



Police and hospital data linkage for traffic injury surveillance: A systematic review

Ali Soltani^{a,e,*}, James Edward Harrison^a, Courtney Ryder^{a,b,g}, Joanne Flavel^{a,c}, Angela Watson^{d,f}

^a Injury Studies, FHMRI, Bedford Park, Flinders University, SA 5042, Australia

^b George Institute for Global Health, Newtown, NSW 2042, Australia

^c Stretton Institute, University of Adelaide, SA 5005, Australia

^d The Australian Centre for Health Services Innovation (AusHSI), Queensland University of Technology, Qld 4000, Australia

^e Urban Planning Department, Shiraz University, Shiraz, Iran

^f School of Public Health & Social Work, Queensland University of Technology, Qld 4000, Australia

^g School of Population Health, UNSW, Kensington, NSW 2052, Australia

ARTICLE INFO

Keywords:

Data linkage

Injury

Traffic crash

Systematic review

PRISMA

ABSTRACT

This systematic review examines studies of traffic injury that involved linkage of police crash data and hospital data and were published from 1994 to 2023 worldwide in English. Inclusion and exclusion criteria were the basis for selecting papers from PubMed, Web of Science, and Scopus, and for identifying additional relevant papers using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and supplementary snowballing ($n = 60$). The selected papers were reviewed in terms of research objectives, data items and sample size included, temporal and spatial coverage, linkage methods and software tools, as well as linkage rates and most significant findings. Many studies found that the number of clinically significant road injury cases was much higher according to hospital data than crash data. Under-estimation of cases in crash data differs by road user type, pedestrian cases commonly being highly under-counted. A limited number of the papers were from low- and middle-income countries. The papers reviewed lack consistency in what was reported and how, which limited comparability.

1. Introduction

1.1. Background and context of data linkage of police data and health data

Road crashes, particularly those involving seriously or fatally injured people, are often included in administrative data collection. Police and other agencies in charge of road safety manage data systems that focus on crashes, including where, when, and how they happen, as well as the vehicles and people involved. Health services, people, and causes are the focus of hospital and health agency data systems. Injuries from road crashes are included in hospital data. The ICD (International Classification of Diseases) classifies injuries by “external causes”.

Data linkage is the process of connecting data from several sources to explain a phenomenon and is commonly used in epidemiological studies

to obtain an overall appreciation of factors impacting on populations (Bohensky et al., 2010). Data linkage can add to the understanding of complicated events, enabling informed decision-making, and valuable insights in public health research (Baker et al., 2023). It can reduce fragmentation of information and open possibilities for discovery, which may enable more efficient resource allocation and better policymaking. Linkage studies can often achieve this without requiring the collection of new data, which is often time-consuming and expensive. Linkage studies have their own challenges, including requirements to obtain ethics committee and custodian approvals and to adhere to rules to protect privacy and confidentiality. In the case of road crashes, it is the process of combining crash data, hospital admissions, and other relevant sources to estimate crash injuries and better understand crash causes and effects and uncover trends and patterns that may influence road injury prevention policies and activities (Kamaluddin et al., 2019; Amorim et al.,

* Corresponding author.

E-mail addresses: ali.soltani@flinders.edu.au (A. Soltani), james.harrison@flinders.edu.au (J. Edward Harrison), Courtney.ryder@flinders.edu.au (C. Ryder), joanne.flavel@adelaide.edu.au (J. Flavel), angela.watson@qut.edu.au (A. Watson).

<https://doi.org/10.1016/j.aap.2023.107426>

Received 14 August 2023; Received in revised form 7 December 2023; Accepted 12 December 2023

Available online 5 January 2024

0001-4575/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

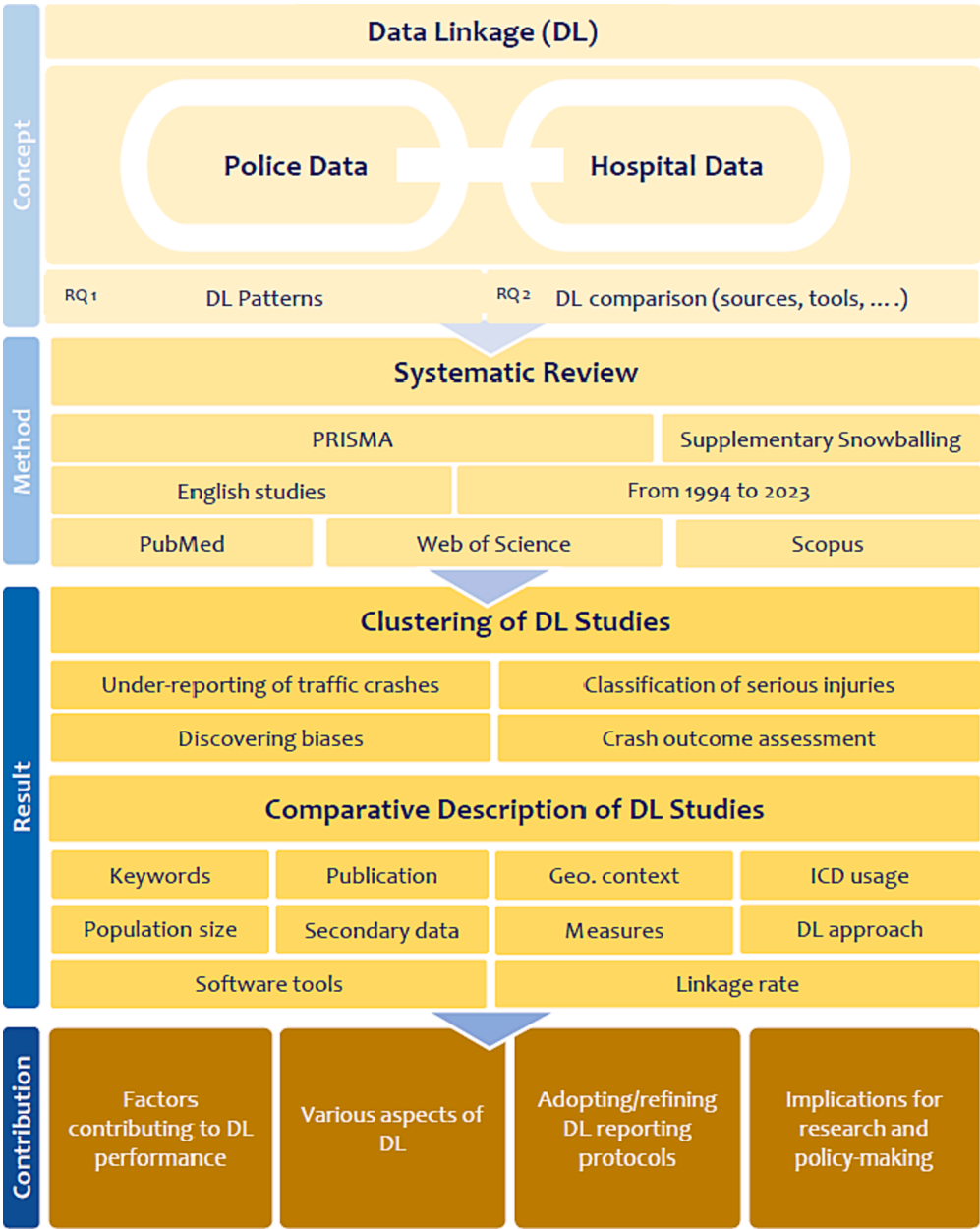


Fig. 1. Research process.

2014). Crash data includes vehicle, road user, and crash details. Hospital data, however, focuses on crash-related injuries and health care usage (Harrison et al., 2019). Linked crash and hospital data may indicate trends in road traffic crashes and injuries and high-risk categories and regions (Paixão et al., 2015a; Paixão et al., 2015b). Data linkage may be utilised to assess road safety, risk variables, and injury severity (Lujic et al., 2008). Data linkage may raise new questions in crash literature (Singleton et al., 2004).

According to Mandacaru et al. (2017), police crash data may estimate crash severity with bias. Evidence suggests data linkage has led to a better knowledge of crash causes such as vehicle speed, driver behaviour, and road conditions, as well as contributing variables like alcohol or drug usage (Hossain et al., 2023). Crash misclassification and database errors make linking crash and traffic data problematic for road safety studies (Imprialou and Quddus, 2019). Police’s “hospitalisation” road category may not adequately represent hospitalisations because of the large variety of severity levels among hospitalised cases (Watson et al., 2013).

Understanding road safety and its effects on public health requires linking police crash data with hospital crash admissions data. The absence of a linkage between these datasets makes it harder to assess injury trends, identify high-risk areas, and devise effective solutions. The linkage of these datasets is not only about better data; it is about generating evidence-based policies to enhance road crash safety and reduce the human and economic toll of road crashes.

Linking crash data with hospital data may also help better evaluate road safety measures, such as the impact of new traffic rules or infrastructure enhancements on collisions and injuries (Urie et al., 2016). Data linkage may help in identifying high-risk road traffic crash zones and individuals, enabling targeted efforts to prevent collisions and injuries (Amoros et al., 2006). Underreporting crashes and injuries are common when using crash data alone, and may cause biases that can be corrected when using linked crash and hospital data (Watson et al., 2015). Linkage of these types of data sets provides further depth and knowledge for researchers and policymakers surrounding traffic crash causes and injuries, additionally it improves data quality and

Table 1
Components of search formula.

Item	Sub-item	Details
Keywords	Main keywords Supplementary keywords	data linkage; police; hospital data; crash; injury casualty; fatality; road; car crash; transport crash; traffic; accident; motor-vehicle; collision; mortality; report; hospital-based; ambulance; emergency; death; registry; health care; linking; linked; record; integrated; integration; underreport; underestimate; misclassification
Operators		“OR”, “AND”
Search formula		((“crash injur*” OR “road crash” OR “road injur*” OR “road fatal*” OR “traffic crash*” OR “Car crash” OR “transport crash” OR “traffic accident*” OR “traffic injur*” OR “car accident” OR “transport accident” OR “traffic casualty” OR “traffic injur*” OR “motor-vehicle injur*” OR “motor-vehicle injur*” OR “traffic collision*” OR “car collision” OR “transport fatal*” OR “road fatal*” OR “traffic mortality” OR “police” OR “police report*” OR “police-report*”) AND (hospital OR “hospital-based” OR ambulance OR emergency OR “death registr*” OR “health care”) AND (linkage OR “linking” OR “Linked hospital” OR “linked ambulance” OR “data linkage” OR “record linkage” OR “linked data” OR “integrated data” OR “data integration” OR “underreport*” OR “under-report*” OR “underreport*” OR “underestimat*” OR “under-estim*” OR “misclassification”))
Time period		1994–2023
Language		English
Document type (included)		Journal paper
Document type (excluded)		Grey literature, conference papers, non-peer reviewed papers, books, book chapters, governmental or consultancy reports, research reports, working papers, theses, and dissertations.
Inclusion criteria		Those papers that align with the scope of the research were considered. Those papers with clear explanation of data linkage process and its components were selected. The papers had to use linked data from at least two datasets of which one of the datasets had to be routinely collected police-reported road crash data. The papers had a full-text versions. The search was limited to the English language and human subjects. There were no restrictions imposed on the nation, kind of road crash, age, race, or gender of crash casualties.
Exclusion criteria		Studies which linked police crash data with non-health sector data were excluded. The studies were excluded if they were case reports, case series, systematic, meta-analysis or narrative review of data linkage. The studies with inadequate data on linkage process and its detailed features were excluded. Review articles of any kind were not included. Papers not written in English were excluded.
Date of search		Tuesday, May 23, 2023

completeness. By linking data from many sources, researchers can correct errors, and fill gaps.

Data linkage studies have added to resources and evidence in several disciplines. One study used data linkage to discover disparities in injury severity among populations by examining variables affecting the gap between police-reported crash data and trauma registry (Hosseinzadeh et al., 2022). Connecting these databases allows researchers to study crash characteristics, variables affecting injury severity, and crash type-specific disparities. In another study (Bohensky et al., 2010), data linkage was used to investigate health disparities and identify populations at high risk for specific health outcomes. By linking disparate health databases, researchers may analyse injury severity disparities across age, gender, urban/rural, and other populations (Moore et al., 2019). Targeting at-risk populations and reducing disparities requires this knowledge.

Furthermore, data linkage enables the tracking of health outcomes over time as well as the evaluation of the effectiveness of public health programmes (Bohensky et al., 2010). This long-term perspective is crucial for understanding intervention impacts, identifying areas for development, and strengthening public health programmes. To ensure data integrity and completeness, researchers have focused on developing trustworthy linkage approaches (Soltani et al., 2022). These initiatives aim to increase the quality of linked datasets as well as the dependability of data linkage study outcomes. Continuous research and advancements in data linkage methods provide valuable insights into health disparities, injury severity, and intervention effectiveness (Alzaffin et al., 2023; Amorim et al., 2014).

Data linkage compares police-reported crash data with trauma data to identify disparities in injury severity among populations (Alzaffin et al., 2023; Soltani et al., 2022). It identifies injury severity variations by location, age, gender, and other characteristics (Moore et al., 2019). Data linkage validates and contrasts injury severity by cross-checking police data with medical diagnosis (Hosseinzadeh et al., 2022). It has extensive uses in public health research, including the study of injury severity, the evaluation of therapies, and the identification of outcome

disparities (Amorim et al., 2014). Data linkage follows health outcomes across time and evaluates intervention effectiveness (Bohensky et al., 2010).

1.2. Research objective/questions

This study compares and explores previously published journal articles on linkage between police crash data and hospital collected crash data in terms of data sources, methods, software tools, and study context; identifies injury patterns among motor vehicle crash casualties; and determines how demographic, geographic, and temporal factors affect linkage likelihood. The research objective is to compare studies linking police crash data with hospital collected crash data, examining data sources, methodologies, software tools, and contextual factors, while investigating injury trends and the impact of demographic, geographic, and temporal variables on this linkage, as well as identifying variables affecting linkage success for enhanced injury monitoring and prevention. This literature review focusses on two main questions:

- What are the dominant patterns of data linkage studies?
- How do data linkage studies compare in terms of data sources, methodology, software tools, research setting, and time span?

1.3. Research contribution

Through systematic analysis, the study aims to explore several aspects of data linkage, such as methodological considerations, data quality and shortcomings, linkage performance norms and standards for reporting studies, and implications for research and policy. Ultimately, the findings will contribute to the advancement of data linkage studies, improve research outcomes and inform policy decisions.

