Journal of Safety Research 58 (2016) 1–14

Contents lists available at ScienceDirect

Journal of Safety Research

journal homepage: www.elsevier.com/locate/jsr

Record linkage for road traffic injuries in Ireland using police hospital and injury claims data

Jack Short, Brian Caulfield *

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

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form 20 April 2016
Accepted 17 May 2016
Available online 27 May 2016

Keywords:
Road safety
Accident records
Data linkage

ABSTRACT

Introduction: 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. *Method:* In this study, anonymized 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. *Results:* The linkage of police and hospital datasets shows results that are similar 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. *Practical application:* The study also underlines the risk in relying on the Lincoln–Petersen capture–recapture estimator to provide an estimate of the total population concerned. *Conclusion:* 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. First, 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). Second, 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). Third, 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 Fig. 1). Fourth, 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.

Defining and accurately counting road traffic injuries are well known problems (Cryer & Langley, 2006; Fingerhut, 2004; Haagsma et al., 2012; Langley & Brenner, 2004). Difficulties with police injury assessments are also amply documented and the biases and understatement well evidenced (Alsop & Langley, 2001; Department for Transport UK, 2006; Jeffrey et al., 2009; Lopez, Rosman, Jelinek, Wilkes, & Sprivulis, 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

* Corresponding author. Tel.: +353 1 896 2534.
E-mail address: brian.caulfield@tcd.ie (B. Caulfield).

http://dx.doi.org/10.1016/j.jsr.2016.05.002
0022-4375/

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	4,090	302,735	74	−42	−18
Japan	5,745	895,326	156	−45	−22
Korea	5,505	352,458	64	−46	−22
New Zealand	375	14,031	37	−19	+39
Norway	208	8,924	43	−39	−25
Spain	2,478	11,6503	47	−57	−16
Sweden	266	23,307	88	−55	+5
Switzerland	327	24,237	74	−45	−17
United Kingdom	1,905	215,700	113	−47	−35
United States	32,885	223,9074	68	−22	−25

Source: IRTAD, (2012).

to determine thresholds for deciding on matches and non-matches. A method is suggested that 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, summarizes 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

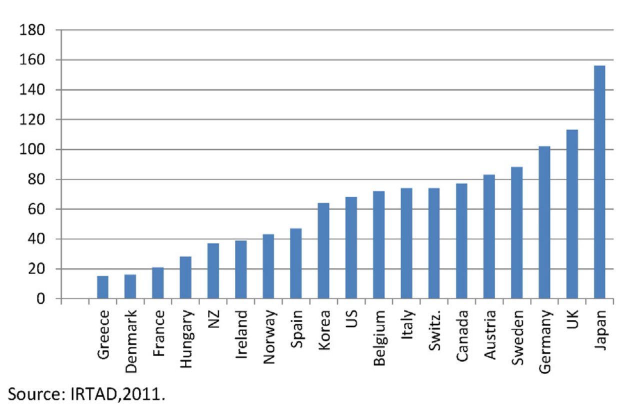

Fig. 1. Injuries per fatality, selected countries 2010.
Source: IRTAD, 2012.

countries where the technique has been used in road safety. Papers from France (Amoros, Martin, & Laumon, 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, Kennedy, Axford, and P (1959) were given a solid mathematical foundation by Fellegi 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 ϒ is formed for each pair of records, one from each set. The comparison vector consists of 1 s for a link and 0 s 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. Fellegi 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 leads 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 program 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, judgments 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