is partly explained by the fact that many are admitted to hospital for observation only. In fact, 68% of those aged 0-14 remain in hospital for 1 day or less.

The linkage shows how a significant number of those hospitalised are not recorded by the police. Over the period 2005 to 2011 there were 5,200 people who were classed as seriously injured clinically (MAIS3+) who were not reported to police. This number is roughly the same as the total police-reported number of serious injuries over the period. In addition there were 17,600 hospitalised people (with MAIS<3) and a further 6,800 hospitalised people without a clinical severity rating who were not reported to the police. This latter fact is because the mapping from the ICD

codes to the AIS (European Centre for Injury Research, University of Navarra, 2006) does not assign an AIS value in around 20% of the cases.

Though Irish law (Irish Statute Book, 2015) requires that vehicle accidents involving serious injuries be reported to the police, the evidence here is that a significant number are not reported. The total of almost 30,000 people hospitalised but not recorded by the police over the seven year period is an illustration of the additional cost of road traffic accidents for society beyond those indicated by official data. It also underlines how the extent of the injury problem cannot adequately be measured unless account is taken of different data sources.

In relation to accident severity, police-reported serious injuries declined by 54% over the period. Over the same period serious injuries of those hospitalised declined by 25%. When police-reported severity is compared to clinical severity for those in the overlap set, it is seen that 18% of police-reported minor injuries were classed as MAIS3+ in 2005 and 27% in 2011. If clinical severity coding did not change over the period this therefore is an indication that police reporting of serious injury may have changed, with fewer similar injuries being classed as serious (MAIS3+) in 2011 compared to 2005. This finding echoes Jeffrey et al. (2009) and Gill et al. (2006) for the UK. The conclusion for Ireland is that the police recorded number of serious injuries is neither an accurate measure of the extent of the serious injury problem nor a reliable indicator of the trend.

In summary, record linkage of police and hospital data shows matching rates of less than 20% of police records and less than 30% of hospital records. These rates vary by road user category but exceed 50% only for seriously injured police reported cases. Police reported injuries to persons under 14 or over 65 are more likely to be linked to a hospital patient than those in other age groups. Pedestrians and hospitalised car drivers and passengers are more likely to be linked to a police recorded crash than other road users.

A confidence interval and statistical errors for the police-hospital linkage over the period 2005-2011 can be calculated using the theory described in Section 2.4. These are as follows:

$$E(\text{No. of matches}) = \sum_1^M p_i = 11,509; \sum_1^n p_i = 10,039 \text{ and } \sum_1^M p_i q_i = 2,108.$$

Using the normal approximation, a 95% credible interval for the number of matches is therefore: $(11,509 - 1.96 \sqrt{2108} - 11,509 + 1.96 \sqrt{2108}) = (11,419 - 11,599)$

And the expected number of false positives and false negatives are:

$$E(FP) = E(FN) = 11,509 - 10,038 = 1,471.$$

With the data in this study the statistical errors are large, with the number of false positives and false negatives each equivalent to almost 13% of the matched file. This is a reflection of the earlier discussion on the variables used and their accuracy. On the other hand, the confidence interval may appear to be quite narrow given the relatively large number of false positives and false negatives. It is due to the fact that the variance $\sum_1^m p_i q_i$ is small compared to n because many of the p values are near 1 and $p_i q_i$ is very small (e.g. $0.99 * 0.01 = 0.0099$) with each $p_i q_i$ therefore making a small contribution to the total variance. As pointed out by Brenner (1994), the false positives and false negatives also tend to cancel each other out.

In summary, the data show that a significant number of people (almost 30,000, equivalent to nearly 50% of the police-recorded total) were hospitalised for transport accidents over the seven year period but were not recorded by the police. Of these, over 5,000 were clinically assessed as being seriously injured. This underlines how official data understate the true extent of the injury problem.

The following section develops this idea further by looking at a third source of data on transport injuries.

4.2 Linking Three Datasets

The probabilistic process described at Section 2.3 above to match police and hospital data was applied also to match police and Injuries Board data as well as hospital and Injuries Board data. Then these three sets of pairs can be combined to analyse the matches and overlaps between the three datasets. This allows the construction, from the three basic datasets, of a single dataset where each case is an injured individual who is in either one, two or all three of the original datasets. This dataset, the Combined Injuries Dataset, provides a set of unique individual injury cases from the original three datasets.

The process to generate this dataset is as follows. Starting with the three datasets for the years 2010-2011 from the police, hospitals and Injuries Board the Linkage-Wiz matching programme and the probabilistic method described at Section 2.3 is applied to these datasets in pairs. This gives three sets of pairs of “best” matches. Table 5 shows the data and the numbers of matches or overlaps obtained including the M (the size of the smaller dataset) and n (the estimated number of matched pairs) values from Section 2.3. The table also shows how the prior probability can be revised based on more precise estimates of the number of matches.

(Table 5 around here)

In the next step, the three sets of generated pairs are combined and filtered for duplicates. Duplicates occur when for example a police case is matched with a hospital case and separately with an IB case. Putting these together gives the set of the individuals in all three datasets. In the final step, the original files from the three sources are added to the file of those in two or three datasets, duplicates eliminated and needed variables are transferred to the retained case record. In the resulting dataset (the Combined Injuries Dataset) each case refers to a single individual and the data for this individual contains the relevant information from each of the original data files.

Figure 3 illustrates the results from the linking of the three datasets for the period 2010-2011. The addition of the Injuries Board data to the two other sets identifies a large number of additional injuries. This additional set has a match rate (or overlap) equivalent to 30% of the police data but only 6% with the hospital data. This is an indication that many of the injuries are minor ones; nevertheless they are all certified by a doctor and, on average, a compensation award of over €20,000 is paid. While there have been allegations of fraud and claims farming (Injuries Board Ireland, 2013) there can be little doubt that the social and economic cost of these injuries is significant and they need to be taken into account in an assessment of the impacts of road traffic. The compensation paid is not a full measure of this impact but amounts to around €150 million annually. The total number of distinct persons injured between the three sources is nearly 49,000, significantly more than in the police and hospital datasets combined and more than three times the official published figure.