1.4. Research process

Firstly, the concept development stage involved reviewing

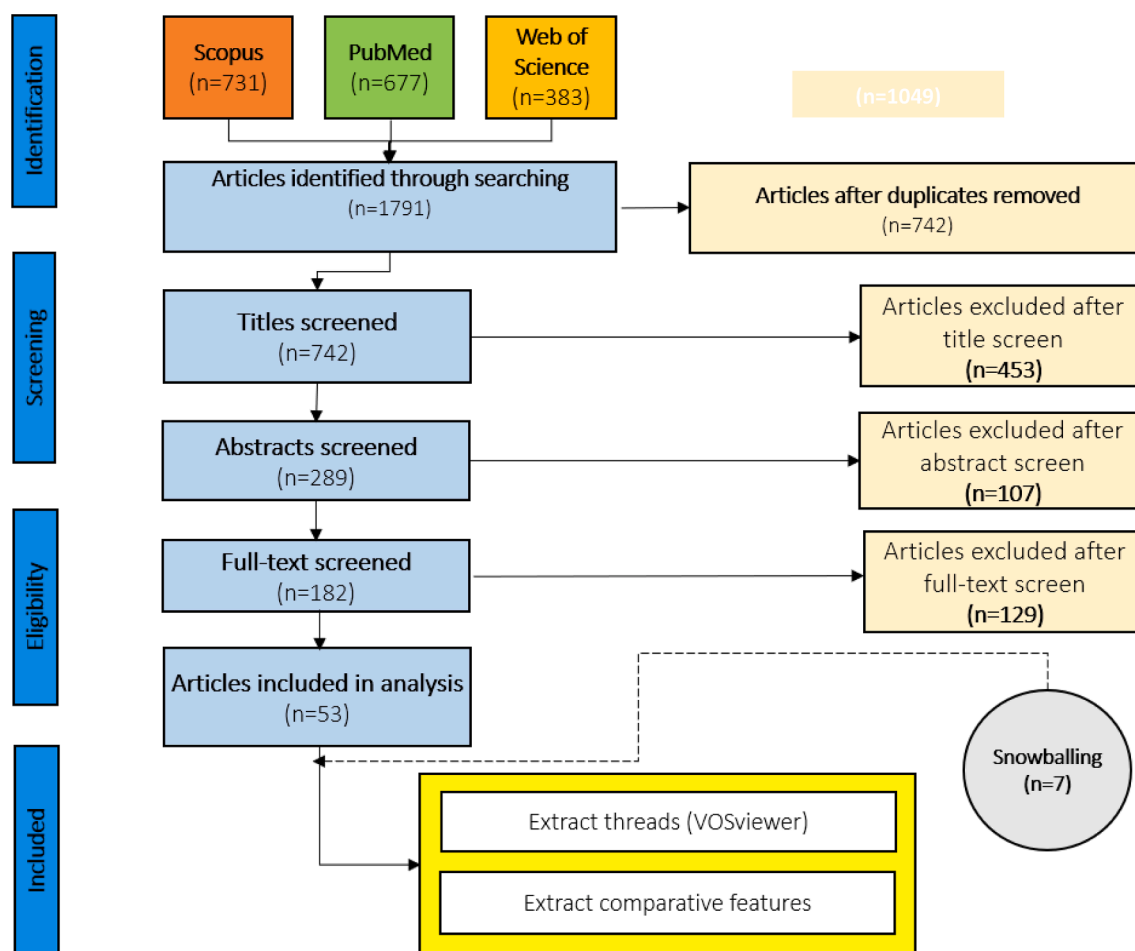


Fig. 2. PRISMA study selection diagram.

background studies and finalizing the two research questions. The existing literature was carefully reviewed in order to identify gaps, and shape the research questions accordingly. Next, the method stage focused on systematic data collection and analysis. The systematic review was conducted using the PRISMA framework and employed snowballing techniques as supplementary. A list of relevant papers on data linkage from well-known databases, including the Web of Science (WOS), Scopus, and PubMed was compiled. The data linkage between police crash data and hospital registry data studies were clustered to identify common themes, then a comparative description to identify trends and gaps was conducted. Finally, the contribution stage highlighted the impact of the research by providing recommendations for standard protocols in data linkage studies. The factors affecting the performance and implications of the findings for research, policy, and healthcare services were identified. The details are shown in Fig. 1.

2. Methodology

2.1. Search process

This paper applied the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology (Page et al., 2021). PRISMA offers a transparent and standardised method for researchers to guarantee the completeness and quality of their systematic review (Salameh et al., 2020). PRISMA improves the reliability and repeatability of systematic reviews in scientific research (Johnson & Hennessy, 2019). It usually entails formulating research questions, doing a thorough literature search, choosing relevant studies based on established inclusion and exclusion criteria, extracting data from the chosen studies,

synthesising and analysing the data, and interpreting the results.

Literature search strategies were developed using keywords related to data linkage between police-reported crash data and hospital data. The selection of keywords was based on the research team's background knowledge and experience. The search was done for studies published between 1st of January 1994, and 23rd of May 2023. Only studies related to data linkage of road traffic injuries were included. The search is detailed in Table 1. The search across all three databases yielded 1791 results, with 742 remaining after removal of duplicates. The titles, keywords and abstracts of these records were evaluated against the eligibility criteria (which resulted in the removal of 453 and 107 papers, respectively), with 182 papers remaining. Based on the full-text assessment of these papers, 129 were eliminated since they did not satisfy the inclusion criteria. The supplementary snowballing as described later resulted in the addition of seven more papers, bringing the total number of papers in the review and synthesis to 60. The procedure followed the PRISMA 2020 statement and its guidelines (Page et al., 2021).

The initial search results from three databases were saved in an Excel document. To detect and delete duplicates, each paper's title, author and year were carefully copied into the "Find" function. This made it possible to remove duplicates, assess the titles, keywords and abstracts against the inclusion and exclusion criteria as detailed in Table 1, and generated a shortlist which was potentially ready for full-text screening. The inclusion and exclusion checklist was used again for the full-text screen, while the reasons of exclusion were noted.

The PRISMA flowchart (Fig. 2) depicts the progression of research from identification to screening for eligible inclusion to final synthesis.

Due to the multidisciplinary nature of the topic, a supplementary search was undertaken using the snowballing technique. Formal

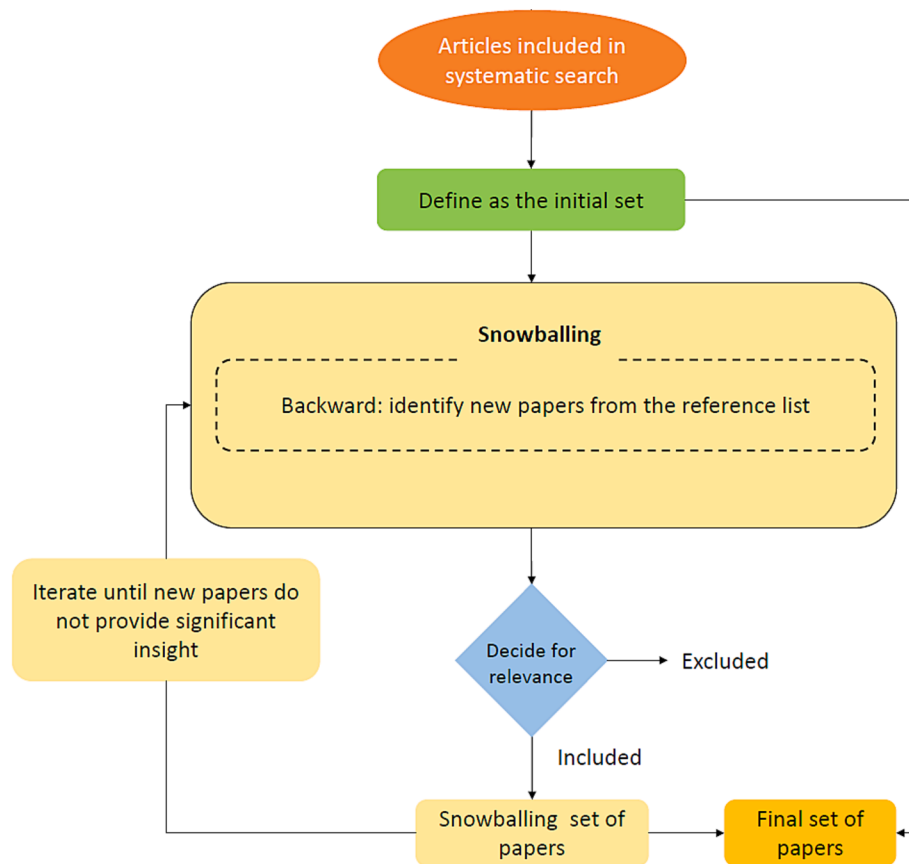


Fig. 3. Supplementary snowball process, adopted from Santoro et al. (2021).

protocol-driven search strategies, may miss key evidence in systematic reviews of complicated and varied evidence – such those assumed for policymaking and management issues (Greenhalgh and Peacock, 2005). Hand searches complement database searches well and should be considered whenever possible (Vassar et al., 2016).

The 53 papers from PRISMA output were taken as the “Start set”, and the process was started with the initial pool of papers. The Backward snowballing was then conducted to identify additional relevant papers. For this purpose, the reference lists of the papers were examined in the start set and excluded papers that did not meet the basic inclusion criteria. Where backward snowballing identified papers had already been examined in the process discussed, those papers were removed from the set of papers based on being found earlier in the search. The remaining papers were considered as potential candidates for inclusion in the review and were further evaluated through abstract and full-text review. This process was repeated until no additional relevant articles were found. In particular, the guidelines suggested by Greenhalgh and Peacock (2005) were followed to ensure that the search methodology was rigorous and systematic (Fig. 3). The full list of nominated papers is provided in Appendix A.

The following features were mined from the selected papers: title; list of authors; year of publication; the objective of linkage; the geographical context; the time coverage; usage of the ICD standard; the used data sources based on the data authority; the size of the datasets (sample or population size); data linkage methods; the software/tool used; the parameters describing the linkage result (e.g., % linked records, match rate; discordance rate); and the statistical methods. These features are then used to compare the studies.

2.2. Clustering analysis

The clustering analysis is based on the content analysis of the selected papers. After downloading the content (title, abstract, keywords) file of all selected papers, they were imported into VOSviewer (Van Eck & Waltman, 2018) and the software made some outputs including clustering map. The repeated key terms in the whole set of papers are grouped into clusters. Clusters are non-overlapping in VOSviewer. Also, clusters do not need to exhaustively cover all key terms in a map. Hence, there may be key terms that do not belong to any cluster. The clusters are named by the authors based on the nature of key terms grouped in the cluster. The key terms are shown in the network visualisation by their label and, by default, by a circle. The weight of a key term determines both the label's size and its circle. A key term's colour is defined by the cluster to which it belongs. Links are shown as lines between key terms.

3. Clustering linkage studies

Based on the research objectives and outcomes of the selected papers, these studies can be classified into four groups (Fig. 4).

- Data Linkage for investigating under-reporting of injurious cases in crash data;
- Data Linkage for estimation and classification of serious injuries;
- Discovering biases in reporting and linking data;
- Data Linkage for crash outcome assessment; discovering the influential factors and evaluating safety policies.

Note that some studies may have multiple or overlapping objectives and could be classified under multiple categories.

Thread A. Data Linkage for investigating under-reporting of injurious cases in crash data:

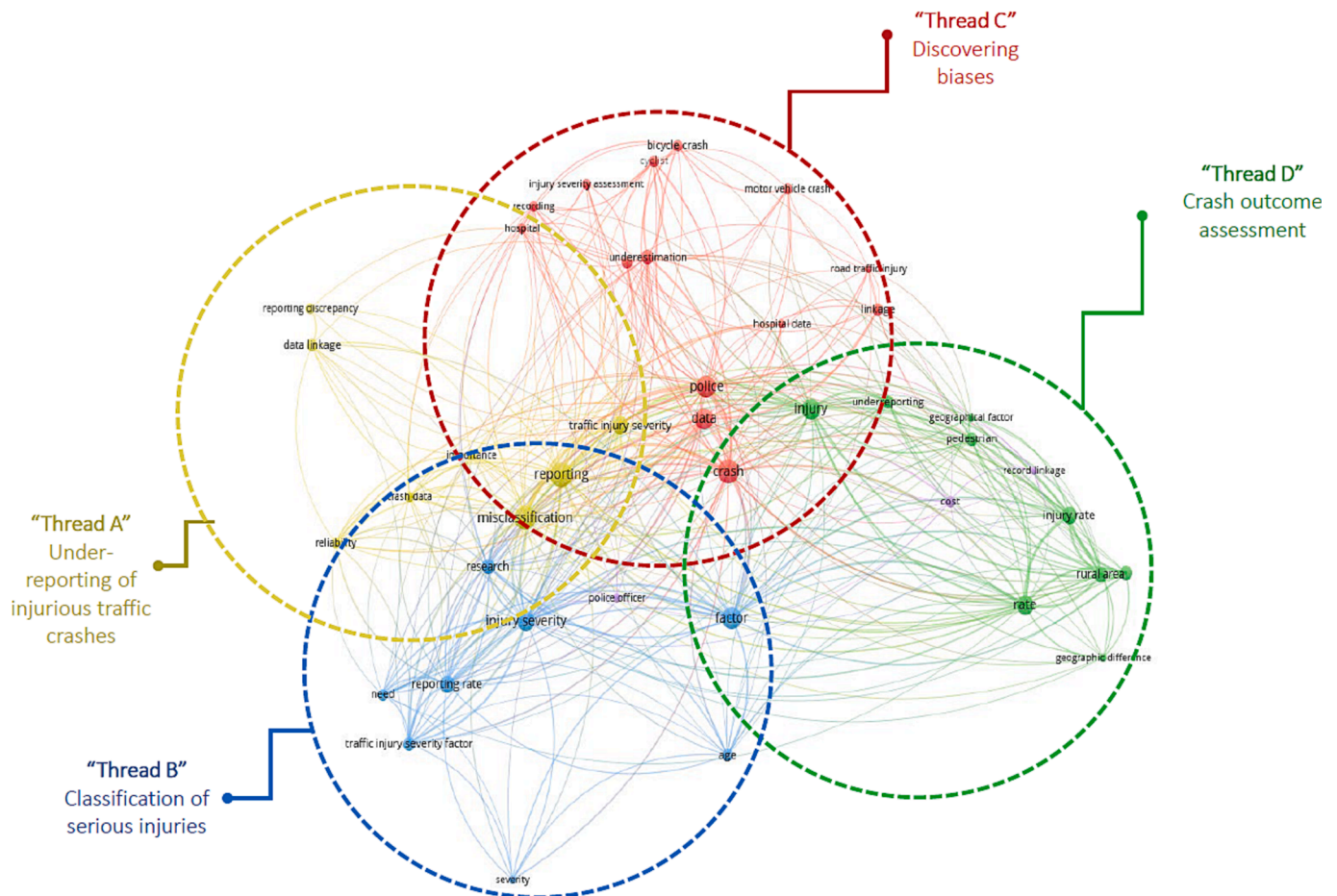


Fig. 4. Clustering of studies.

This type of study focused on the extent of under-counting of injurious road crash incidents in crash data (Watson et al., 2015; Lujic et al., 2008; Rosman and Knuiman, 1994). It uses hospital data as a denominator and investigates how many hospital records can be linked to a road crash record, indicating cases that may not have been reported to the police (Hosseinzadeh, 2022; Milani et al., 2015). The studies in this category have mostly focused on the following themes: the extent of underreporting and linked data analysis; variation in underreporting by road user group; impact of injury severity and characteristics on underreporting; factors contributing to underreporting; policy and resource allocation implications of underreporting and need for improved integration and data reform.

Using linked police crash and hospital data, these studies consistently find higher incidence rates of injuries compared to police data alone, indicating substantial underestimation of injurious road crash incidence by the latter source (Watson et al., 2015; Lujic et al., 2008; Rosman and Knuiman, 1994; Hosseinzadeh, 2022; Milani et al., 2015). Vulnerable road users, including pedal cyclists and motorcyclists, experience lower linkage rates between police and hospital data compared to motor vehicle occupants, suggesting higher levels of underreporting for these groups (Conderino et al., 2017; Soltani et al., 2022). Less severe injuries, such as not requiring emergency room treatment, are more likely to be underreported in police records. Single-vehicle crashes also tend to be underrepresented in police-reported data (Schneider et al., 2023; Wilson et al., 2012).

Age and gender play a role in underreporting, with younger road users, particularly males, being underrepresented in police-reported data. Disparities in reporting based on race have also been observed (Janstrup et al., 2016). The underreporting of road crash injuries has

implications for policy decisions and resource allocation. Policymakers should be cautious when relying solely on police-reported data for road safety evaluations and resource allocation (Ferreira et al., 2015; Amorós et al., 2006). Studies emphasize the need for better integration between police and hospital databases to enhance the reliability of crash data and address the underreporting issue (Ferreira et al., 2015; Jeffrey et al., 2009).

Causes of under-enumeration include a lack of awareness or knowledge about reporting procedures (Wang et al., 2019; Dandona et al., 2008; Rosman & Kuniman, 1998), fear of legal consequences (Schneider et al., 2023; Rosman & Kuniman, 1998), and perceived difficulties in the reporting process (Elvik, 2010; Janstrup et al., 2016; Dandona et al., 2008). These factors can lead to inaccurate understanding of crash patterns (Weijermars et al., 2016; Rosman & Kuniman, 1998) and inadequate resource allocation for road safety initiatives (Schneider et al., 2023). To counter these issues, strategies involve implementing public awareness campaigns to educate the public about reporting procedures (Weijermars et al., 2016; Alsop & Longley, 2001), streamlining reporting processes to reduce perceived difficulty (Elvik, 2010; Janstrup et al., 2016; Dandona et al., 2008), and encouraging anonymous reporting to alleviate concerns about legal consequences (Schneider et al., 2023; Loo & Wong, 2012). These efforts aim to enhance reporting accuracy and reliability, thereby improving road crash data collection and analysis.

Thread B. Data Linkage for estimation and classification of serious injuries: Similar to the previous type, the studies in this cluster use methods that involve linking road crash and hospital records (Lombardi et al., 2022). This cluster can be split into two different types 1) misclassification of serious injuries; and 2) estimation of serious

injuries. The major difference is that the misclassification studies use police-reported crashes as the denominator, but the estimation studies do not. The most frequent themes discussed in this cluster include extent of underreporting and linked data analysis; importance of data integration and quality; discrepancies in injury severity assessment; accuracy of police-reported injury severity; estimation of serious injuries and the misclassification of injuries.

Thread B1: Misclassification of serious injuries:

These studies aimed to identify and correct classification or reclassify injury severity (serious injury and sometimes death), by using police data as the denominator and linking to hospital records. Some studies in this cluster investigate the discrepancy between injury severity classification by different sources (police, hospital), highlighting the subjective nature of injury severity classification in police reports (Lombardi et al., 2022; Janstrup et al., 2016; Kudryavtsev et al., 2013). They emphasize the importance of using multiple data sources and considering clinical assessments for accurate injury severity classification (Lombardi et al., 2022; Lopez et al., 2000).

Lopez et al. (2000) studied how trauma registry and police records may enhance severity categorisation and police reports in Western Australia. The research emphasises data integration and several injury monitoring data sources for underrepresented communities. Tsui et al. (2009) compared injury severity score (ISS) and police reports of traffic crash injury severity and hospital stay using binary logistic regression. This study found that police records exaggerated injury severity and exceeded the 12-hour hospital stay outcomes. Ferreria et al. (2015) used linkage to compare case severity according to police with severity according to hospital data, finding that police data often overstated severity.

Research reveals discrepancies between injury severity assessments by different sources, emphasizing the need for multiple data sources and considering clinical assessments (Lombardi et al., 2022; Janstrup et al., 2016; Kudryavtsev et al., 2013; Lopez et al., 2000). Various studies use methods involving linking police crash and hospital records to correct misclassification, or reclassify injury severity (Lombardi et al., 2022; Weijermars et al., 2016; Tsui et al., 2009). Kudryavtsev et al. (2013) investigated the differences in injury severity assessment between police, hospital, and in-depth crash investigations and discussed the implications for crash prevention and injury mitigation.

Thread B2: Estimation of serious injuries:

This category of studies in cluster B investigates better estimation of serious injury, commonly by including cases not reported to police as well as those that are (to include those cases not reported to police). These studies use methods that include some that link road crash and hospital records, as well as some that do not (hospital records only). This category may employ capture-recapture methodologies, which help estimate the total number of cases by accounting for the overlap between different data sources (Marrison and Stone, 2000). For instance, Weijermars et al. (2016) examined Dutch road injuries using capture-recapture for estimating unknown populations using linked data and found mopeds had the lowest hospital discharge record completeness. Some studies in this group also examine bias or at least compare how reporting may differ when using linked data versus non-linked data (either hospital data alone or police data alone) (Watson et al., 2015).

Studies finding that police-reported injury severity can be inaccurate and may overestimate or underestimate severity (Watson et al., 2015; Lopez et al., 2000). Studies using linked police crash and hospital data have consistently found higher performance of linked data in estimation of severe injuries. Linkage rates were observed to be lower for pedal cyclists and motorcyclists (Lujic et al., 2008; Lombardi et al., 2022; Janstrup et al., 2016; Kudryavtsev et al., 2013; Watson et al., 2015). Studies highlight the significance of integrating different data sources and improving data quality to accurately predict road traffic serious injuries (Paixão et al., 2015a; Reurings et al., 2011).

The underreporting of road crash injuries, particularly severe ones, has implications for data quality and reliability (Watson et al., 2015;

Lombardi et al., 2022). Amorim et al. (2014) explored integrating crash data with emergency medical services (EMS) monitoring records to improve serious injury estimations. A study linking police and hospital records found that crash-injury severities modified when using linked data vs police-reported data (Alzaffin et al., 2023). Weijermars et al. (2016) examined Dutch road injuries using capture-recapture method for estimating unknown populations and found mopeds had the lowest hospital discharge record completeness. The number of major injuries is determined by combining police data with hospital discharge record data. Following the linking of both datasets, an estimated approach influenced by capture-recapture was employed to estimate the number of major traffic injuries.

The challenges related to the misclassification of serious injuries often arise from the use of severity scales (e.g., Abbreviated Injury Scale) to classify injuries (Sciortino et al., 2005; Lopez et al., 2000; Couto et al., 2016). Variability in injury severity classifications, can lead to inconsistencies (Short and Caulfield, 2016; Tainter et al., 2020; Soltani et al., 2022). The many types of injuries, such as fractures or internal injuries, contributes to the complexity (Lopez et al., 2000; Short and Caulfield, 2016). Functional impacts on injured people, like long-term disabilities, further complicate the accurate assessment of injury severity (Lopez et al., 2000; Amoros et al., 2006; Short and Caulfield, 2016; Dandona et al., 2008). These complexities often result in incomplete or inaccurate medical data, affecting the estimation process (Conderino et al., 2017; Lopez et al., 2000; Dandona et al., 2008). To address these challenges, researchers suggest enhanced data collection from medical sources (Mock et al., 2005; Schneider et al., 2023; Amoros et al., 2006), collaboration with healthcare providers (Sasser et al., 2012; Short and Caulfield, 2016; Abdulatif, 2017), and utilization of predictive modelling techniques (Tainter et al., 2020; Wilson et al., 2012). These strategies aim to improve the accuracy of estimating and classifying serious injuries, thereby enhancing the quality of road crash data analysis.

Thread C. Discovering biases in reporting and linking data:

Studies in this category examine various biases related to reporting and linking data in road crash incidents (Hosseinzadeh et al., 2022; Wilson et al., 2012; Lujic et al., 2008; Cryer et al., 2001). This type can be an aspect of studies in thread (A) or (B). For example, researchers may investigate biases in misclassifying serious injuries or biases in under-reporting. Some studies in Thread B2 also examine biases, as described before. The most common themes in this category include biases and reporting in linked data studies; selection bias and linked data utilization; challenges in data linkage and bias mitigation; comparison of linked data vs. non-linked data; bias and underreporting in linked data; variation in injury reporting; and factors influencing biases in reporting. Some examples and findings are listed below:

Factors such as injury severity, speed, and pedestrians were found to influence both crash data records and trauma registry (Janstrup et al., 2016). Suboptimal or inadequate data links may limit the value of data linkage, resulting in systematic bias (Bohensky et al., 2010). Disparities in variables like gender, age, and road user type have been identified between crash and hospital collected datasets (Cryer et al., 2001; Schneider et al., 2023). Common characteristics found across studies on selectivity bias in linking crash data to public health data included gender and indicators of injury severity including speed and crash type (Hosseinzadeh et al., 2022).

Understanding divergency in the reporting rate of road crashes is crucial for devising effective policy actions to increase reporting rates and reduce bias in road safety analysis (Soltani et al., 2022). Due to the separation of linkage and analysis methods, researchers using linked data have trouble assessing bias (Harron et al., 2017). The crash dataset is also susceptible to selection bias due to limited sample and data numbers for each condition (Morgan et al., 2020). Tarko and Azam (2011) explored Indiana pedestrian crash-related selection bias in a linked crash and hospital data set when predicting low injury levels in pedestrian-involved crashes, with age, gender, crash type, and road

layout configuration influencing record presence. A person with a more severe injury is more likely to be sent to the hospital and associated with hospital admission data than other people with less severe injuries, which might lead to selection bias in the linked data (Tarko and Azam, 2011; Giles, 2001).

Underreporting, misclassification, poor data quality, and slow data exchange may introduce bias in police data (Brubacher et al., 2019). Additionally, studies have shown that severe injuries are twice as likely to be reported by clinicians compared to police-reported data (Short and Caulfield, 2016). Moore (1997) conducted crash-hospital data linkage in Alaska, revealing significant biases based on geographical location and crash type, but not age and gender. Cryer et al. (2001) observed significant disparities in gender, age, and type of road user in crash and hospital collected datasets.

Schneider et al. (2023) found differences between police-reported crashes and EMS data, indicating the importance of integrating crash data with emergency medical services monitoring records (Schneider et al., 2023; Amorim et al., 2014). Lujic et al. (2008) analysed the reporting rates of road traffic injuries to the police and hospital databases and identified factors influencing injury classification such as age, collision type, weather conditions, and individual characteristics. Sciortino et al. (2005) mapped and assessed fatal police severe injury data, revealing disparities in reporting based on race and underreporting of pedestrian injuries. Various factors, such as age, gender, crash type, and geographical location, were found to influence injury reporting rates (Hosseinzadeh et al., 2022; Moore, 1997; Wang et al., 2019).

The biases in data collection and reporting within road crash analysis can arise from under-reporting certain demographics, such as age and gender (Amorim et al., 2014; Alzaffin et al., 2023; Jeffrey et al., 2008; Couto et al., 2016), geographic or jurisdictional disparities (Couto et al., 2016; Kudryavtsev et al., 2013; Abdulatif, 2017), reporting bias due to external factors (Abay, 2015), and errors in data entry or transcription (Conderino et al., 2017; Lopez et al., 2000; Short and Caulfield, 2016). These biases can lead to differing reporting standards and protocols (Brubacher et al., 2019; Ferreira et al., 2015; Conderino et al., 2017; Lopez et al., 2000; Soltani et al., 2022) and incomplete or missing data fields (Conderino et al., 2017; Lopez et al., 2000; Amoros et al., 2006). To minimize these discrepancies and biases, strategies involve standardizing data collection practices (Alzaffin et al., 2023; Conderino et al., 2017), regular data quality audits and checks (Brubacher et al., 2019; Kudryavtsev et al., 2013; Abdulatif, 2017), training and guidelines for data collectors (Lopez et al., 2000; Tainter et al., 2020; Alsop & Longley, 2001), and cross-checking with external data sources (Amoros et al., 2006; Wilson et al., 2012). By implementing these approaches, road crash analysts aim to minimize biases and ensure more accurate and consistent data collection and reporting practices.

Thread D. Data linkage for crash outcome assessment; discovering the influential factors and evaluating safety policies

The studies in this cluster highlight the value of using data linkage and integration of data from multiple sources as tools for studying crash outcomes and consequences. Some studies discuss the reliability of different data sources (police, hospital, EMS) and suggest manual examination of heuristic algorithm results for data reliability. Furthermore, some studies in this cluster are interested in modelling injury severity, body regions, or injury nature on the basis of various crash, infrastructure, or driver/rider characteristics. These almost exclusively use all road crash records that link to another health source (for the injury information), but they must have a police record. Some of the common themes investigated in this cluster include value of data linkage and integration; modelling crash outcomes; disparities and rural injury rates; predicting the cost of road traffic injuries and insights for marginalized groups.

Abulatif (2017) discovered that combining crash and hospital data may assist planning, monitoring, and evaluation of road safety initiatives using verified road crash data in five Brazilian cities. Soltani et al. (2022) identified injury trends and informed traffic fatality and injury

prevention measures using linked data. The analysis informs focused initiatives for road safety improvement. Some studies focus on developing predictive models for injury severity based on various factors (Tarko and Azam, 2011; Amoros et al., 2006). Some studies investigated disadvantaged groups and rural injury rates, highlighting factors such as EMS response times, speeding, and trauma system deficiencies (Li et al., 2008; Santiago et al., 2020; Lombardi et al., 2022; Wang et al., 2019; Weijermars, 2016).

This cluster includes some studies on disadvantaged groups and rural injury rates, highlighting factors including prolonged EMS response times, higher speeds, drunk driving, and trauma system deficiencies (Li et al., 2008; Wang et al., 2019; Lateef, 2010). They may also examine spatial inequalities, emphasising greater death rates in rural traffic crashes and the need to discover environmental factors to reduce road traffic injuries (Gonzalez et al., 2006; Mitchel et al., 2015; Li et al., 2008). Santiago et al. (2020) assessed the medical expenditures of personal injuries, nighttime crashes with and without illumination, and weekend vs weekday crashes. Bambach and Mitchel (2015) estimated human recovery costs of badly wounded road crash casualties using linked data.

Studies in this cluster suggested integrating of different data sources to enhance data quality for road crash outcome assessment. This involves incorporating diverse data sets, such as police reports, hospital records, and other relevant sources (Amorim et al., 2014; Mandacaru et al., 2017; Mitchell et al., 2015; Gonzalez et al., 2006; Lujic et al., 2008; Janstrup et al., 2016; Curry et al., 2019; Jeffrey et al., 2008), often aiming to overcome challenges related to differing reporting standards and protocols (Brubacher et al., 2019; Ferreira et al., 2015; Conderino et al., 2017; Lopez et al., 2000; Soltani et al., 2022) and incomplete or inaccurate data (Conderino et al., 2017; Lopez et al., 2000; Amoros et al., 2006). Researchers propose strategies such as utilizing technology like sensors and cameras to enhance data collection (Mitchell et al., 2015; Lopez et al., 2000; Abdulatif, 2017), automating data collection processes (Short and Caulfield, 2016; Alsop & Longley, 2001; Amoros et al., 2006), employing data validation and verification processes (Amorim et al., 2014; Mitchell et al., 2015; Gonzalez et al., 2006; Conderino et al., 2017; Lopez et al., 2000; Couto et al., 2016) and integration of data into existing systems (Kudryavtsev et al. 2013; Soltani et al., 2022). Additionally, strategies include linking data from different agencies or jurisdictions (Amorim et al., 2014; Mandacaru et al., 2017; Mitchell et al., 2015; Gonzalez et al., 2006; Jeffrey et al., 2008) and utilizing data fusion techniques (Watson et al., 2015; Alzaffin et al., 2023) to improve the integration and overall quality of road crash data. These approaches collectively aim to enhance the reliability and comprehensiveness of data used in road crash analysis, leading to more accurate and insightful findings.

4. Comparative description of linkage studies

The analysis conducted in this section consisted of two main components. Firstly, a macroscopic bibliometric analysis was performed to identify emerging topics and examine the temporal and geographical distribution of data linkage studies on a global scale. Secondly, a comparative description of individual studies was provided, focusing on research characteristics such as sample size, methodology, data sources, software tools, and linkage standards.

4.1. Document data: trends, and collaboration

Over the course of nearly three decades, the dataset of 60 papers located by this systematic review has witnessed growth, with an annual growth rate of 6.37 %. This indicates an increase in the number of journal papers being produced within the field of data linkage. On average, each paper in the collection has an age of 9.4 years, suggesting that the dataset represents a mix of recent and relatively older publications. The average citations per document is 22.77, highlighting the

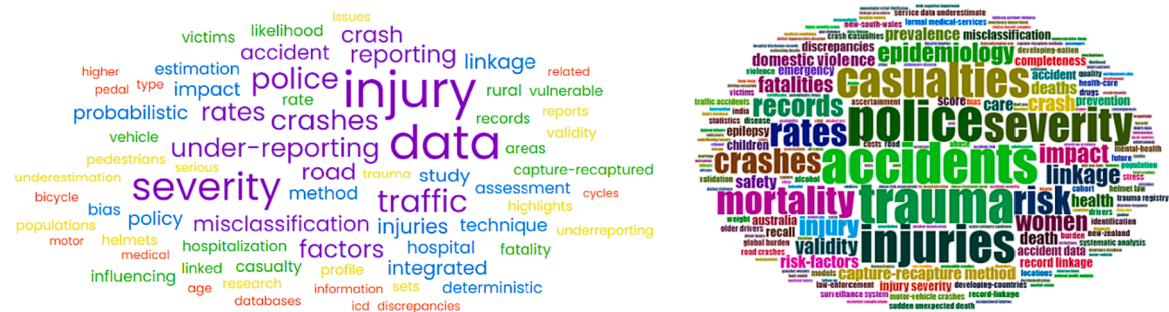


Fig. 5. Word cloud of keywords mentioned in the abstract (left); word cloud of author keyword (right).