(Figure 3 around here)

The Combined Injuries Dataset allows analysis of the total number of injuries by gender, age, location, the number of dataset appearances and injury severity. Table 6 shows these aggregates and the ascertainment rates for the three datasets. Persons appearing in all datasets compared to those in just one are more inclined to be male (62% against 54%), to be slightly older (35 years old compared to 33, with 9% over 65 compared to 5%) and are much more likely to be pedestrians (24%

compared to 5%). This latter point suggests that if police are called to a pedestrian accident, it is relatively more likely to involve hospitalisation and a compensation claim.

(Table 6 about here)

The ascertainment rates (the shares of the totals accounted for by each of the datasets) show how this third source dominates the others; the rates for the police and hospitals are less than a half of the figures seen in Table 4. This illustrates the risk of assessing injuries only on the basis of the two main sources. The ascertainment rates vary by road user group with the Injuries Board rates being higher for females, for those of middle age and those in Dublin and lower for the youngest and oldest age groups. The hospital and police rates reflect the earlier discussion on the two dataset linkage except that the new rates are much lower because of the addition of the third dataset. The greater predominance of females and people aged from 35-54 in the Injuries Board data may, tentatively, be interpreted as a tendency for these groups to be less at fault for accidents than some of the other groups more predominant in police data. The significance of the oldest and youngest age groups in the hospital data illustrates at least partially, the tendency to admit these people to hospital for observation.

Comparing severity between the datasets is not directly possible as no single measure applies across the datasets. The different measures - a serious injury as assessed by police at the scene, Maximum Abbreviated Injury Score (MAIS) and length of stay (LoS) for hospital patients and size of compensation award for Injury Board claimants can indicate more severe injuries in each of the datasets. For the different measures, a threshold to define an injury as serious is needed. Police data already has the serious and minor categorisation and MAIS3+ is an internationally accepted breakpoint for clinically coded injuries. Length of stay (LoS) is set at 6 days or more (LoS6+); this is used in several countries as the definition of a serious injury (International Transport Forum, 2012). For the Injuries Board data, a threshold of €20,000 is used and injuries for which awards are above this level are defined to be serious. This is not an insignificant amount, and, for example, it is close to the annual disposable income per person in Ireland (Central Statistics Office, Ireland, 2014). When the same individual is in more than one dataset, the different severity assessments can be compared. Table 7 summarises the results showing how the different serious injury indicators are correlated but by no means fully so. For example, a hospitalised person with a police reported serious injury has an equal probability of being classed as MAIS3+ or MAIS<3. A hospitalised person seems to have a larger chance of getting a large compensation award with a MAIS<3 than with MAIS3+ as does a person with a minor rather than serious police reported injury. However, this should be treated with care because of the risk of biases in the Injuries Board data because of the zero award category. In each of the overlap datasets, between 30% and 40% of the total cases are shown to be seriously injured by one measure and not seriously injured by the other.

(Table 7 around here)

A composite severity measure, which defines a serious injury as one that is serious by *any* of the measures, can also be analysed. This composite measure (called Bad Injury) shows in Table 8 how the likelihood of appearing in more than one dataset increases with the severity of the injuries. It shows also that, by this measure, there are about 5 times as many people seriously injured compared to police-reported data. This is illustrative of the complexity of the injury problem, underlining how any single measure will not capture the full range of impacts.

(Table 8 around here)

4.3 Capture-recapture Estimates

Capture-recapture methods are used to estimate population sizes from a number of samples of the population (Amstrup, 2005; ITF, 2012). A wide variety of models are available and there is an extensive literature especially for estimating animal populations. In road safety research a simple estimator called the Lincoln-Petersen (L-P) is often used to estimate the total population of fatalities or injuries based on the records on two lists, usually police and a medical source. Examples of its use in safety research include Amorós et al.(2007), Meuleniers et al.(2006), Tercero & Anderson (2004), Morrison & Stone (2000), Razzak & Luby (1998). The four conditions for correct use of the L-P estimator are that the population is closed, that there is perfect identification of individuals, that the samples are independent and that in each sample the selection probability is constant (IWGDMF, 1995). Under these conditions the population size is estimated by taking the product of the two file sizes and dividing by the overlap. Because of its simplicity there is a temptation to use the L-P estimator even though the conditions for its correct use may not be met.

Two illustrations of the risk of using the L-P estimator can be given from the data in this study.

First, L-P estimates might be made of the size of the injured population using the police and hospital data from this study. These are shown in Table 9 for different subgroups of the population. Because this study includes a third data set there is a broader view of the total injured population as shown by the Combined Injuries Dataset which is the set of all unique injuries in the three datasets after applying record linkage. It is not a complete set of all injuries but it might be considered as a reasonable first approximation. The numbers in the Combined Injuries Dataset can be compared with the estimates obtained using the L-P estimator. The results are also shown in Table 9.

(Table 9 around here)

First, notice that the totals in the Combined Injuries Dataset and the L-P estimates are actually reasonably close for the total and for some specific groups (for the year 2011, for females, for those aged 25-34). The L-P estimator overestimates the total in the Combined Injuries Dataset total by only 4% but the difference ranges from very large overestimate for some counties, through a 28% overestimate for 0-14 year olds to a 23% underestimate for those aged from 35-44. Though the aggregates are close, it can be seen that there are quite large differences in the components of these aggregates. The third dataset from the Injuries Board has different characteristics from the police and hospital data and the L-P estimator cannot pick these up. This lesson that the unseen cases may not resemble the seen ones has been forcibly underlined in the past, for example by Cormack (1999) and Cormack et al.(2000).

A second illustration of the problems with using the L-P estimator is to consider the three datasets as samples from the injured population. Then taking the datasets in pairs, three separate (L-P) estimates can be calculated. Table 10 summarises the results, showing for each pair of datasets, the numbers in each set, the overlaps and the L-P estimator. Assuming data were available only from police and hospitals, the L-P estimator gives an estimated injured population of over 51,000 (in the first line of Table 10). If the third source is included and the L-P estimator is calculated from police and Injuries Board data, the estimate of the total injured population doubles to over 100,000. When it is applied to the totals in the hospital and Injuries Board datasets it more than trebles to over 150,000. If the conditions for correct use of the L-P estimator were met (that the samples were independent random samples from the total population.(Hook & Regal, 1997; IWGDMF, 1995).), then these estimates should all be the same.