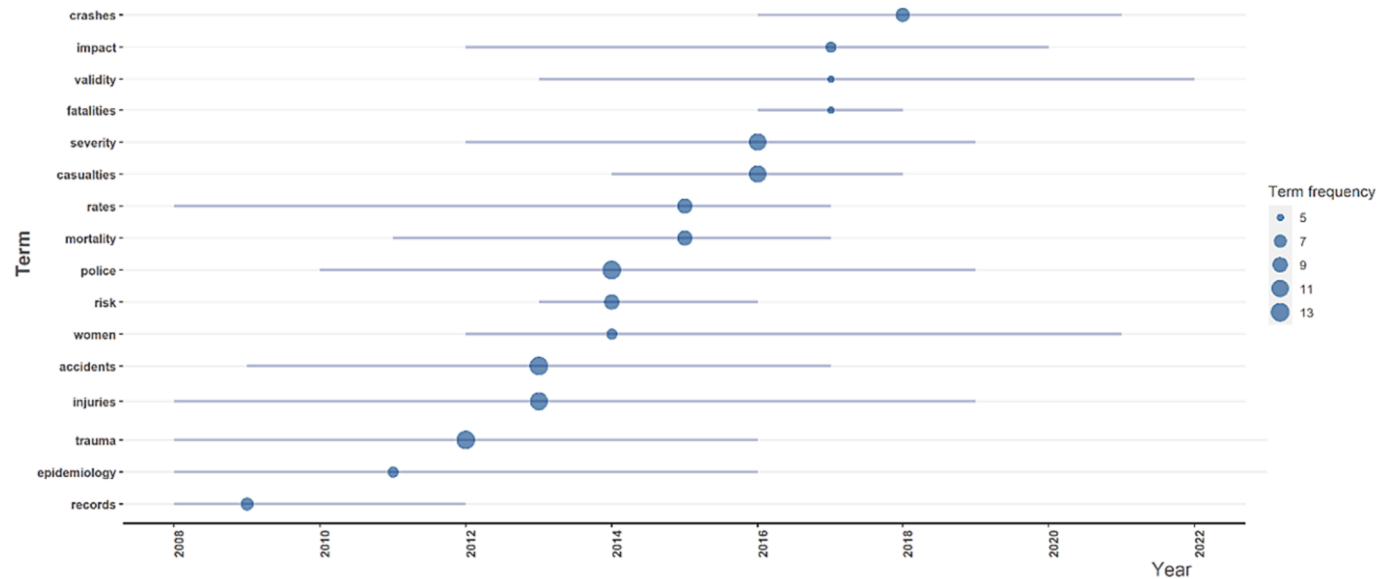


Fig. 6. The temporal trend of the keywords.

level of influence and recognition received by the documents in the collection. This indicates that the works are being cited by other researchers, implying impact on the field. Additionally, the references section of the documents contains a total of 1,715 citations, reflecting the breadth of research and external sources referenced within this dataset. The countries with highest citation frequency of linkage studies are: Australia; USA; France; Sweden, New Zealand, and China.

The dataset involves 222 authors who have contributed to the 60 documents. Among these, three authors have authored single-authored documents, indicating the presence of individual research endeavors. Collaboration is also evident, as each document has an average of 3.7 co-authors. This indicates a high degree of teamwork and cooperation within the field, fostering a culture of collaborative research. International collaboration plays a significant role within the dataset, with approximately 21 % of the co-authorships involving researchers from different countries. This highlights the global nature of the field of study, as researchers from various nations come together to contribute their expertise and perspectives.

4.2. Frequent keywords and trend topics

The dataset consists of documents with various subject matters, as indicated by both Keywords Plus (ID) and Author's Keywords (DE). The presence of 165 Keywords Plus and 184 Author's Keywords suggests the diversity of topics covered in the collection, with authors utilizing specific terms to highlight the focus and themes of their work. The word cloud graph visually represents the relative size of each term, reflecting

its frequency and significance throughout the dataset (Fig. 5). As anticipated, the terms “injury”, “severity”, “data”, “crashes”, “roads”, “police” and “linkage” emerge as the most often occurring words in the abstract of the papers, given they were included in the search query. However, the situation is somewhat different in the content of the papers, and words such as “accidents”, “police”, “severity”, “causalities”, “crashes”, “injuries”, “trauma”, “mortality”, “epidemiology” and “linkage” were the most frequent words. In addition to the aforementioned terms, the literature on data linkage often addressed the concepts of traffic, factors, misclassification, rates, and underreporting. The use of word clouds to display keywords aids in discerning the subject matter and primary emphasis of a certain research endeavour. These terms exhibit a strong interconnection.

Fig. 6 shows the trend of topics in data linkage literature over time, highlighting shifts in research focus within the field. Early studies (before 2012) focused on trauma, epidemiology, and records, investigating the impact of traumatic events, disease patterns, and data integration. From 2012 to 2014, the research then turned to injuries, accidents, women, risk, and police, examining factors contributing to injuries, gender-specific considerations, risk assessment, and law enforcement involvement. Between 2014 and 2018, the studies explored mortality, rates, validity, and impact, analysing mortality trends, calculating rates, assessing data validity, and evaluating the overall impact of data linkage on health outcomes. From 2016 to 2018, the focus narrowed to casualties, severity, and fatalities, assessing the consequences and severity of incidents, such as accidents, and their impact on individuals and communities. Since 2018, researchers have

Table 2
Inconsistency in keywords and terms used in data linkage literature.

Concept	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Crashes	Crashes	Collisions	Road accidents	Traffic accidents	Motor vehicle crashes	Incidents		
Injury severity	Injury severity	Severity of injuries	Crash severity	Traffic injury severity				
Police data	Police data	Police reports	Police-reported information	Police-reported data	Police records	Police reporting		
Data linkage	Data linkage	Linked police-hospital data	Record linkage	Linked dataset	Linked database	Linked data	Police-hospitalisation linked data	
Hospital data	Hospital data	Hospital or hospitalisation records)	Hospital discharge data	Medical records	Prehospital data	Hospital injuries data	Patient data	Hospital admission data
Injury patterns	Injury patterns	Injury locations	Injury sites	Injury types	Injury characteristics	Injury specifics		
Data linkage	Data linkage	Record linkage	Linked data	Data integration	Data fusion			
Under reporting	Under reporting	Reporting bias	Misclassification	Under estimation	Over estimation			
Injuries	Injuries	Non-fatal injuries	Casualties	Non-fatal traffic injuries	Traffic-related injuries			
Data quality	Data quality	Data accuracy	Data verification	Data completeness	Data validity	Data integrity		

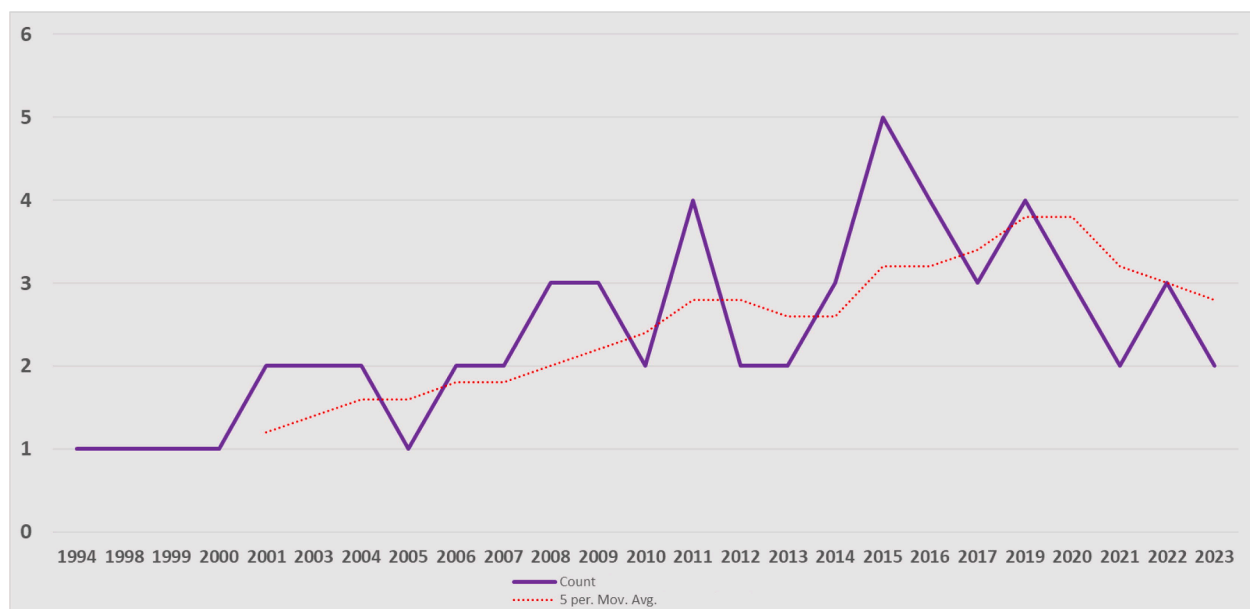


Fig. 7. Temporal distribution of selected studies.

increasingly focused on crashes, investigating causes, consequences, and prevention measures related to collisions in various contexts. This graph reveals a dynamic landscape, with evolving research priorities in data linkage literature. From trauma to specific areas like injuries, mortality rates, casualties, and ultimately crashes, these shifts reflect changing research interests and emerging challenges in the field.

On the other hand, different terms or words are commonly used to refer to the same concepts in this literature. Table 2 shows various concepts related to data linkage, analysis, and terminology used in the context of crash incidents, injury severity, police data, hospital data, road user types, and data quality. These terms frequently refer to the same concept, but they may be used interchangeably in various studies and publications. For example, “crash” can also be referred to as “collision,” “road accident,” “traffic accident,” or “motor vehicle crash.” Similarly, “injury severity” may be described as “severity of injuries,” “crash severity,” or “traffic injury severity.” These alternative terminologies may help researchers, practitioners, and policymakers

understand and communicate effectively within their respective fields of interest and expertise. However, they also carry potential for ambiguity and miscommunication. A detailed version of this Table including the references to each term is provided in Appendix B.

4.3. Studies by year and time-coverage

The selected papers include works published from 1994 (the earliest journal paper in the included set) to 2023 with diverse study themes (Fig. 7). While there was a limited quantity of papers published before the year 2000, there has been an evident upward trend in the number of publications after that time. On average, 2.5 articles are published yearly, and the annual number of papers in this area has tended to increase. The number of linked injury records varied between studies. Rosman and Knuiman (1994), included 18,544 linked injury records, whereas Schneider et al. (2023) had only 375. One of the earliest cases of research was conducted in Western Australia and revealed a high degree

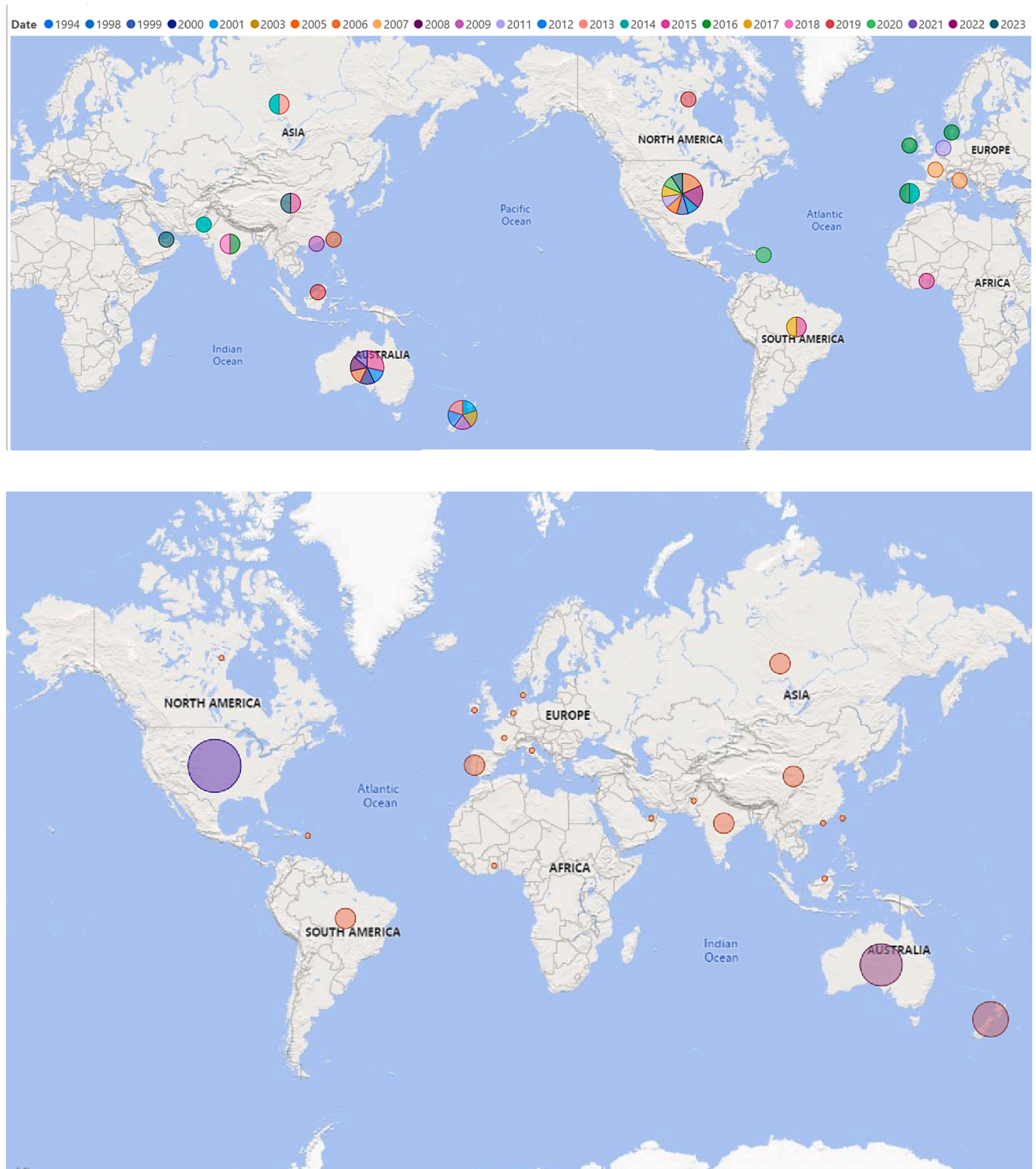


Fig. 8. Geographical distribution of selected studies by date of publication (top); by frequency (down) (generated by MS Power BI, version 2.122.746.0).

of concordance between the two sources linked regarding crash type and road user classification; nevertheless, police-reported injury severity levels were found to be less consistent (Rosman and Knuiman, 1994). Using linked police and emergency department data, a recent study from Milwaukee, United States, was successful in improving the estimation of cyclist injury (Schneider et al., 2023). The highest share of papers belongs to the decade of 2004 to 2014.

The time periods of data examined in the studies range from as early as 1991 to as recent as 2019–2023. Study duration varied, with some being one or two years, while others studied a longer time period. Reurings et al. (2011) with a time period of 15 years, from 1993 to 2008; Weijermars et al. (2016) with a time period of 10 years, from 2000 to 2009; and Lujic et al. (2008) with a time period of 8 years, from 1999 to 2006 are the longest-term studies. These studies reveal longer-term

traffic safety trends.

4.4. Studies by geographical context

The studies ranged in location across 25 nations, mostly in high-income countries ($n = 45$ papers) such as the US, Australia, Canada, and France, some middle-income countries ($n = 11$ papers) such as Brazil, China, Malaysia, and lower middle-income countries ($n = 4$ papers), such as Ghana, India, and Pakistan. This classification is based on the World Bank classification of economies (World Bank, 2023). Since 1992, the US Department of Transportation (DOT) has supported state-level data linkage programmes for standard reporting (Benavente et al., 2006). The US National Center for Injury Prevention and Control monitors injury trends and road safety measures using linked police crash data and hospital records (CDC, 2018). The EU encourages data linkage in road safety research and monitoring (European Commission, n.d.). Data linkage activities in Australia started in the 1970s and gathered momentum with the formation of the Western Australian Data Linkage System in 1995 and the Centre for Health Record Linkage in New South Wales in 2006. Australia has led the use of linked data and the development of linkage systems for over fifty years (Harrison et al., 2019; Smith and Flack, 2021) due to the Population Health Research Network, a nationwide network of data linkage units.

Data linkage projects can grow to encompass bigger regions within a country or whole nation, providing academics and policymakers with more extensive and diversified information. The distribution of studies across country income categories reflects each country's research resources, but some studies were conducted in regions with lower income levels. The significant gap between the number of studies conducted in high-income countries and fewer studies conducted in the rest of the world highlights the potential for health inequities due to variations in income classification and geographical location of research to inform reductions in road injuries. These inequities can result in disparities in health outcomes between different income groups or regions, as well as biases in the study findings. Overall, the studies demonstrate a global interest in the issue of data linkage (Fig. 8).

4.5. ICD usage

The WHO designed the international classification of diseases (ICD) to standardise diagnostic coding, universal health data reporting, and cross-country health outcome comparison (Harrison et al., 2021). It enhances healthcare billing, resource allocation, and cost-cutting (Hong and Zeng, 2023; De Coster et al., 2006). The usage of ICD codes in data linkage studies standardizes data and allows trustworthy and comparable study findings. Use of ICD codes also allows discovery of new health trends and patterns that traditional data collection and analysis may have overlooked. The comparison of selected papers shows that 55 % of research papers employed the ICD standard. Moreover, the usage of ICD codes is more common in North America, Europe and Oceania cases. Several studies have used ICD codes directly (Bambach and Mitchell, 2015), while others have opted to convert them to a local system, as shown by Ferreira et al. (2015) in Porto.

4.6. Population (Sample) coverage

Curry et al. (2019), Couto et al. (2016), and Brubacher et al. (2019) had the most cases included with the samples of 10,352,998; 40,000 and 15,022 respectively. In contrast, Cryer et al. (2001), Lopez et al. (2000), and Salifu & Ackaah (2012) have the lowest participant counts in their projects with case numbers of 2666, 497 and 400 cases respectively. Although minor injuries are extremely common and the systematic bias against reporting them is not fixed even when we have a large data set, larger studies are more likely to include them.

It is important to note that there is inconsistency in reporting the sample size in linkage studies. While some studies report the sample size

of the linked cases (Brubacher et al., 2019), others may only provide information on the number of police data cases (Alzaffin et al., 2023) or the number of the hospital/ambulance/EMC cases (Cryer et al., 2001; Ceklic et al., 2021) or the number of both police and hospital cases (Rosman and Knuiman, 1994). Furthermore, the inclusion criteria for selecting linked injury records in most studies is absent. This variation in reporting can make it challenging to directly compare the sample sizes across different studies.

4.7. Secondary data used

While crash/police data provides insights into the characteristics of crashes and the individuals/vehicles involved, healthcare data primarily focuses on injury diagnosis, treatments, and outcomes. Both data sources offer valuable information, but they have distinct emphases and serve different purposes in understanding and addressing the implications of crashes. The selected studies involved the utilization of linked data from multiple datasets, with one dataset being police-reported road crash data collected routinely. Various healthcare data sources were employed, including hospital records, discharge data, emergency department data, patient care reports, trauma registries, and death data. The primary objective of these research works was to develop comprehensive and accurate road crash information systems to support injury prevention efforts. The term "data fusion" was used in some studies to describe the integration of data from multiple sources to create a more comprehensive dataset (Alzaffin et al., 2023). Other studies referred to this process as the creation of "integrated datasets", which aimed to enhance the understanding of road traffic injuries, their severity, and the contributing factors (Curry et al., 2019; Abulatif et al., 2018; Amorim et al., 2014). Some research included workers compensation, insurance, birth, death, and injury monitoring data (Santiago et al., 2020; Boufous and Williamson, 2009). The three most frequent sources of linkage data are: police data; death data and medical data including hospital admission data; EMS and trauma registry data (Hosseinzadeh & Kluger, 2021; Amoros et al., 2006; Sciortino et al., 2005; Cirera et al., 2001).

4.8. Measures used

Linkage studies utilised a variety of measures based on their objectives. Mortality rates, injury severity indicators, hospital admission rates and expenses, crash rates and characteristics, and crash data completeness and correctness are employed (Dandona et al., 2008; Cirera et al., 2001). Using crash conditions and risk measures, several studies have focused on crash types or demographics of injured persons (Alzaffin et al., 2023; Ferreira et al., 2015; Amoros et al., 2006). Due to the variety of road traffic crashes and study topics, selecting the most common measures is challenging. These studies demonstrate traffic injury complexity with their many measures.

Injury Severity: Several studies focused on measuring the severity of injuries resulting from road traffic crashes. This includes metrics such as Police Reported Levels of Injury Severity, Abbreviated Injury Scale (AIS), MAIS Estimates, Injury Severity Score (ISS), Severity of 'Injuries (measured by MAIS) (Tarko and Azam, 2011; Kudryavtsev et al., 2013), hospital recovery costs (Bambach and Mitchell, 2015), and length of stay in the hospital (Ferreira et al., 2015). Standardised tools for classifying and measuring injury severity are provided by measures like the ISS, and New Injury Severity Score (NISS), based on the AIS. To calculate the likelihood of survival for a person with numerous injuries, two particular scoring systems—ICISS (ICD-based Injury Severity Score) and TMPM (Trauma Mortality Prediction Model) are based on the ICD injury classification. Physiological scoring systems, such as the Glasgow Coma Scale (GCS) and the Revised Trauma Score (RTS), evaluate injury severity based on physiological factors as distinct from anatomical injuries. GCS measures the degree of consciousness and neurological function, whereas the RTS combines vital indicators (such as blood pressure, respiration rate, and GCS score) to offer an overall estimate of

injury severity. The chance of survival based on the severity of injuries is represented by survival fractions, which are given as proportions or percentages. These valuations can provide light on the overall survival rates for various injury kinds and severities and are frequently derived from large-scale research or databases. One measure of the severity of an injury sometimes used is the length of hospitalisation or the time spent in a medical institution. Longer stays might indicate injuries or consequences requiring more extensive medical treatment. The presence of long-lasting or permanent mechanical effects from an injury is evaluated by the durable mechanical markers. Examples include the need for adaptive equipment (such as a wheelchair), the presence of chronic pain or another physical disability, or the necessity for continuing medical procedures.

Number of Injuries or Casualties: Many researchers examine the number of injuries or casualties resulting from road traffic crashes. This metric includes various aspects such as the number of road injuries, number of serious road injuries, number of non-fatal crashes, mortality rates, and number of deaths (Benavente et al., 2006; Paixão et al., 2015a).

Reporting and Underreporting: Metrics related to the reporting of road traffic crashes and injuries are also common. This includes metrics like underreporting rates of road crash injuries, amount of underreporting of road crash injuries to police, matching degree between police and hospital records, and reporting rates by age, gender, and demographic factors (Moore, 1997; Cryer et al., 2001; Alsop and Langley, 2001; Sciortino et al., 2005; Amoros et al., 2006; Gonzalez et al., 2006; Watson et al., 2015).

Crash Characteristics: Several studies focus on understanding the circumstances and characteristics of crashes. This includes metrics such as crash type, crash circumstances, crash location, crash time, and number of vehicles involved (Rosman & Knuiman, 1994; Stutts and Hunter, 1999; Tin et al., 2013; Mitchell et al., 2015).

Hospital-related Metrics: Metrics related to hospital records and health outcomes are also prevalent. These include metrics like hospital inpatient records, health outcomes of young driver crashes, and hospital admission (Moore, 1997).

4.9. Data linkage approach

Data linkage incorporates numerous approaches, including probabilistic, deterministic, manual, and customised methods (Dusetzina et al., 2014). The choice of method and algorithm is influenced by factors such as available resources, research questions, and the quality and quantity of variables to uniquely identify individuals. When linking police and hospital data, typical algorithms include probabilistic record linkage techniques and occurrence-based conceptual algorithms (Amorim et al., 2014; Benavente et al., 2006). Occurrence-based algorithms utilise predefined principles to establish connections between records based on co-occurrence, with the goal of identifying plausible matches and minimising spurious matches (Amorim et al., 2014). The probabilistic method assigns probabilities to potential matches based on shared characteristics between records, utilising algorithms like the Fellegi-Sunter algorithm or probabilistic record linkage method (Victor & Mera, 2001). Numerous studies have utilised the probabilistic method, such as Edwards & Gutierrez (2023), Soltani et al. (2022), and Wilson et al. (2012). In contrast, the deterministic method creates direct matches between records based on predetermined rules or precise matching criteria (Shlomo, 2019). Several studies, such as Tainter (2020) and Kamaluddin (2019), have utilised deterministic methods. Studies like Watson et al. (2015) and Amorim et al. (2014) have utilised both deterministic and probabilistic approaches.

The differences in linkage rates across studies can be related to the methods employed for linking, the types of data being linked, and the presence of sufficient data items. Probabilistic methods affect linkage rates differently than deterministic methods, with probabilistic methods offering more flexibility and accommodating errors. The nature of

records, such as differences in data collecting and data granularity between police crash and hospital admission records, can be difficult. Successful matches can be strongly influenced by the availability of unique IDs or dependable matching criteria, as well as data quality and consistency. Contextual factors, population features, and variations in data recording practices affect linkage rates. Variations in linkage terminology and definitions might cause misunderstanding and inconsistency when comparing outcomes across studies.

Record linkage can be separated from the analysis of linked records to safeguard subjects' privacy and adhere to ethical standards. This is accomplished by involving two distinct groups: the Investigators group, which is responsible for accessing original sensitive data, and the linkage group, which is responsible for the actual record linkage, using project person numbers (PPN) in lieu of personally identifiable information (Dusetzina et al., 2014). This separation protects the privacy of individuals while allowing for efficient and secure data linkage.

4.10. Software tools and statistical tests used

Among the selected studies, LinkSolv and Linkage Wiz were the software tools utilised most frequently. Automatch, SPSS, SAS, Microsoft SQL, MapInfo, manual, R and R-studio, and self-written programmes were also utilized. The statistical methods included both descriptive (such as frequencies, proportions) and inferential tests such as Chi-square (15 studies), Logistic regression (six studies), Cramer's V (five studies), Ordered probit regression (three studies), and Poisson regression (two studies).

4.11. Linkage quality and linkage rate

The quality of linkage is critical in record linkage studies as it determines the extent to which linked records accurately represent the same individuals or entities across different datasets. Assessing and ensuring linkage quality is essential for drawing valid conclusions and making reliable inferences from linked data (Elstad et al., 2023). Errors in data linkage can introduce bias and lead to inaccurate results if not reported properly (Harron et al., 2017). These errors may arise from factors such as the absence of unique identifiers, missing identification information, or flaws in the linkage program itself (Christen & Goiser, 2007). To address these issues, transparency, and documentation of the linkage process, including limitations and steps taken to address errors, are crucial. Sensitivity analyses and evaluating the robustness of linkage results can further help assess the impact of potential errors (Harron et al., 2020; Harron et al., 2017).

Linkage rate is the percentage of records linked across data sources to produce a more complete database. In many cases, however, the methods for assessing the completeness of linkage were also inconsistent (Bohensky et al., 2010). The "Separation Principle" refers to arrangements in when linkage and analysis are separated and data transformity are organised so that the investigator never has access to direct identifiers (ABS, 2010). In an extension of separation, direct identifiers are encrypted before being provided to the linkage service, all records for a study being encrypted in the same way (Lim et al., 2022).

In the selected studies, the terminology used to describe the matching procedure can vary, leading to potential confusion when comparing outcomes. Different measures of the linkage process are introduced. Most commonly used measures are: linkage rate (Rosman & Knuiman, 1994; Conderino et al., 2017), share of linked records (Ceklic et al., 2021), success rate (Cryer et al., 2001), match rate (Hosseinzadeh et al., 2022; Tarko and Azam, 2011; Kudryavtsev et al., 2013; Langley et al., 2003; Santiago et al., 2020; Tin Tin et al., 2013), matching rates (Tainter et al., 2020; Kamaluddin et al., 2019), matching records (Lujic et al., 2008), match probability (Curry et al., 2019), discordance rate (Tsui et al., 2009; Edwards & Gutierrez, 2023; Soltani et al., 2022; Watson et al., 2015), reporting rate (Amorim et al., 2014) and false positive/false negative rates (Amorim et al., 2014).

Table A1

List of nominated papers.

DOI
https://doi.org/10.1016/0001-4575(94)90091-4
https://doi.org/10.1037/e439752008-001
https://doi.org/10.1016/S0001-4575(99)00007-X
https://doi.org/10.1016/S0001-4575(99)00130-X
https://doi.org/10.1016/S0001-4575(00)00049-X
https://doi.org/10.1080/10286580108902567
https://doi.org/10.1136/ip.7.3.234
https://doi.org/10.1136/ip.9.4.376
https://doi.org/10.1016/j.aap.2005.06.010
https://doi.org/10.1177/0361198106195300121
https://doi.org/10.1016/j.aap.2009.01.015
https://doi.org/10.1097/01.ta.0000229816.16305.94
https://doi.org/10.1016/S0001-4575(98)00051-7
https://doi.org/10.1016/j.aap.2005.12.015
https://doi.org/10.1016/j.aap.2005.11.006
https://doi.org/10.1136/ip.2008.019638
https://doi.org/10.1016/j.jsr.2008.10.008
https://doi.org/10.1111/j.1753-6405.2008.00162.x
https://doi.org/10.1136/ip.2008.018630
https://doi.org/10.1080/15389580802593699
https://doi.org/10.1016/j.aap.2008.09.005
https://doi.org/10.1177/1010539509356808
https://doi.org/10.2471/BLT.10.080317
https://doi.org/10.1016/j.annepidem.2011.05.007
https://doi.org/10.1016/j.aap.2011.03.027
https://doi.org/10.1080/17457300.2011.628752
https://doi.org/10.1016/j.aap.2011.03.007
https://doi.org/10.1016/j.aap.2012.12.022
https://doi.org/10.1186/1471-2458-13-420
https://doi.org/10.3141/2432-02
https://doi.org/10.1016/j.aap.2015.09.013
https://doi.org/10.1016/j.ssci.2015.05.013
https://doi.org/10.1016/j.aap.2015.04.039
https://doi.org/10.1590/1980-5497201500010009
https://doi.org/10.1590/0102-311X00081314
https://doi.org/10.1016/j.aap.2015.06.011
https://doi.org/10.1016/j.jsr.2016.03.008
https://doi.org/10.1080/15389588.2015.1128533
https://doi.org/10.1016/j.jsr.2016.05.002
https://doi.org/10.1080/15389588.2015.1042577
https://doi.org/10.5123/S1679-49742018000200018
https://doi.org/10.1016/j.aap.2017.02.011
https://doi.org/10.1016/j.aap.2017.06.018
https://doi.org/10.1016/j.jth.2019.100606
https://doi.org/10.1080/15389588.2019.1679552
https://doi.org/10.1016/j.aap.2017.02.022
https://doi.org/10.1080/17457300.2018.1476385
https://doi.org/10.1016/S2468-2667(19)30057-X
https://doi.org/10.1177/0361198120935439
https://doi.org/10.1016/j.jsr.2020.04.012
https://doi.org/10.1080/10903127.2020.1771487
https://doi.org/10.1016/j.aap.2021.106053
https://doi.org/10.1016/j.aap.2021.106053
https://doi.org/10.1080/15389588.2022.2083612
https://doi.org/10.1016/j.jsr.2022.08.002
https://doi.org/10.1016/j.aap.2022.106897
https://doi.org/10.1080/15389588.2022.2143236
https://doi.org/10.1177/03611981221151073

The term “match rate” refers to the proportion or percentage of records from one dataset that successfully match or link with records from another dataset (Sciortino et al., 2005). Similar to this, the term “linkage rate” refers to the proportion or percentage of records that have been successfully linked or matched between the two databases (Conderino et al., 2017). The term “discordance rate” often refers to the percentage or rate of records that do not match or align between the two databases (Watson et al., 2015). The accuracy of the linkage process is also evaluated using the words “false positive rate” and “false negative rate” (Mitchell, et al., 2014; Conderino et al., 2017). In contrast to the false negative rate, which reflects the proportion or percentage of missed matches or links that are not detected by the linkage process, the false positive rate is the proportion or percentage of wrong matches or links

that are incorrectly identified by the linkage process.

However, the measures of linkage rate, do not necessarily reflect the quality of the linkage. In other words, achieving a high linkage rate does not guarantee a high-quality linkage (Bohensky, 2015). In these cases, false negatives are almost impossible to determine and therefore linkage quality is also impossible to determine (at least for researchers). As studies use different denominators, comparing the ‘data linkage rates’ can be challenging. Some are interested in how many injury records link with a hospital record - police data being the denominator (Amorim et al., 2014), while others are mainly interested in how many road crash injury hospital records link to a road crash record - hospital data being the denominator (Rosman & Knuiman, 1994; Ceklic et al., 2021). Some studies are interested in both (Short and Caulfield, 2016). The denominator or ‘direction’ of the linkage really matters in how these rates (or % linked) are interpreted and what the implications are. There are also some who are expecting full ascertainment and so the ‘linkage rate’ cited is seen as a reflection of linkage quality and effectiveness.

In their population-based research on urban India, Dandona et al. (2008) reported a 22 % under-reporting of fatal road traffic injuries, and Cryer et al. (2001) found that manual linkage to hospital admissions data in pedestrian crashes was 72 % successful. Linkage rates vary by injury type and mode of travel. Rosman & Knuiman (1994) reported a 64 % hospital-to-police link rate, but the proportion varied from 29 % for motorcyclists in single-vehicle accidents to 79 % for motor vehicle drivers. Santiago et al. (2020) reported a significant difference in match rates between the deterministic approach (52.38 %) and the probabilistic method (70.41 %). The match rate between the two databases of hospital and police records was 70.5 % in average in a data linkage project in Malaysia, the matching rate was highest for the highest severity level (Kamaluddin et al., 2019). Wilson et al. (2012) reported an average connection rate of 46.2 %, while linking rates varied by injury severity, injury site, injury type, age and duration of hospital stay. Linkage rates for moderate injuries varied by age, injury type, and length of hospital stay. Only age and length of hospital stay were substantially linked with linkage rate variations in major injuries. The linking rate rose with increasing degrees of injury severity and was significantly lower for specific ethnic groups of those injured (Rosman & Knuiman, 1994).