(Table 10 around here)

Clearly, some assumptions underlying the correct use of the L-P estimator are not met. Specifically the assumptions on independence between samples and on homogeneous probabilities for each data set are not correct. From the linked data it has already been seen that certain characteristics in one dataset are associated with higher or lower probabilities of presence in another. The most clear cut example is of those seriously injured in police-reported crashes who are much more likely to be admitted to hospital. In the other direction, cyclists' presence in the hospital dataset is associated with a very low likelihood of being in the police dataset. The Injuries Board dataset is more likely to have urban residents and people aged between 35 and 64. This lack of independence also implies heterogeneous appearance probabilities.

Since the conditions for application of the L-P estimator are not met the true number of people injured could be more or less than the numbers indicated by the formula. In such cases, theory recommends stratifying the population into homogeneous groups and then applying the estimator to these groups (IWGDMF, 1995). Injury severity is an obvious candidate for such stratification. However, this is not possible with the data used here as there is not a consistent definition of a serious injury across the datasets. An illustrative approach could be to use each dataset's definition of serious injury and then apply the L-P estimator. This is shown also in Table 10 and results in very different estimates of the number of seriously injured people. In short, little credence can be placed on such calculations.

These examples demonstrate that the L-P estimator cannot be relied upon when the independence and homogeneity conditions are not met. The third dataset available here illustrates how it is not possible to make deductions about the total injured population from the police and hospital data alone. For three or more datasets other modelling techniques, which allow for heterogeneity within the samples and lack of independence between them as in the literature on animal populations (Otis et al. 1978; Pledger, 2000; Chao 2001; Rivest and Baillergeon 2007; Huggins and Whang, 2011) might provide improved estimates but are beyond the scope of this paper.

5 Discussion and Conclusions

The following discussion examines first some limitations of the methods and data presented above. Then the results are summarised and some policy implications are drawn out.

- Limitations

To the authors' knowledge this is the first attempt in Ireland to use record linkage to bring injury data from different sources together and to provide an assessment on the extent of the traffic injury problem. The three datasets provide significantly more information than official data and show how the scale of the injury problem is far greater than officially indicated. However, work to understand injuries is at an early stage and there are a number of data and other limitations to the work that need to be understood and remedied in the future.

The first group of limitations concerns the data. The datasets are obtained in different ways for different purposes. The statistical units used do not have unambiguous definitions. Moreover, individuals do not have unique identifiers and some of the variables available for comparison in record linkage are not ideal. However, there is scope to improve all three datasets to allow improved quality of matching. Specifically, if police data can provide full date of birth, the county of residence and can eliminate non-responses on key variables, there would be a significant improvement to the power of the matching process. Hospital data is already of good quality with a high completion rate and few obvious errors. A key improvement here will be to the coding of traffic and non-traffic cases. At present, including all V codes in the analysis overstates the true number of road traffic

injuries. But excluding all accidents coded as non-traffic accidents would be incorrect. At present about 40% of hospital patients are categorised as non-traffic injuries. If these were all non-traffic accidents the aggregate in the combined dataset for the number of traffic injuries for 2010 and 2011 would fall by around 4,000 from 49,000 to 45,000. Without additional information the true number of non-traffic injuries cannot be known but based on the matching results it is unlikely to be more than 2,000. In any case these injuries are relevant in a general policy sense as they require public resources and have significant personal and social costs. More generally, understanding and dealing effectively with serious injuries will benefit greatly from improved trauma data and analysis. A review of trauma arrangements has been undertaken in Ireland (Deasy et al., 2016) and participation in the Trauma Audit and Research Network (TARN) as well as use of standardised templates like the Utstein template (Ringdal et al., 2008) should all contribute to this aim in the future.

The Injuries Board is a relatively new data source and at present contains limited information. It is also subject to administrative influences and public knowledge on the system. For the future, Injuries Board data will be improved if the transport mode is included and some clinical coding is undertaken so that a clearer understanding of the severity of the injuries can be obtained. More clarity in the cases where no awards are paid by the IB would also help. These improvements to the three datasets will strengthen the linkages by increasing the m values (the probability of a link given a true match) and reducing the u values (the probability of a link given a non-match), thereby increasing the posterior probabilities and reducing both kinds of statistical error.

Each dataset has its own characteristics and biases. Bringing the three datasets together probably reduces these biases but since there is no 'gold standard' consisting of a true dataset of injuries it is not possible to say what biases remain. Injuries not reported to the police not involving hospitalisation and where no injury claim is made would still be excluded. One evident category that these could include is less serious single vehicle accidents. At present there are no other national datasets on injuries apart from the three analysed here. Other sources like general practitioners, Accident and Emergency units, Insurance companies or ambulance services could provide additional insights but there are at present no data from them.

The second set of limitations concerns the process to match the datasets. This process, particularly for the three datasets, is complicated, with its combination of probabilistic matching using a record linkage programme, enhancement to the matching and conversion of the scores into probabilities in order to select "best" sets and finally deterministic matching at the second stage to create the set of unique individuals-the Combined Injuries Dataset. Different matching software or a different combination of matching methods could possibly improve or simplify the process and could also give different answers. Tests on the sensitivity of the results to the weights showed that the results would indeed vary but not dramatically. The majority of pairs or triples deemed to be matches have high posterior probabilities and match on all, or all except one, or at most two, of the variables. Varying the parameters slightly does not alter most of these matches. Another programme or an improved matching process would, in all likelihood, generate a matched set with at least 75% ,and probably more, of the same pairs and triples.

The estimated numbers of false positives and false negatives for the 2010-2011 linkage are each over 10% of the size of the smaller of the two datasets. Reducing these requires especially better quality data on the existing variables as set out above. At present the linkages in the Irish case may be at the lower end of the scale internationally in terms of statistical precision because of the smaller number of comparable data fields and their weaker discriminatory power. Despite the high statistical errors, the confidence intervals for total matches are quite narrow and the numbers are probably a reasonable starting point for discussion and analysis.