The reporting deficiency in data linkage reports may arise from a lack of accessible information about the specific data linkage process used in the dataset (Palamuthusingam et al., 2019). In some projects, the data used were linked by a trusted third party, leading to uncertainties about the precise details of the linkage process (Elstad et al., 2023). The reviewed studies provided limited clarity on how much the authors knew about specific aspects of the linkage process employed for their dataset. Insufficient information was often provided concerning the dataset’s origin, linkage variables utilized, methods employed, and the assessment of linkage outcomes. The absence of insights into the decision-making during the linkage process hampers a comprehensive understanding of the dataset used for analysis (John et al., 2014).

5. Discussion

5.1. Underestimation and bias in reporting vulnerable groups

Many studies have found that police-reported data underestimates road traffic injuries and by different degrees for different road user types thus integrating multiple sources improves injury statistics. Lombardi et al. (2022) revealed that hospital data identified underrepresented groups better than police data, stressing the need for multiple data sources. Lopez et al. (2000) found police reports and trauma registry data enhance severity categorisation. Hospital-EMS record integration may improve injury data accuracy, but its limitations require more investigation.

The selected studies have found that police-reported injury data undercounts injuries especially for non-motorised modes and for

Table B1

Evolution of key terms in the area of data linkage of road crash data and hospital crash data.

Concept	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Crashes	Crashes (Janstrup et al., 2016; Lopez et al., 2000; Kamaluddin et al., 2019; Wilson et al., 2012; Benavente et al., 2006; Mitchell et al., 2015)	Collisions (Soltani et al., 2022; Stutts and Hunter, 1999; Alsop 2001; Sciortino et al., 2005)	Road accidents (Ferreira et al., 2015; Barchielli et al., 2002; Aptel et al., 1999; Couto et al., 2016)	Traffic accidents (Hu et al., 2011; Paixão et al., 2015a; Weijermars 2016; Amorim et al., 2014; Abdulatif 2017)	Motor vehicle crashes (Weijermars 2016; Wilson et al., 2012; Alsop 2001; Alzaffin et al., 2023; Ceklic et al., 2021; Tainter et al., 2020)	Incidents (Edwards 2023;		
Injury severity	Injury severity (Cryer et al., 2001; Bambach et al., 2015; Alzaffin et al., 2023; Roseman & Knuiman, 1994; Hosseinzadeh & Kluge, 2022; Tarko & Azam, 2011)	Severity of injuries (Barchielli et al., 2002)	Crash severity (Tainter et al., 2020)	Traffic injury severity (Tarko & Azam, 2011)				
Police data	Police data (Barchielli et al., 2002)	Police reports (Janstrup et al., 2016; Paixão et al., 2015a; Wilson et al., 2012)	Police-reported information (McDonald et al., 2009)	Police-reported data (Hu et al., 2011; Schneider et al., 2023; Hosseinzadeh & Kluge, 2022)	Police records (Lujic et al., 2008; Kamaluddin et al., 2019; Lateef)	Police reporting (Jeffrey et al., 2008)		
Data linkage	Data linkage (Lujic et al., 2008; Kamaluddin et al., 2019; Wilson et al., 2012; Curry et al., 2019)	Linked police-hospital data (Alzaffin et al., 2023)	Record linkage (Tainter et al., 2020)	Linked dataset (Wilson et al., 2012)	Linked database (Cryer et al., 2001; Couto et al., 2016)	Linked data (Watson et al., 2015; Ceklic et al., 2021)	Police-hospitalisation linked data (Mitchell et al., 2015)	
Hospital data	Hospital data (Curry et al., 2019; Bambach et al., 2015; Sciortino et al., 2005; Aptel et al., 1999)	Hospital or hospitalisation records (Kamaluddin et al., 2019; Wilson et al., 2012; Edwards 2023; Bambach et al., 2015; Amorim et al., 2014; McDonald et al., 2009)	Hospital discharge data (Weijermars 2016; Wilson et al., 2012; McDonald et al., 2009)	Medical records (Ferreira et al., 2015; Janstrup et al., 2016; Benavente et al., 2006; Roseman & Knuiman, 1994)	Prehospital data (Gonzalez et al., 2006)	Hospital injuries data (Janstrup et al., 2016)	Patient data (Edwards, 2023)	Hospital admission data (Cryer et al., 2001)
Injury patterns	Injury patterns (Weijermars 2016; Cryer et al., 2001; Roseman & Knuiman, 1994)	Injury location (Lateef, 2017; Boufous et al., 2008; Kudryva et al., 2013; Hosseinzadeh & Kluge, 2022)	Injury sites (Wilson et al., 2012)	Injury types (Wilson et al., 2012)	Injury characteristics (Wilson et al., 2012)	Injury specifics (Benavente et al., 2006)		
Data linkage	Data linkage (Edwards 2023; Benavente et al., 2006; Amorim et al., 2014; Tainter et al., 2021)	Record linkage (Barchielli et al., 2002)	Linked data (Wilson et al., 2012; Tarko & Azam, 2011)	Data integration (Conderino et al., 2017; Abdulatif 2017;	Data fusion (Alzaffin et al., 2023)			
Under reporting	Under reporting (Ferreira et al., 2015; Janstrup et al., 2016; Jeffrey et al., 2008; Wilson et al., 2012; Lateef; Edwards 2023; Dandona 2008; Amoros et al., 2006)	Reporting bias (Janstrup et al., 2016; Cryer et al., 2001; Watson et al., 2015; Sciortino et al., 2005; Amoros et al., 2006)	Misclassification (Ferreira et al., 2015; Bambach et al., 2015; Watson et al., 2015; Couto et al., 2016)	Under estimation	Over estimation			
Injuries	Injuries (Sciortino et al., 2005; Alzaffin et al., 2023)	Non-fatal injuries (Tainter et al., 2020)	Casualty (Ferreira et al., 2015; Jeffrey et al., 2008; Cryer et al., 2001; Amoros et al., 2006)	Non-fatal traffic injuries (Ferreira et al., 2015; Aptel et al., 1999)	Traffic-related injuries (Conderino et al., 2017)			
Data quality	Data quality (Conderino et al., 2017; Jeffrey et al., 2008; Benavente et al., 2006; Mandacaru et al., 2017)	Data accuracy (Conderino et al., 2017; Bambach et al., 2015; Abdulatif 2017; McDonald et al., 2009)	Data verification (McDonald et al., 2009; Kudryva et al., 2013)	Data completeness (Conderino et al., 2017; Barchielli et al., 2002)	Data validity (Wilson et al., 2012; Amorim et al., 2014)Data reliability (Abdulatif 2017; Kudryva et al., 2013)	Data integrity (Bambach et al., 2015)		

socially-disadvantaged groups. Short and Caulfield (2016) found twice as many clinically significant injuries according to hospital records as were reported by police, while Tin et al. (2013) found that at least 70 % of self-reported pedal cyclist crashes were not reported to police. Police records are commonly not sufficient to assess injury diagnosis. These studies also found police crash data underreports pedestrian injuries. Sciortino et al. (2005) found that police crash records underreported pedestrian injuries by 21 %, notably African-American and female pedestrians. These studies demonstrate the importance of trustworthy road traffic injury data for road safety policies and activities. They emphasise the need for combining multiple data sources and approaches to adequately capture road traffic injuries and reduce under-reporting. Vehicle type, casualty age, gender, and region may impact injury reporting.

The selected studies highlight the need to consider differences between road user types especially vulnerable road users. Wang et al. (2019) found that pedestrians were the most vulnerable road users in a study in China, whereas Weijermars et al. (2016) found that Dutch moped riders had the lowest hospital discharge record completeness. Governments and public health officials may use these trends to implement policies that reduce injuries and fatalities.

5.2. Method of linkage

The selected studies underscore the significance of probabilistic record linkage. The absence of unique identifiers in datasets poses challenges for accurate linkage, often necessitating the adoption of probabilistic record linkage as the preferred method (Amorim et al., 2014). Santiago et al. (2020) demonstrated the superior efficiency and accuracy of probabilistic record linkage compared to deterministic linking. Short and Caulfield (2016) employed probabilistic record linkage to estimate the size of injury datasets and discovered that clinically significant injuries were twice as numerous as those reported by the police. Probabilistic record linkage accommodates imperfections in identifiers and has the potential to capture more genuine matches.

Most included studies did not provide a rationale for the chosen linkage approach, making it difficult to ascertain the prevailing data linkage method. Each method has its merits and drawbacks, contingent on the research topic and available data. The selection of a linkage method depends on factors such as data quality, availability of direct identifiers, research objectives and subjective judgment (Dusetzina et al., 2014). A linkage algorithm's accuracy is measured using specificity, sensitivity, positive predictive value, and negative predictive value (Dusetzina et al., 2014).

Deterministic methods are recommended for information-rich scenarios with high-quality identifiers due to their ease of implementation and interpretation (Haneef et al., 2022). In information-poor or low-quality data situations, probabilistic methods outperform deterministic methods. Even in information-rich scenarios, researchers studying rare diseases may opt for probabilistic methods to maximize sample size. Ultimately, the decision hinges on the researchers' goals, budget, timeline, and available resources, representing the "art" of record linkage. Each researcher must carefully weigh the advantages and disadvantages to select the most suitable method for their project's context and research question (Dusetzina et al., 2014).

There are two broad reasons to use multiple data sources, combined by record linkage, when studying road crash injuries. One reason is to improve case ascertainment. The other is to add to what is known about ascertained cases (e.g. police data may be the richest source of information about crashes, while hospital data may be the richest source on injuries sustained in the crashes). Under some circumstances, linkage of multiple data sources, all of which record crashes in the same population, can be seen as equivalent, or nearly equivalent, to the capture-recapture method (Morrison and Stone, 2000). Conceiving a linked data study as a capture-recapture project carries the benefit of encouraging consideration of whether the data sources used satisfy the

assumptions required for validity of that method, notably independence of the sources, adequate quality of the data and non-transience of the population to which the data refer (Morrison and Stone, 2000; Jarvis et al., 2000). Morrison and Stone (2000) concluded that while the police and health sources they used for road fatalities were nearly complete, the sources they used for non-fatal cases were much less complete and exhibited properties that led them to conclude that ascertainment-adjusted rates should be viewed with caution. A similar caveat is likely to apply to the findings of many other road injury linkage projects, though not all provide as clear a statement of the problem. While the analogy with capture-recapture is useful, it is not exact. Crash data and hospital data are inherently different sources and can be expected to have overlapping but different scope of inclusion.

5.3. Ambiguity in terminology of data linkage

The terminology and language may confuse linkage and matching studies. This also applies to hospital data descriptions, which may be complicated. Some studies refer to hospital data as a general term (Sciortino et al., 2005; Aptel et al., 1999), while others may refer to discharge data or inpatient/outpatient data or hospital admission/stay data (Conderino et al., 2017; Short & Caulfield, 2016; Tsui et al., 2009). Linkage project metrics vary, which may make it difficult to compare studies. This variance in metrics may make linkage results hard to compare. Another concern is the heterogeneity of previous studies' case number reporting. Certain studies may provide information on the volume of police and/or hospital data records (Ferreira et al., 2015), whereas others may not explicitly disclose the number of records (Abulatif et al., 2018). However, a majority of studies have reported both the size of crash data from law enforcement agencies and health sector data (Ceklic et al., 2021; Curry et al., 2019; Conderino et al., 2017; Amorim et al., 2014; Dandona et al., 2008; Cryer et al., 2001). This poses a challenge in determining the magnitude of records and doing comparative analysis with other research investigations.

In linkage studies, it is important to consider the limitations of relying solely on the linkage rate as an indicator of linkage quality. Researchers should strive for transparency in the linkage process, acknowledge potential errors, and implement measures to ensure the reliability and validity of the linked data. Given these issues, global linkage initiatives' methodologies and results should be reported in a standardised manner. By adopting standard terminology and metrics, researchers may be able to better compare studies, find linkage, and find matching best practises. Data utilisation and the analytical findings may improve.

5.4. Standardisation of injury coding

The studies emphasised diagnosis and injury severity classification. Over half of studies used the ICD to classify and code injuries, allowing for more standardised comparable assessments (Short and Caulfield, 2016; Sciortino et al., 2005). For instance, Bambach and Mitchell (2015) used the ICD-10 injury diagnoses S00-T89 to determine the severity of the injury by calculating the Survival Risk Ratio for each individual code. Using the ISS, Tsui et al. (2009) revealed that police injury grading differs from the 12-h hospital stay criteria. Coding criteria may have influenced injury counts in several studies. Moore et al. (2016) noted that the decentralised system caused jurisdictional variances in data linked, linkage methods, and approval procedures, making multi- and cross-jurisdictional research more difficult than with a uniform national approach.

5.5. The gap between middle-income and high-income countries

Data linkage has helped identify and classify road traffic injuries in middle-income and high-income countries, but it also presents challenges. High-income countries have more resources to collect and

handle data than low-income countries (Imprialou & Quddus, 2019; Kamaluddin et al., 2019; Tercero & Andersson, 2004). Low- and middle-income countries need trustworthy data systems to identify road safety issues and track progress (Mandacaru et al., 2017), given their larger proportion of road crash injuries and fatalities, for example in the context of Iran (Soltani and Askari, 2017; Soltani and Askari, 2014). Data linkage may improve road safety in low- and middle-income countries, but many nations need more data infrastructure and capacity training (Heydari et al., 2019). Most low-income countries have people, infrastructural, and budgetary restrictions that hinder big data management (Luna et al., 2014). These issues may be addressed by government agencies, academic institutions, and other stakeholders (Suriyawongpaisal & Kanchanasut, 2003). Gutierrez et al. (2022) found that local epidemiological data sources in low- and middle-income countries may improve global estimates and boost local stakeholders' trust.

5.6. Standard reporting

Documentation and support for researchers regarding data linkage methodology play a crucial role in ensuring transparency and accessibility of linked data (National Academies of Sciences, Engineering, and Medicine, 2022). Some studies examined the shortcomings and challenges of data linkage in identifying road traffic injuries, such as the lack of regulation and national standardisation of reporting instruments (Paixão et al., 2015b; Reurings and Stipdonk, 2011). A common critique of such studies is the lack of detailed information on the linkage process, which affects transparency and reproducibility (Elstad et al., 2023; Sonesson et al., 2023). Without understanding the methodology and potential sources of error, accurate interpretation of results becomes challenging. To enhance transparency and enable readers to gauge the quality and potential biases introduced by the linkage process, studies should improve the reporting of data linkage information. Providing comprehensive details about the data linkage process is essential for the accurate interpretation and reproducibility of study results.

Insufficient reporting of data linkage can introduce unknown bias, compromising the validity and reliability of study findings (Bohensky et al., 2011). By contrast, efforts to enhance data linkage reporting can contribute to increased openness and replicability in the research (Mars et al., 2016; Gilbert et al., 2018). Better reporting of methods and sources would be valuable for researchers seeking to replicate study findings, facilitating knowledge transfer from high-income countries to lower income countries, and assisting policymakers in gaining insights from the data linkage literature. To ensure proper interpretation, it is crucial to provide comprehensive information about the data linkage process in research publications. This includes details on dataset origin, variables used for linkage, methods employed, the rate of missing data, and linkage quality evaluation. Transparent reporting promotes understanding, replication, and critical evaluation of study findings (Mourby et al., 2019).

5.7. Data linkage challenges

There are challenges involved in the linkage of clinical data with road traffic injury data. These challenges include ethical, legal, cultural, and technical issues. Mitchell et al. (2014) state that the ethical challenges associated with the use of personal data include authorization, transparency, and the potential for damage to individuals or organisations. Legal challenges include privacy restrictions, data ownership and exchange restrictions, and prohibitions on the use of personal information (Kalkman et al., 2022). Sharing or using personal data for research may be illegal (Tulu et al., 2013; Ameratunga et al., 2006; Suriyawongpaisal & Kanchanasut, 2003). Data linkage projects require careful consideration of privacy concerns and the need to protect individual privacy (Elstad et al., 2023).