Notwithstanding the data and process difficulties, record linkage provides useful insights. The Combined Injuries Dataset is a helpful construct when clerical review of potential matches is not possible. It is an addition to methods like the matching standards method used in the literature and allows analysis of a specific data set. It facilitated the second stage matching which would have been more complex without it. The method can work with the scores from matching programmes as long as these scores are correctly based on the m and u values from the data, allowing them to be converted into probabilities from the formula in Section 2. The matching is undertaken like sampling without replacement, where once a pair is matched it is not considered further. The statistical and matching implications of this need further study, possibly including simulations using other matching programmes. Different programme algorithms may give different results. But from the general discussion above, the majority of pairs would not be affected. But such work will be useful as matching of three or more sets is likely to become more common. It should be noted also that the Combined Dataset can also be adapted by restricting it, for example to perfect matches, thereby allowing analyses where false positives are unlikely.

- Results

The rather complicated technical aspects of the linking should not result in sight being lost of the main points emerging from the analysis, which are rather robust. The police and hospital data over the period 2005-2011 indicate matching rates of about 20% of the police total and about 30% of the hospital total with significant variation between different categories. The total number of injuries identified between the two sources is over 88,000, 50% more than the official data. Police ascertainment of the number of injuries identified by the two sets varies significantly. The lowest is for cyclists where the police ascertainment rate is only 26% and the highest for car occupants where it is nearly 80%.

An illustration of the potential use of record linkage is illustrated in considering those seriously injured in the police and hospitals datasets. When the overlap sets are examined it can be seen that there is a growth in the share of police reported minor injuries that are classed as MAIS3+ and an increase in the likelihood of a police-reported serious injury being MAIS3+. This shows that the police severity measure is not consistent over time, and appears to be changing towards the inclusion only of more serious injuries (as measured by MAIS) but with significant misclassification of clinically serious injuries.

The third data source, the Injuries Board, identifies a large number of additional injuries not seen in police or hospital sources. These injuries show structural differences from the other datasets and this new dataset has a low overlap especially with hospital data. It contains more women, more people aged from 35-54 and has a more urban concentration. This third dataset demonstrates how using only police and hospital injury data results in a significant underestimation of the true total. The total number of injuries between the three sources is about 25,000 annually, over three times the number identified by police reports and five times the number hospitalised. The implications are that the injury problem is both significantly larger and structurally different than official figures show.

The analysis shows how reasonable definitions of serious injury can be made for each of the datasets used here. The linkages show that if someone is seriously injured by one measure the odds of being seriously injured by another increase. But these serious injury datasets by no means overlap and all possible combinations of people being seriously injured by one measure and not by another actually occur. This underlines that measuring serious injuries is complex and cannot be satisfactorily undertaken by a single measure.

The dataset provides a more comprehensive set of injuries than the traditional police-reported injuries or the less frequently used hospital or Injury Board data. Nevertheless, it does not include all traffic injuries, missing those not reported to the police and that do not involve a hospital stay or an injury claim. In Ireland, as in most countries, there are no comprehensive national datasets on injuries and the true total remains unknown.

Attempts to estimate the total number of injuries using capture recapture methods and in particular the L-P estimator are likely to be inaccurate. The independence and homogeneity conditions necessary for correct use of the estimator are not met and even stratification of the population into more homogeneous groups does not provide estimates that could be relied on. The extrapolation of characteristics of those in the existing datasets to those outside would not appear to be justified. The conclusion is that data from other sources are needed to obtain a reliable picture of the injury problem.

- Implications for Policy

Road traffic Injuries are a subject of growing interest in road safety policy. But most countries do not have a comprehensive view of the extent and nature of the problem. International data are not comparable and at national level the different sources have problems of compatibility and completeness. Countries are at different stages with some already having extensive experience in combining data sources and sophisticated and continually improving systems to study them, for example in Sweden with the STRADA system (Howard & Linder, 2014) or the United States with CIREN data (Elliott et al, 2010). However, many other countries, including Ireland are at an early stage in bringing different sources together and still rely entirely on the police for official injury data. In this sense, the present study can contribute to a process gradually to improve the data and better understand the problem.

There are also some more specific policy conclusions. The most evident is that police data alone are not a satisfactory way to measure the extent or nature of the road traffic injury problem. The hospital data identify additional and different injuries with limited overlaps. A key feature of the hospital data is the significant numbers of vulnerable users especially cyclists and motorcyclists who are not seen in the police data. Comparisons between injury severity measures for those in both datasets, show that the number of clinically serious injuries is about twice the police-reported number and that the police definition of serious injury is changing over time. Of the vulnerable traffic users, cyclists are the most striking with far larger numbers hospitalised than shown by police. This underlines how the national policy to encourage cycling needs to pay more attention to safety.

When the third dataset is added to the police and hospital cases, a large number of entirely new injuries is seen. Even if many of these are minor accidents they have a large cost for the individuals and for society. These data need to be taken into account in injury statistics but also improved to provide more reliable and complete information. This implies that new injury indicators are needed since existing data do not capture the extent and variety of the problem. While international comparability will be improved by the use of MAIS3+ as the indicator of serious injuries, it will not provide a comprehensive summary of the injury problem. Other indicators will be needed and record linkage could be used to provide new data series, even informal ones, to monitor trends and progress. Efforts to improve the data and continue to strengthen the linkages would therefore be worthwhile. While the third dataset used here from the Injuries Board is specific to Ireland other countries may have data from sources other than police and hospitals; the lesson here is that these need to be studied, using record linkage if possible, as they seem likely to include injuries not seen in the other datasets.

The evidence that the numbers are far greater than the official data indicate implies that reducing injuries needs to play a more important role in road safety strategy. Policy measures under consideration to reduce fatalities could obviously also contribute to reducing injuries. Among these are helmets for cyclists, lower urban speed limits, stronger measures to protect pedestrians, cyclists and motorcyclists. The large number of injuries previously not counted in economic analysis of the benefits of additional measures mean that the returns on some specific measures will be higher than previously considered. One particular injury category, with a large number of claims and high costs is soft tissue or whiplash injuries and in this regard measures to understand better the severity of the injuries and also to reduce the number of low speed car collisions could be cost effective. Given the large number of these injuries, such measures would appear to be justified and have higher socio-economic returns than previously thought, thereby strengthening the case for their introduction.

A final conclusion follows from the fact that, though the evidence presented here shows that the number of injuries is much higher than police-reported figures, not enough is known on the nature of injury accidents nor on their consequences. Research is needed better to understand these injuries and assess the policies to reduce their number and gravity.

Acknowledgements

The Road Safety Authority, the Injuries Board and the Economic and Social Research Institute are thanked for the provision of the data for this work. Thanks are also due to the anonymous referees for their helpful suggestions.

References

- Alsop, J., & Langley, J. (2001). Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accid Anal Prev*, 33, 353-359.
- Association for the Advancement of Automotive Medicine (2008). The Abbreviated Injury Scale 2005. Update 2008 *Des Plaines, IL*.
- Amstrup, S. C., McDonald, T. L., Manly, B. F. J. (editors) (2005). *Handbook of Capture-Recapture Analysis*: Princeton Press.
- Amoros, E., Martin, J. L., & Laumon, B. (2007). Estimating non-fatal road casualties in a large French county, using the capture-recapture method. *Accid Anal Prev*, 39, 483-490.
- Bhalla, K., Shahraz, S., Naghavi, M., Lozano, R., & Murray, C. (2008). Estimating the distribution of external causes in hospital data from injury diagnosis. *Accident Analysis & Prevention*, 40(6), 1822-1829.
- Brenner, H. (1994). Application of capture-recapture methods for disease monitoring: potential effects of imperfect record linkage. *Methods Inf Med*, 33(5), 502-506.
- Central Statistics Office, Ireland (2014). Survey on Income and Living Conditions, 2012 results.
- Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological*

- and *Environmental Statistics*, 6(2), 158-175.
- Clark, D. E. (2004). Practical introduction to record linkage for injury research. *Injury Prevention*, 10, 186-191.
- Cryer, C., & Langley, J. (2006). Developing valid indicators of injury incidence for "all injuries". *Injury Prevention*, 12, 202-207.
- Cormack, R. M. (1999). Problems with using capture-recapture in epidemiology; An example of a measles epidemic. *Journal of Clinical Epidemiology*, 52(10), 917-926.
- Cormack, R. M., Chang, Y. F. and Smith, G. S. (2000). Estimating deaths from industrial injury by capture-recapture: a cautionary tale. *International Journal of Epidemiology*, 29(6), 1053-1059.
- Deasy, C., Cronin, M., Cahill, F., Geary, U., Houlihan, P., Woodford, M., Crowley, P. et al. (2016) Implementing Major Trauma Audit in Ireland. *Injury*, 47(1), 166-172.
- Department for Transport UK. (2006). Underreporting of Road Casualties ; *Road Safety Research Report 69*.
- Department for Transport UK. (2012a). Reported Road Casualties in Great Britain; a valuation of road accidents and casualties in Great Britain in 2011.
- Department for Transport UK. (2012b). Linking Police and Hospital data on Road Accidents in England, 1999 to 2009.
- European Centre for Injury Research, University of Navarra. Mapping from ICD-10 to AIS.
- Economic and Social Research Institute. (2013). Activity in Acute Public Hospitals in Ireland, 2012 Annual Report.
- Elliott, M. R., Resler, A., Flannagan, C. A., & Rupp, J. D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis & Prevention*, 42(2), 530-539.
- European Commission. (2008). Safety Net ; Estimation of real number of road accident casualties; final report.
- Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.
- Felligi, I. P. (1997). Record Linkage and Public Policy – A Dynamic Evolution. Proceedings of an International Workshop and Exposition. Federal Committee on Statistical Methodology; Arlington, VA, March 20-21, 1997.
- Fingerhut, L. A. (2004). International Collaborative Effort on Injury Statistics: 10 year review.. *Injury Prevention*, 10, 264-267.
- Gennarelli, T. A., & Wodzin, E. (2006). AIS; a contemporary injury scale. *Injury Prevention*, 37(12), 1083-1091.

- Gill, M., Goldacre mJ., & DG., Y. (2006). Changes in Safety on Englands Roads;analysis of hospital statistics. *Br Med J*, 333, 73.
- Healthcare Pricing Office (2015). <http://www.hpo.ie/>
- Hook, E. B. & Regal, R. R. (1997). Validity of methods for model selection, weighting for model uncertainty, and small sample adjustment in capture-recapture estimation. *American Journal of Epidemiology*, 145(12), 1138-1144.
- Hook, E. B., & Regal, R. R. (1999). Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *J Clin Epidemiol*, 52(10), 917-926; discussion 929-933.
- Howard, C., & Linder, A. (2014). Review of Swedish experiences concerning analysis of people injured in traffic accidents: VTI.
- Huggins, R. and Hwang, W. H. (2011). A Review of the Use of Conditional Likelihood in Capture-Recapture Experiments. *International Statistical Review*, 79(3), 385-400.
- Injuries BoardIreland (2013). Annual Report and Accounts, 2012.
- Injuries Board (2015). Information on Making a Claim
<https://piab-injuriesboard.ie/Pages/Step1.aspx>
- IRTAD (2012) Annual report, 2011. Paris: OECD.
- International Transport Forum, (2012). IRTAD. Reporting on Serious Road Casualties. Paris: OECD.
- Irish Statute Book (2015). Road Traffic Act, 1961 10106.
http://www.irishstatutebook.ie/ResultsTitle.html?q=Road+Traffic+Act+1961&search_type=all&button=Search (Accessed, 26 Sept.2015).
- IWGDMF (1995). International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: history and theoretical development. *American Journal ofEpidemiology*, 142(10), 1047-1058.
- Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Stat Med*, 14(5-7), 491-498.
- Jeffrey, S., Stone, D. H., Blamey, A., Clark, D., Cooper, C., Dickson, K., Major, K. (2009). An evaluation of police reporting of road casualties. *Injury Prevention*, 15(1), 13-18.
- Kudryavtsev, A.V., Ermolina, M., Lund, J., Kleshchinov, N., Grijibovski, A., Nilssen, O., Ytterstad, B., Road traffic fatalities in Arkhangelsk, Russia in 2005–2010: Reliability of police and healthcare data. *Accident Analysis and Prevention*, 53, 46-53.
- Langley, J., & Brenner, R. (2004). What is an Injury? *Injury Prevention*,, 10, 69-71.
- Linkage-Wiz. (2013). User Manual.
- Lopez, D. G., Rosman, D. L., Jelinek, G. A., Wilkes, G. J., & Sprivulis, P. C. (2000). Complementing police road-crash records with trauma registry data — an initial evaluation. *Accident Analysis & Prevention*, 32(6), 771-777.