Cultural differences in data sharing and privacy, as well as

confidence and privacy concerns, are challenges (Dhillon and Torkzadeh, 2006). Combining data from multiple systems, formats, and quality, and missing data are technological challenges. Regulations, procedures, investments in new technology and infrastructure, stakeholder engagement, and communication concerns are further challenges (Engin et al., 2020). These risks include erroneous or incomplete data from human or computer mistakes in collection, storage, and categorization (Clark, 2004). Leakage of data compromises people's privacy by disclosing their names and sensitive information (Nair & Tyagi, 2021). Discrimination is another potential risk (Gopal et al., 2021; Prince & Schwarcz, 2019). Legal difficulties related to personal information collection, usage, and sharing, data ownership, and access, as well as abuse of linked information by those with access, are also a risk (Tikkinen-Piri et al., 2018). Protections, ethical standards, and stakeholders should be included in research design and regulatory systems to reduce these risks.

Police-hospital data linkage systems face challenges such as data harmonisation, anonymizing sensitive police data, and accessing confidential health information (Brophy et al., 2023). Data harmonisation across diverse police systems, anonymizing narrative-based sensitive police reports, accessing confidential health data, ensuring data accuracy and completeness, and assessing data linkage benefits are challenges in different countries (Brophy et al., 2023; Cherry et al., 2018). However, not all studies reviewed in this paper reported obtaining approval from human research ethics committees, or described other aspects of project operation relevant to the privacy and safety of the people whose records were used.

5.8. Cutting-edge technologies in data linkage

Data linkage has become simpler with the rapid growth of data science techniques such as machine learning, data mining, and big data (Raghupathi & Raghupathi, 2014). These methods may identify duplication and normalise variables across data sets (Vatsalan et al., 2017). These techniques may reveal traffic crash causes and injuries, which may inform road safety laws and measures. Advanced machine learning algorithms may uncover patterns and trends in enormous data sets that statistical methods may overlook (Gkoulalas-Divanis et al., 2021). Natural language processing can assess police and medical reports. Data visualisation techniques may help stakeholders and decision-makers understand complex data. Interactive dashboards enable data analysis and report customisation (Raghupathi & Raghupathi, 2014). Machine learning and big data may identify road safety threats including crashes and injuries in real time (Torbaghan et al., 2022). Prediction algorithms may detect motor vehicle collision and injury risk variables via data mining.

Improved accuracy, scalability, flexibility, automated feature engineering, real-time linkage, handling missing data, learning complex relationships, decreased human error, cost-efficiency, and continuous improvement are just a few of the benefits that machine learning brings to data linkage projects (Ranbaduge et al., 2023; Martinez-Garcia and Hernández-Lemus, 2022). By automating the linking process, adjusting to different data formats, and scaling effectively to accommodate huge, complicated datasets, it performs better than traditional approaches, eventually lowering mistakes and costs while enabling real-time updates and advanced pattern identification.

5.9. Research limitations and future research

This systematic review was limited to English language studies, which resulted in the omission of several papers published in diverse geographical and social contexts. Our search for published papers was limited to three widely used bibliographic databases (PubMed, Web of Science, Scopus). Use of additional sources, particularly the Transport Research International Database (TRID) as a leading worldwide authority in transportation research, may have improved

comprehensiveness. Additionally, our review did not extend to the grey literature, book chapters, industry or governmental reports, non-refereed working papers, and conference proceedings. Inclusion of which could have provided further insights. Grey literature may provide data not found within commercially published literature, providing an important forum for disseminating studies with null or negative results that might not otherwise be disseminated. Grey literature may thus reduce publication bias, increase reviews' comprehensiveness and timeliness, and foster a balanced picture of available evidence (Paez, 2017). The sophisticated technical aspects involved in data linking, including algorithms, conceptual models, and complex technological parameters and issues, can result in papers being available online but not being included in journals or peer-reviewed publications. Hence, we would recommend that future research initiatives embrace a more inclusive methodology by exploring various types of grey literature. Implementing this comprehensive approach will guarantee a more comprehensive evaluation, avoiding the omission of crucial observations and enhancing the thoroughness of analysis in data linkage investigations. The scope of this research was limited to data linkage papers that provided sufficient detail on the linkage process and its outcomes. As a result, some studies with only brief descriptions of data linkage but with a focus on the applications of linked data for analysing traffic crashes were not included. Adding further datasets—like pre-hospital/ambulance and insurance data—could broaden the study's focus and advance knowledge of the subject.

Future research directions in data linkage include improving methods for accuracy, timeliness, and efficiency, enhancing data quality and completeness, standardizing data collection, and investigating factors influencing injury severity discrepancies. Harmonizing police and hospital data collection, storage, categorisation, and addressing missing data are crucial considerations. Understanding how data linkage is utilized across high-income, middle-income, and low-income countries, as well as within regions and nations, is essential for developing flexible data linkage methods. Examining the impact of data linkage on the quality and scope of injury research and its contribution to evidence-based policymaking is also a research priority. Additionally, assessing biases in estimates of vulnerable road users and marginalized groups is important for targeting injury prevention strategies effectively. Moreover, there is a need to fully leverage cutting-edge technologies like machine learning and artificial intelligence in order to overcome existing challenges and unlocking the full potential of data linkage studies.

6. Conclusion

Improving the accuracy and completeness of injury data is the main goal of data linkage studies in the field of road traffic injury research. These studies highlight the persistent bias and underreporting in police-reported data, highlighting how important it is to integrate data from many sources—especially hospital records—to address these problems. Our analysis highlights the superiority of hospital data over police data in identifying marginalised and vulnerable populations, an important finding for developing public health and traffic safety strategies. Furthermore, these studies emphasise the necessity of considering variations across different types of road users, with an emphasis on the increased susceptibility of pedestrians and specialised user groups.

To standardise injury severity categorisation and data and enable more comparative evaluations, the adoption of standardised coding systems such as ICD is essential. Moreover, the increasing assimilation of sophisticated data science methods, such as machine learning, is indicative of the changing terrain of data linkage approaches that adjust to the requirements of every research setting. Studies of data linking show a great deal of variation in several areas, including data sources, techniques, software, environments, and time spans. There is a rising trend towards merging primary data sources, such as hospital records and police reports, to improve the comprehensiveness of data on traffic

injuries. Because many datasets lack homogeneous unique IDs, probabilistic record linking emerges as the ideal technique, with the option between probabilistic and deterministic linkage methods driven by contextual factors like data quality, research goals, and resource availability. While the discussion of software tools used is limited, it signals an area for potential research and standardisation. Research settings encompass a spectrum from high-income to middle-income countries, reflecting resource disparities and challenges. The temporal scope of investigations varies, with some studies focused on short-term analyses while others take a longitudinal perspective to capture evolving trends and patterns.

Our investigation ends with a focus on rural road safety concerns and vulnerable road users, such as young people, motorcyclists, and pedal cyclists. Working together, legislators and public health specialists may create sophisticated linking plans that will reduce traffic injuries worldwide. The journey ahead entails negotiating challenges with linking rates, geographical variations, harmonising law enforcement and healthcare data gathering procedures, and setting strict quality evaluation criteria. Particularly in less-studied economic contexts, data linking practises might differ greatly between nations and regions. Further work is required to develop adaptable methods and advance fair road safety initiatives. When recognised limitations are addressed, data linkage becomes a powerful instrument for improving traffic safety and minimising injury. To support research replication and evidence-based policymaking, standardising data linking practises and improving reporting protocols are essential first steps. Research integrity and practical value can be greatly increased by reporting data connection projects with specific suggestions for data sources, matching criteria, and quality control. Prospects for data-driven road safety solutions appear bright when efforts are directed towards improving data link procedures.

Policymakers and public health authorities may use this study to help develop focused measures that will lower the number of road fatalities and injuries. In addition to highlighting the value of data linking for improvement, it also emphasises the advantages of combining data from other sources, such as ambulance records, to combat underreporting and get complete crash information, including exact location data. The incorporation of information from many sources, such as hospital records and ambulance data, holds potential to enhance approximations and expand comprehension of traffic-related injuries. This will enable the creation of precisely focused interventions and well-informed policymakers.

Credit authorship contribution statement

Ali Soltani: Methodology, Software, Investigation, Data analysis, Visualization, Writing – review & editing. **James Edward Harrison:** Conceptualization, Resources, Validation, Writing – original draft, Writing – review & editing. **Courtney Ryder:** Methodology, Supervision, Validation, Writing – review & editing. **Joanne Flavel:** Methodology, Investigation, Validation, Writing – review & editing. **Angela Watson:** Conceptualization, Resources, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors would like to acknowledge Austroads (Grant #SAG6132) for funding the road injury linkage project in Australia that provided the inspiration for this paper.

Data availability

Data will be made available on request.

Appendix

Table A1
Table B1

References

- Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. *Transp. Res. A Policy Pract.* 71, 31–45.
- Abulatif, L. I., & Grupo Técnico de Parceiros do Projeto Vida no Trânsito (2018). Data Integration Process: an information management model for multiple databases on traffic accidents in Brazil. Processo de integração de dados - um modelo de gestão da informação para múltiplas bases de dados de acidentes de trânsito no Brasil. *Epidemiologia e serviços de saúde : revista do Sistema Unico de Saude do Brasil*, 27(2), e2017160.
- National Academies of Sciences, Engineering, and Medicine. (2022). Improving Consent and Response in Longitudinal Studies of Aging: Proceedings of a Workshop. Washington, DC: The National Academies Press.
- Alsop, J., Langley, J., 2001. Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accid. Anal. Prev.* 33 (3), 353–359. [https://doi.org/10.1016/S0001-4575\(00\)00049-X](https://doi.org/10.1016/S0001-4575(00)00049-X).
- Alzaffin, K., Kaye, S.A., Watson, A., Haque, M.M., 2023. A data fusion approach of police-hospital linked data to examine injury severity of motor vehicle crashes. *Accid. Anal. Prev.* 179, 106897 <https://doi.org/10.1016/j.aap.2022.106897>.
- Ameratunga, S., Hajar, M., Norton, R., 2006. Road-traffic injuries: confronting disparities to address a global-health problem. *Lancet* 367 (9521), 1533–1540.
- Amorim, M., Ferreira, S., Couto, A., 2014. Linking police and hospital road accident records: how consistent can it be? *Transp. Res. Rec.* 2432 (1), 10–16. <https://doi.org/10.3141/2432-02>.
- Amoros, E., Martin, J.L., Laumon, B., 2006. Under-reporting of road crash casualties in France. *Accid. Anal. Prev.* 38 (4), 627–635. <https://doi.org/10.1016/j.aap.2005.11.006>.
- Aptel, I., Salmi, L.R., Masson, F., Bourdè, A., Henrion, G., Erny, P., 1999. Road accident statistics: discrepancies between police and hospital data in a French island. *Accid. Anal. Prev.* 31 (1–2), 101–108.
- Australian Bureau of Statistics (ABS). (2010). Creating a safe and effective environment for data integration. A Guide for data integration projects involving Commonwealth data for statistical and research purposes. Retrieved from <https://statisticaldataintegration.abs.gov.au/topics/applying-the-separation-principle>.
- Baker, S.R., Heaton, L.J., McGrath, C., 2023. Evolution and development of methodologies in social and behavioural science research in relation to oral health. *Commun. Dent. Oral Epidemiol.* 51 (1), 46–57.
- Bambach, M.R., Mitchell, R.J., 2015. Estimating the human recovery costs of seriously injured road crash casualties. *Accid. Anal. Prev.* 85, 177–185. <https://doi.org/10.1016/j.aap.2015.09.013>.
- Benavente, M., Knodler, M.A., Rothenberg, H., 2006. Case study assessment of crash data challenges: linking databases for analysis of injury specifics and crash compatibility issues. *Transp. Res. Rec.* 1953 (1), 180–186. <https://doi.org/10.1177/0361198106195300121>.
- Bohensky, M., 2015. Bias in data linkage studies. *Methodol. Develop. Data Linkage* 63–82.
- Bohensky, M.A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D.V., Scott, I., Brand, C. A., 2010. Data linkage: a powerful research tool with potential problems. *BMC Health Serv. Res.* 10 (1), 1–7. <https://doi.org/10.1186/1472-6963-10-346>.
- Bohensky, M.A., Jolley, D., Sundararajan, V., Evans, S., Ibrahim, J., Brand, C., 2011. Development and validation of reporting guidelines for studies involving data linkage. *Aust. N. Z. J. Public Health* 35 (5), 486–489.
- Boufous, S., Williamson, A., 2009. Factors affecting the severity of work related traffic crashes in drivers receiving a worker's compensation claim. *Accid. Anal. Prev.* 41 (3), 467–473. <https://doi.org/10.1016/j.aap.2009.01.015>.
- Boufous, S., Finch, C., Hayen, A., Williamson, A., 2008. The impact of environmental, vehicle and driver characteristics on injury severity in older drivers hospitalized as a result of a traffic crash. *J. Saf. Res.* 39 (1), 65–72.
- Brophy, S., Bandyopadhyay, A., Kennedy, N., Jones, H., Harrington, K., Evans, J., Bellis, M., Rowe, B., Spasic, I., McEnerney, C., & Moore, S. C. (2023). Unlocking data to inform public health policy and practice (NETSCC ID number 133680). Data Science Building, Swansea University, Swansea University, SA2 8PP.
- Brubacher, J.R., Chan, H., Erdelyi, S., 2019. Injury severity in police collision reports correlates poorly with requirement for hospital admission. *J. Transp. Health* 14, 100606. <https://doi.org/10.1016/j.jth.2019.100606>.
- Ceklic, E., Tohira, H., Ball, S., Brown, E., Brink, D., Bailey, P., Finn, J., 2021. Motor vehicle crash characteristics that are predictive of high acuity patients: an analysis of linked ambulance and crash data. *Prehosp. Emerg. Care* 25 (3), 351–360. <https://doi.org/10.1080/10903127.2020.1771487>.
- Cherry, C., Hezaveh, A. M., Noltenius, M., Khattak, A., Merlin, L., Dumbaugh, E., Ragland, D., Sandt, L., 2018. Completing the Picture of Traffic Injuries: Understanding Data Needs and Opportunities for Road Safety (Report No. CSCR-84). Federal Highway Administration. Retrieved from <https://rosap.nhtl.bts.gov/view/dot/62333>.
- Christen, P., Goiser, K., 2007. Quality and complexity measures for data linkage and deduplication. *Quality Measur. Data Mining* 127–151.
- Cirera, E.V.A., Plasencia, A., Ferrando, J., Arribas, P., 2001. Probabilistic linkage of police and emergency department sources of information on motor-vehicle injury cases: a proposal for improvement. *Traffic Inj. Prev.* 2 (3), 229–237. <https://doi.org/10.1080/10286580108902567>.
- Clark, D.E., 2004. Practical introduction to record linkage for injury research. *Inj. Prev.* 10 (3), 186–191.
- Conderino, S., Fung, L., Sedlar, S., Norton, J.M., 2017. Linkage of traffic crash and hospitalization records with limited identifiers for enhanced public health surveillance. *Accid. Anal. Prev.* 101, 117–123. <https://doi.org/10.1016/j.aap.2017.02.011>.
- Couto, A., Amorim, M., Ferreira, S., 2016. Reporting road victims: Assessing and correcting data issues through distinct injury scales. *J. Saf. Res.* 57, 39–45. <https://doi.org/10.1016/j.jsr.2016.03.008>.
- Cryer, P.C., Westrup, S., Cook, A.C., Ashwell, V., Bridger, P., Clarke, C., 2001. Investigation of bias after data linkage of hospital admissions data to police road traffic crash reports. *Inj. Prev.* 7 (3), 234–241.
- Curry, A.E., Pfeiffer, M.R., Carey, M.E., Cook, L.J., 2019. Catalyzing traffic safety advancements via data linkage: development of the New Jersey Safety and Health Outcomes (NJ-SHO) data warehouse. *Traffic Inj. Prev.* 20 (sup2), S151–S155. <https://doi.org/10.1080/15389588.2019.1679552>.
- Dandona, R., Kumar, G.A., Ameer, M.A., Reddy, G.B., Dandona, L., 2008. Under-reporting of road traffic injuries to the police: results from two data sources in urban India. *Inj. Prev.* 14 (6), 360–365. <https://doi.org/10.1136/ip.2008.019638>.
- De Coster, C., Quan, H., Finlayson, A., Gao, M., Halfon, P., Humphries, K.H., Johansen, H., Lix, L.M., Luthi, J.C., Ma, J., Romano, P.S., 2006. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv. Res.* 6, 1–6.
- Dhillon, G., Torkzadeh, G., 2006. Value-focused assessment of information system security in organizations. *Inf. Syst. J.* 16, 293–314. <https://doi.org/10.1111/j.1365-2575.2006.00219.x>.
- Dusetzina, S.B., Tyree, S., Meyer, A.M., Meyer, A., Green, L. and Carpenter, W.R., 2014. Linking data for health services research: a framework and instructional guide.
- Edwards, M., Gutierrez, M., 2023. The incidence burden of unreported pedestrian crashes in Illinois. *Traffic Inj. Prev.* 24 (1), 82–88. <https://doi.org/10.1080/15389588.2022.2143236>.
- Elstad, M., Ahmed, S., Røislien, J., Douiri, A., 2023. Evaluation of the reported data linkage process and associated quality issues for linked routinely collected healthcare data in multimorbidity research: a systematic methodology review. *BMJ Open* 13 (5), e069212.
- Elvik, R., 2010. Improving road safety by better reporting of accidents. *Accid. Anal. Prev.* 42 (6), 1620–1626. <https://doi.org/10.1016/j.aap.2010.03.012>.
- Engin, Z., van Dijk, J., Lan, T., Longley, P.A., Treleaven, P., Batty, M., Penn, A., 2020. Data-driven urban management: mapping the landscape. *J. Urban Manage.* 9 (2), 140–150.
- European Commission. (n.d.). Road safety: Statistics and analysis. Retrieved March 3, 2023, from <https://road-safety.transport.ec.europa.eu/statistics-and-analysis/data-and-analysis/en>.
- Ferreira, S., Falcão, L., Couto, A., Amorim, M., 2015. The quality of the injury severity classification by the police: An important step for a reliable assessment. *Saf. Sci.* 79, 88–93. <https://doi.org/10.1016/j.ssci.2015.05.013>.
- Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.C., Smith, P., Goldstein, H., 2018. GUILD: guidance for information about linking data sets. *J. Public Health* 40 (1), 191–198.
- Gkoulalas-Divanis, A., Vatsalan, D., Karapiperis, D., Kantarcioglu, M., 2021. Modern privacy-preserving record linkage techniques: an overview. *IEEE Trans. Inf. Forensics Secur.* 16, 4966–4987.
- Gonzalez, R.P., Cummings, G., Mulekar, M., Rodning, C.B., 2006. Increased mortality in rural vehicle trauma: identifying contributing factors through data linkage. *J. Trauma* 61 (2), 404–409. <https://doi.org/10.1097/01.ta.0000229816.16305.94>.
- Gopal, D.P., Chetty, U., O'Donnell, P., Gajria, C., Blackadder-Weinstein, J., 2021. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthcare Journal* 8 (1), 40.
- Greenhalgh, T., Peacock, R., 2005. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ* 331 (7524), 1064–1065.
- Gutierrez, H., Mitra, S., Neki, K., Mbugua, L.W., Balasubramanian, R., Winer, M., Bhalla, K., 2022. Comparing estimates of road traffic deaths and non-fatal road traffic injuries in Cambodia. *Inj. Prev.* 28 (4), 340–346.
- Haneef, R., Tijhuis, M., Thiébaud, R., Májek, O., Pristaš, I., Tolonen, H., Galloway, A., 2022. Methodological guidelines to estimate population-based health indicators using linked data and/or machine learning techniques. *Arch. Public Health* 80 (1), 9.
- Harrison, J.E., Watson, A., Vallmuur, K., 2019. A national approach to measuring non-fatal crash outcomes. Austroads Publication No. AP-R599-19, Sydney. Harrison, J.E., Watson, A., Flavel, J., 2023. A national approach to measuring non-fatal crash outcomes stage 2. Austroads Publication no: AP-R694-23, Sydney.
- Harrison, J.E., Weber, S., Jakob, R., Chute, C.G., 2021. ICD-11: an international classification of diseases for the twenty-first century. *BMC Med. Inf. Decis. Making* 21 (6), 1–10.
- Harron, K.L., Doidge, J.C., Knight, H.E., Gilbert, R.E., Goldstein, H., Cromwell, D.A., van der Meulen, J.H., 2017. A guide to evaluating linkage quality for the analysis of linked data. *Int. J. Epidemiol.* 46 (5), 1699–1710. <https://doi.org/10.1093/ije/dyx177>.
- Harron, K., Doidge, J.C., Goldstein, H., 2020. Assessing data linkage quality in cohort studies. *Ann. Hum. Biol.* 47 (2), 218–226.
- Heydari, S., Hickford, A., McIlroy, R., Turner, J., Bachani, A.M., 2019. Road safety in low-income countries: state of knowledge and future directions. *Sustainability* 11 (22), 6249.