- Ministry of Transport New Zealand (2012). Social cost of road crashes and injuries June 2011 update.
- Meuleners, L. B., Lee, A. H., Cercarelli, L. R., & Legge, M. (2006). Estimating crashes involving heavy vehicles in Western Australia, 1999-2000: a capture-recapture method. *Accid Anal Prev*, 38(1), 170-174. doi: 10.1016/j.aap.2005.09.003
- Morrison, A., & Stone, D. H. (2000). Capture-recapture: a useful methodological tool for counting traffic related injuries? *Injury Prevention*, 6(4), 299-304.
- Newcombe, H. B., Kennedy, J. M., Axford, S. J., & P., J. (1959). Automatic Linkage of Vital Records. *American Association for the Advancement of Science*, 130, 954-959.
- Otis, D. L., Burnham K. P., White, G. C. and Anderson D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.
- Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2), 434-442.
- Razzak, J. A., & Luby, S. P. (1998). Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method. *Int J Epidemiol*, 27(5), 866-870.
- Reurings, M. C. B., & Stipdonk, H. L. (2011). Estimating the Number of Serious Road Injuries in The Netherlands. *Annals of Epidemiology*, 21(9), 648-653.
- Ringdal, K. G., Coats, T. J., Lefering, R., Di Bartolomeo, S., Steen, P. A., Røise, O., Lossius, H. M. (2008). The Utstein template for uniform reporting of data following major trauma: A joint revision by SCANTEM, TARN, DGU-TR and RITG. *Scand J Trauma Resusc Emerg Med*, 16(1), 1-19. doi: 10.1186/1757-7241-16-7
- Rivest, L. P. and Baillargeon, S. (2007). Applications and extensions of Chao's moment estimator for the size of a closed population. *Biometrics*, 63(4), 999-1006.
- RSA (Road Safety Authority, Ireland)(2013). Road Collision Facts, 2011.
- Short, J., Caulfield, B, The safety challenge of increased cycling, *Transport Policy*, 33, 2014, 154 – 165.
- SWOV (2014). Road Crash Costs. *Fact Sheets*.
http://www.swov.nl/rapport/Factsheets/UK/FS_Costs.pdf (accessed 26 Sept 2015)
- UNECE (2009) United Nations, Economic Commission for Europe Glossary of Transport Statistics.
- Winkler, W. E. (1999). The State of Record Linkage and Current Research Problems. Governmental Statistical papers, Canada. <https://www.census.gov/srd/papers/pdf/rr99-04.pdf> (Accessed 26 Sept. 2015).
- WHO (1990). World Health Organisation. International Classification of Diseases, 10th Revision. Geneva. <http://www.who.int/classifications/icd/en/>. (Accessed, 26 Sept. 2015).

List of Tables

Table 1. International Fatality and Injury Ratios and Trends

Table 2. Datasets and Comparison Variables for Three Datasets

Table 3. Illustration of m and u values and Agreement and Disagreement Weights for Police-Hospital Linkage

Table 4. Totals and Matching and Ascertainment Rates for Police-Hospital Linkage, 2005-2011

Table 5. Dataset Sizes, Overlaps and Prior Probabilities for Linkage 2010-2011

Table 6. Injuries and Ascertainment Rates from Three Datasets 2010-2011

Table 7. Injury Severity for Dataset Overlaps

Table 8. Bad Injuries for Combined Injuries Dataset, 2010-2011

Table 9. L-P Estimates Compared with Combined Injuries Dataset

Table 10. L-P Estimates of Injuries from Pairs of Datasets

Table 1. International Fatality and Injury Ratios and Trends

Country	Fatalities (F)	Injuries (I)	Ratio (I/F)	% Change , 2000 - 2010	
				Fatalities	Injuries
Austria	552	45,858	83	-43	-16
Belgium	840	60,380	72	-36	-15
Canada	2,227	170,629	77	-24	-21
Denmark	255	4,153	16	-49	-52
France	3,992	84,461	21	-51	-45
Germany	3,648	371,170	102	-51	-25
Greece	1,258	19,108	15	-38	-35
Hungary	740	20,917	28	-38	-7
Ireland	212	8,270	39	-49	-25
Italy	4090	302,735	74	-42	-18
Japan	5745	895,326	156	-45	-22
Korea	5505	352,458	64	-46	-22
New Zealand	375	14,031	37	-19	+39
Norway	208	8,924	43	-39	-25
Spain	2478	11,6503	47	-57	-16
Sweden	266	23,307	88	-55	+5
Switzerland	327	24,237	74	-45	-17
United Kingdom	1905	215,700	113	-47	-35
United States	32885	223,9074	68	-22	-25

Source: IRTAD, 2012.

Table 2. Datasets and Comparison Variables for Three Datasets

Data Source	File Sizes	Comparison Variable				
		Incident Date	Age	Gender	Location	Mode
Police	15,131	Exact	In Years 5% not stated	5% not stated	County of incident	Police classification
Hospitals	9,559	Exact or 1 or 2 days later	Month and Year of Birth	all	County of residence	ICD-10 classification
Injury Claims	32,492	Exact	Full date of birth (3% not stated)	all	County of residence	none

Sources: police hospital and Injuries Board datasets.

Table 3. Illustration of m and u values and Agreement and Disagreement Weights for Police-Hospital linkage

Variable	m value	u value	m/u	Agreement Weight $\text{Log}_2(m/u)$	Disagreement Weight $\text{Log}_2(1-m)/(1-u)$
Date	.97	.0005	1,940	10.9	-5.1
Age	.91	.01	91	6.5	-3.5
Gender	.94	.5	1.88	.9	-3.1
Mode	.88	.2	4.4	2.1	-2.7
County	.85	.038	22.4	4.5	-2.7

Sources: police, hospital and Injuries Board datasets; linkage program.

Notes:

Table 4. Totals and Matching and Ascertainment Rates for Police-Hospital linkage, 2005-2011

Category		No. of Cases ⁽¹⁾				Match Rates ⁽²⁾ (%)		Ascertainment Rates ⁽³⁾ (%)	
		Police	Hosp.	Matches	Total ⁽⁴⁾	Police	Hosp.	Police	Hosp.
Gender	Male	30,966	26,368	7,266	50,068	23.5	27.6	61.8	52.7
	Female	23,044	14,453	4,107	33,390	17.8	28.4	69.0	43.3
Road User Type	Pedestrian	6,795	4,414	1,956	9,253	28.9	44.3	73.4	47.7
	Cyclist	2,133	6,478	531	8,080	24.8	8.2	26.4	80.2
	Motor-C	2,885	3,875	1,042	5,718	36.1	26.9	50.5	67.8
	Car	40,690	17,869	7,022	51,537	17.3	39.3	79.0	34.7
	Other	6,291	8,186	959	13,518	15.2	11.7	46.5	60.6
Year	2005	8,600	6,842	1,704	13,738	19.8	24.9	62.6	49.8
	2006	8,342	6,464	1,795	13,011	21.5	27.8	64.1	49.7
	2007	7,677	6,505	1,646	12,536	21.4	25.3	61.2	51.9
	2008	9,521	6,028	1,746	13,803	18.3	29.0	69.0	43.7
	2009	9,563	5,423	1,766	13,225	18.5	32.6	72.3	41.0
	2010	8,053	5,011	1,514	11,550	18.8	30.2	69.7	43.4
	2011	7,078	4,548	1,338	10,288	18.9	29.4	68.8	44.2
Severity	Serious (P)	5,295		2,708	5,295	51.1			
	Minor (P)	53,499		8,801	53,499	16.5			
	MAIS 3+		7,667	2,500	7,667		32.6		
	MAIS<3		23,856	6,399	23,856		26.8		
Age	0-14	4,611	8,067	1,425	11,253	30.9	17.7	41.0	71.7
	15-24	15,325	10,797	3,380	22,742	22.1	31.3	67.4	47.5
	25-34	12,679	7,575	2,037	18,217	16.1	26.9	69.6	41.6
	35-44	8,130	4,825	1,677	11,278	20.6	34.8	72.1	42.8
	45-54	5,629	3,583	1,143	8,069	20.3	31.9	69.8	44.4
	55-64	3,808	2,587	795	5,600	20.9	30.7	68.0	46.2
	65+	3,828	3,387	1,049	6,166	27.4	31.0	62.1	54.9
Collision Type	Single Vehicle(P)	13,280		3,316		25.0			
	Multiple Vehicle(P)	27,289		4,256		15.6			
	Crash with Object (H)		4,357		1,801		41.3		
	With other vehicle (H)		7,220		3,030		42.0		
	Traffic (H)		22,522	8,194	22,522		36.4		
	Non-Traffic		18,300	2,834	18,300		15.5		
County	Dublin	11,647	7,108	1,797	16,958	15.4	25.3	68.7	41.9
	Cork	6,143	4,444	1,535	9,052	25.0	34.5	67.9	49.1
	Galway	3,160	2,332	593	4,899	18.8	25.4	64.5	47.6
	Donegal	2,888	2,065	721	4,232	25.0	34.9	68.0	48.8
	Kildare	2,193	1,730	388	3,535	17.7	22.4	62.0	48.9
	Limerick	3,151	1,440	459	4,132	14.6	31.9	76.3	34.8
	Wicklow	1,828	1,050	265	2,613	14.5	25.2	70.0	40.2
	Wexford	2,146	1,784	555	3,375	25.9	31.1	63.6	52.9
	Others	25,628	18,869	5,196	39,301	20.3	27.5	65.2	47.5
Total		58,784	40,822	11,509	88,097	19.6	28.2	66.7	46.3

Notes:

- (1) The sum of the number of cases shown in the different categories does not necessarily give the same aggregate due to cases where category is not stated.
- (2) Match Rates for police and hospitals are the number of matches as a percentage of the police and hospital totals for each of the categories.
- (3) Ascertainment rate for police and hospitals are the percentages of the number of distinct cases (shown in the total column) that police or hospital cases in the category account for.
- (4) The total is the sum of the police and hospital cases minus the number of matches.

Table 5. Dataset Sizes, Overlaps and Prior Probabilities for Linkage 2010-2011

Dataset 1	Dataset 2	Dataset Size		Overlap ⁽²⁾	Prior probability	
		Dataset 1	Dataset 2 ⁽¹⁾		Estimate ⁽³⁾	Revised
Police	Hospital	15,131	9,559	2,861	1.38×10^{-5}	1.978×10^{-5}
Injuries Board	Police	32,492	15,131	4,853	1.017×10^{-5}	9.871×10^{-6}
Injuries Board-	Hospital	32,492	9,559	2,054	6.439×10^{-6}	6.613×10^{-6}

Sources: police, hospitals and Injuries Board datasets; linkage results.

Notes:

- (1) Dataset 2 is the smaller dataset in each linkage and its size is the maximum number of matches (the value M in Section 2.3).
- (2) Overlap is the sum of the posterior probabilities (n) as explained in Section 2.3.
- (3) The prior probability estimates are calculated on the basis of 2000, 5000 and 2000 matches respectively. The revised figure use the actual overlaps.

Table 6. Injuries and Ascertainment Rates from Three Datasets 2010-2011

Category	Police	Hospitals	Injuries Board	Combined	Ascertainment Rates(%)		
					Police	Hosp.	IB
2010	8,053	5,011	16,160	24,777	32.5	20.2	65.2
2011	7,078	4,548	16,332	24,004	29.5	18.9	68.0
Male	8,379	6329	16,878	26,762	31.3	23.6	63.1
Female	6,376	3230	15,614	21,736	29.3	14.9	71.8
0-14	1,356	2,134	2,987	5,618	24.1	38.0	53.2
15-24	3,840	2,123	6,226	10,149	37.8	20.9	61.3
25-34	3,519	1689	8,321	11,769	29.9	14.4	70.7
35-44	2,306	1127	6,351	8,383	27.5	13.4	75.8
45-54	1,610	955	4,089	5,631	28.6	17.0	72.6
55-64	1,111	683	2,364	3,474	32.0	19.7	68.0
65+	1,095	848	1,233	2,540	43.1	33.4	48.5
Bad Injury ⁽¹⁾	2,410	3,235	5,113	7,789	30.9	41.5	65.6
County ⁽²⁾							
Dublin	3,293	1,671	8,993	11,995	27.4	13.9	75.0
Cork	1,619	1,080	3,605	5,252	30.8	20.6	68.6
Galway	939	500	1,985	2,881	32.6	17.4	68.9
Donegal	669	440	937	1,657	40.4	26.6	56.5
Kildare	614	479	1,146	1,910	32.1	25.1	60.0
Limerick	791	377	2,286	2,975	26.6	12.7	76.8
Wicklow	507	244	663	1,231	41.2	19.8	53.9
Wexford	480	462	622	1,269	37.8	36.4	49.0
Others	6,219	4,306	12,255	19,611	31.7	22.0	67.5
Total	15,131	9,559	32,492	48,781	31.0	19.6	66.6

Sources: Police, Hospital and Injuries Board datasets, Combined Injuries dataset

Notes:

(1) Bad Injury is defined separately for each dataset; For Police it is the police assessment; For Hospitals, MAIS3+ and for IB it is an award over 20K Euros.

(2) Data from eight counties, those with the largest totals, are shown.

Table 7. Injury Severity⁽¹⁾ for Dataset Overlaps

Police-Hospital Overlap			
	Serious (%)	Minor (%)	Total (%)
MAIS3+	272 (44.3)	413 (18.4)	685 (23.9)
MAIS<3	276 (45.0)	1,260 (56.1)	1,536 (53.7)
MAIS NS	66 (10.7)	574 (25.5)	640 (22.4)
Total	614 (100)	2,247 (100)	2,861 (100)
Police-Injuries Board Overlap			
	Serious(%)	Minor (%)	Total(%)
Big Award	105 (24.6)	962 (21.7)	1,067 (22.0)
None or small	321 (75.3)	3,465 (78.3)	3,786 (78.0)
Total	426 (100)	4,427 (100)	4,853 (100)
Hospitals-Injuries Board Overlap			
	Big Award(%)	None or Small(%)	Total(%)
MAIS3+	113 (22.6)	395 (25.4)	508 (24.7)
MAIS<3	276 (55.3)	793 (51.1)	1,069 (52.0)
MAIS NS	110 (22.0)	367 (23.6)	477 (23.2)
Total	499 (100)	1,555 (100)	2,054 (100)

Source: Combined Injuries dataset.

(1) The severity threshold is defined separately for each dataset.

Table 8. Bad Injuries for Combined Injuries Dataset 2010-2011

Dataset Appearances		Bad Injury		Total
		No	Yes	
One Dataset	Number	38,690	3,249	41,939
	%	92.3	7.7	100.0
Two Datasets	Number	4,510	1,253	5,763
	%	78.3	21.7	100.0
Three Datasets	Number	605	634	1,239
	%	48.8	51.2	100.0
Total	Number	43,805	5,136	48,941
	%	89.5	10.5	100.0

Source: Combined Injuries dataset

Table 9. L-P Estimates Compared with Combined Injuries Dataset

Category	Police	Hospitals	Overlap	L-P	Combined	Ratio
2010	8,053	5,011	1,519	26,566	24,748	1.06
2011	7,078	4,548	1,332	24,167	23,983	1.01
Male	8,397	6,329	1,867	28,465	26,736	1.06
Female	6,454	3,230	994	20,972	21,715	0.97
0-14	1,362	2,134	405	7,177	5,612	1.28
15-24	3,832	2,123	698	11,655	10,125	1.15
25-34	3,516	1,689	490	12,119	11,760	1.03
35-44	2,308	1,127	405	6,423	8,379	0.77
45-54	1,614	955	320	4,817	5,630	0.86
55-64	1,113	683	237	3,208	3,472	0.92
65+	1,094	848	306	3,032	2,536	1.20
Bad Injury	2,594	3,533	1,461	6,273	8,029	0.78
Dublin	3,299	1,671	476	11,581	11,979	0.97
Cork	1,611	1,080	383	4,543	5,242	0.87
Galway	924	500	149	3,101	2,879	1.08
Donegal	663	440	171	1,706	1,656	1.03
Kildare	618	479	124	2,387	1,909	1.25
Limerick	794	377	124	2,414	2,971	0.81
Wicklow	500	244	47	2,596	1,231	2.11
Wexford	477	462	139	1,585	1,268	1.25
Others	6,245	4,306	1,239	21,704	19,778	1.10
Total	15,131	9,559	2,852	50,714	48,731	1.04

Sources: police hospitals and Injuries Board datasets; linkage results.

Table 10. L-P Estimates of Total Injuries from Pairs of Datasets

Data Source1	Cat.	Number	Data Source2	Cat.	Number	Overlap	Total	L-P Estimate
Police	All	15,131	Hospitals	All	9,559	2,822	21,868	51,467
Police	All	15,131	IB	All	32,492	4,773	42,850	103,010
Hospital	All	9,559	IB	All	32,492	2,061	39,990	150,699
Police	Bad ⁽¹⁾	2,410	Hospital	Bad	3,235	1,274	4,371	6,119
Police	Bad	2,410	IB	Bad	5,113	1,427	6,096	8,635
Hospital	Bad	3,235	IB	Bad	5,113	890	7,458	18,584

Sources: police hospitals and Injuries Board datasets; linkage results.

(1) Bad Injury is the composite serious Injury defined at Section 4.2

Figures for Paper

Fig.1 Injuries per Fatality, Selected Countries 2010

Source: IRTAD, 2011.

Fig.2 Probabilities of Matches Police-Hospital Linkage 2005-2011

Fig. 3 Linkage Results with Overlaps

***Biographical sketch of all authors**

Jack Short has been an expert on transport policy internationally for over 25 years. He is currently conducting research on road safety at Trinity College Dublin. Formerly he was the Secretary General of the International Transport Forum, an OECD affiliated intergovernmental organisation in Paris. In Ireland he previously worked in the Central Statistics Office, the Departments of Transport and Finance and the Economic and Social Research Institute.

Dr Brian Caulfield is an Assistant Professor in the Department of Civil, Structural and Environmental Engineering, Trinity College Dublin. He has been active in transport research for over 10 years and currently leads a number of EPA and EU FP7 projects. He has published over 40 international peer reviewed journal papers and has presented at over 50 national and international conferences. Dr Caulfield is also the current chair of the Irish Transport Research Network.