- Hong, Y., Zeng, M.L., 2023. International classification of diseases (ICD). KO Knowledge Organizat. 49 (7), 496–528.
- Hossain, A., Sun, X., Islam, S., Alam, S., Hossain, M.M., 2023. Identifying roadway departure crash patterns on rural two-lane highways under different lighting conditions: Association knowledge using data mining approach. *J. Saf. Res.*
- Hosseinzadeh, A., Kluger, R., 2021. Do EMS times associate with injury severity? *Accid. Anal. Prev.* 153, 106053 <https://doi.org/10.1016/j.aap.2021.106053>.
- Hosseinzadeh, A., Karimpour, A., Kluger, R., Orthober, R., 2022. Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records. *J. Saf. Res.* 81, 21–35. <https://doi.org/10.1016/j.jsr.2022.01.003>.
- Hosseinzadeh, Aryan, Data linkage for crash-injury outcome assessment. (2022). Electronic Theses and Dissertations. Paper 3805. <https://doi.org/10.18297/etd/3805>.
- Hu, G., Baker, T., Baker, S.P., 2011. Comparing road traffic mortality rates from police-reported data and death registration data in China. *Bull. World Health Organ.* 89 (1), 41–45. <https://doi.org/10.2471/BLT.10.080317>.
- Imprialou, M., Quddus, M., 2019. Crash data quality for road safety research: Current state and future directions. *Accid. Anal. Prev.* 130, 84–90.
- Janstrup, K.H., Kaplan, S., Hels, T., Lauritsen, J., Prato, C.G., 2016. Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. *Traffic Inj. Prev.* 17 (6), 580–584. <https://doi.org/10.1080/15389588.2015.1128533>.
- Jarvis, S.N., Lowe, P.J., Avery, A., Levene, S., Cormack, R.M., 2000. Children are not goldfish—mark/recapture techniques and their application to injury data. *Inj. Prev.* 6 (1), 46–50. <https://doi.org/10.1136/ip.6.1.46>.
- Jeffrey, S., Stone, D.H., Blamey, A., Clark, D., Cooper, C., Dickson, K., Mackenzie, M., Major, K., 2008. An evaluation of police reporting of road casualties. *Injury Prevent.: J. Int. Soc. Child Adolescent Injury Prevent.* 15 (1), 13–18. <https://doi.org/10.1136/ip.2008.018630>.
- John, A., Dennis, M., Kosnes, L., Gunnell, D., Scourfield, J., Ford, D.V., Lloyd, K., 2014. Suicide Information Database-Cymru: a protocol for a population-based, routinely collected data linkage study to explore risks and patterns of healthcare contact prior to suicide to identify opportunities for intervention. *BMJ Open* 4 (11), e006780.
- Johnson, B.T., Hennessy, E.A., 2019. Systematic reviews and meta-analyses in the health sciences: Best practice methods for research syntheses. *Soc. Sci. Med.* 233, 237–251.
- Kalkman, S., van Delden, J., Banerjee, A., Tyl, B., Mostert, M., van Thiel, G., 2022. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *J. Med. Ethics* 48 (1), 3–13.
- Kamaluddin, N.A., Abd Rahman, M.F., Vårheli, A., 2019. Matching of police and hospital road crash casualty records—a data-linkage study in Malaysia. *Int. J. Inj. Contr. Saf. Promot.* 26 (1), 52–59.
- Kudryavtsev, A.V., Kleshchinov, N., Ermolina, M., Lund, J., Grjibovski, A.M., Nilssen, O., Ytterstad, B., 2013. Road traffic fatalities in Arkhangelsk, Russia in 2005–2010: Reliability of police and healthcare data. *Accid. Anal. Prev.* 53, 46–54. <https://doi.org/10.1016/j.aap.2012.12.022>.
- Langley, J.D., Dow, N., Stephenson, S., Kypri, K., 2003. Missing cyclists. *Inj. Prev.* 9 (4), 376–379.
- Lateef, M.U., 2010. Estimation of fatalities due to road traffic crashes in Karachi, Pakistan, using capture-recapture method. *Asia Pacific J. Public Health* 22 (3), 332–341. <https://doi.org/10.1177/1010539509356808>.
- Lim, D., Randall, S., Robinson, S., Thomas, E., Williamson, J., Chakera, A., Boyd, J., 2022. Unlocking potential within health systems using privacy-preserving record linkage: exploring chronic kidney disease outcomes through linked data modelling. *Appl. Clin. Inf.* 13 (04), 901–909.
- Lombardi, L.R., Pfeiffer, M.R., Metzger, K.B., Myers, R.K., Curry, A.E., 2022. Improving identification of crash injuries: state-wide integration of hospital discharge and crash report data. *Traffic Inj. Prev.* 23 (sup1), S130–S136. <https://doi.org/10.1080/15389588.2022.2083612>.
- Loo, B.P.Y., Wong, S.C., 2012. Under-reporting of road traffic accidents in Hong Kong. *Accid. Anal. Prev.* 49, 374–381. <https://doi.org/10.1016/j.aap.2012.05.019>.
- Lopez, D.G., Rosman, D.L., Jelinek, G.A., Wilkes, G.J., Sprivilis, P.C., 2000. Complementing police road-crash records with trauma registry data—an initial evaluation. *Accid. Anal. Prev.* 32 (6), 771–777. [https://doi.org/10.1016/S0001-4575\(99\)00130-X](https://doi.org/10.1016/S0001-4575(99)00130-X).
- Lujic, S., Finch, C., Boufous, S., Hayen, A., Dunsmuir, W., 2008. How comparable are road traffic crash cases in hospital admissions data and police records? An examination of data linkage rates. *Aust. N. Z. J. Public Health* 32, 28–33. <https://doi.org/10.1111/j.1753-6405.2008.00162.x>.
- Mandacaru, P.M.P., Andrade, A.L., Rocha, M.S., Aguiar, F.P., Nogueira, M.S.M., Girodo, A.M., Pedrosa, A.A.G., Oliveira, V.L.A., Alves, M.M.M., Paixão, L.M.M.M., Malta, D.C., Silva, M.M.A., Morais Neto, O.L., 2017. Qualifying information on deaths and serious injuries caused by road traffic in five Brazilian capitals using record linkage. *Accid. Anal. Prev.* 106, 392–398. <https://doi.org/10.1016/j.aap.2017.06.018>.
- Mars, B., Cornish, R., Heron, J., Boyd, A., Crane, C., Hawton, K., Gunnell, D., 2016. Using data linkage to investigate inconsistent reporting of self-harm and questionnaire non-response. *Arch. Suicide Res.* 20 (2), 113–141.
- Martinez-Garcia, M., Hernández-Lemus, E., 2022. Data integration challenges for machine learning in precision medicine. *Front. Med.* 8, 784455.
- McDonald, G., Davie, G., Langley, J., 2009. Validity of police-reported information on injury severity for those hospitalized from motor vehicle traffic crashes. *Traffic Inj. Prev.* 10 (2), 184–190. <https://doi.org/10.1080/15389580802593699>.
- Milani, J., Kindelberger, J., Bergen, G., Novicki, E. J., Burch, C., Ho, S. M., & West, B. A. (2015, August). Assessment of characteristics of state data linkage systems. (Report No. DOT HS 812 180). Washington, DC: National Highway Traffic Safety Administration, and Atlanta: Centers for Disease Control and Prevention.
- Mitchell, R.J., Cameron, C.M., Bambach, M.R., 2014. Data linkage for injury surveillance and research in Australia: perils, pitfalls and potential. *Aust. N. Z. J. Public Health* 38 (3), 275–280.
- Mitchell, R.J., Senserrick, T., Bambach, M.R., Mattos, G., 2015. Comparison of novice and full-licensed driver common crash types in New South Wales, Australia, 2001–2011. *Accid. Anal. Prev.* 81, 204–210. <https://doi.org/10.1016/j.aap.2015.04.039>.
- Mock, C., Lormand, J. D., Goosen, J., Joshipura, M., Peden, M., & Guidelines for Essential Trauma Care. (2005). Guidelines for essential trauma care. Geneva: World Health Organization.
- Moore, M.A., 1997. Comparison of young and adult driver crashes in Alaska using linked traffic crash and hospital data. *Alaska Med.* 39 (4), 95–102.
- Moore, M., Conrick, K.M., Fuentes, M., Rowhani-Rahbar, A., Graves, J.M., Patil, D., Vavilala, M.S., 2019. Research on injury disparities: a scoping review. *Health Equity* 3 (1), 504–511.
- Moore, H.C., Guiver, T., Woollacott, A., De Klerk, N., Gidding, H.F., 2016. Establishing a process for conducting cross-jurisdictional record linkage in Australia. *Aust. N. Z. J. Public Health* 40 (2), 159–164.
- Morgan, K., Page, N., Brown, R., et al., 2020. Sources of potential bias when combining routine data linkage and a national survey of secondary school-aged children: A record linkage study. *BMC Med. Res. Method.* 20 (1), 178. <https://doi.org/10.1186/s12874-020-01064-1>.
- Morrison, A., Stone, D.H., 2000. Capture-recapture: a useful methodological tool for counting traffic related injuries? *Inj. Prev.* 6 (4), 299–304. <https://doi.org/10.1136/ip.6.4.299>.
- Mourby, M.J., Doidge, J., Jones, K.H., Aidinlis, S., Smith, H., Bell, J., Kaye, J., 2019. Health data linkage for UK public interest research: key obstacles and solutions. *Int. J. Popul. Data Sci.* 4 (1).
- Nair, M.M., Tyagi, A.K., 2021. Privacy: History, statistics, policy, laws, preservation and threat analysis. *J. Inf. Assurance Security* 16 (1).
- Paez, A., 2017. Gray literature: An important resource in systematic reviews. *J. Evid. Based Med.* 10 (3), 233–240.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, n71.
- Paixão, L.M., Gontijo, E.D., Mingoti, S.A., Costa, D.A.D.S., Friche, A.A.D.L., Caiassa, W.T., 2015a. Urban road traffic deaths: data linkage and identification of high-risk population sub-groups. *Cad. Saude Publica.* 31, 92–106.
- Paixão, L.M., Gontijo, E.D., Drumond, E.deF., Friche, A.A., Caiassa, W.T., 2015b. Traffic accidents in Belo Horizonte: the view from three different sources. *Rev. Brasil. Epidemiol.* = *Brazil. J. Epidemiol.* 18 (1), 108–122. <https://doi.org/10.1590/1980-5497201500010009>, 2008 to 2010.
- Palamuthusingam, D., Johnson, D.W., Hawley, C., Pascoe, E., Fahim, M., 2019. Health data linkage research in Australia remains challenging. *Intern. Med. J.* 49 (4), 539–544.
- Prince, A.E., Schwarcz, D., 2019. Proxy discrimination in the age of artificial intelligence and big data. *Iowa J. Rev.* 105, 1257.
- Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. *Health Inf. Sci. Syst.* 2, 1–10.
- Ranbaduge, T., Vatsalan, D., Ding, M., 2023. Privacy-preserving deep learning based record linkage. *IEEE Trans. Knowl. Data Eng.*
- Reurings, M.C., Stipdonk, H.L., 2011. Estimating the number of serious road injuries in the Netherlands. *Ann. Epidemiol.* 21 (9), 648–653. <https://doi.org/10.1016/j.annepidem.2011.05.007>.
- Rosman, D.L., Knuiman, M.W., 1994. A comparison of hospital and police road injury data. *Accid. Anal. Prev.* 26 (2), 215–222. [https://doi.org/10.1016/0001-4575\(94\)90091-4](https://doi.org/10.1016/0001-4575(94)90091-4).
- Salameh, J.P., Bossuyt, P.M., McGrath, T.A., Thombs, B.D., Hyde, C.J., Macaskill, P., McInnes, M.D., 2020. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *bmj* 370.
- Santiago, J.D., Colón Jordán, H., Valdés, D., 2020. Record linkage of crashes with injuries and medical cost in Puerto Rico. *Transp. Res. Rec.* 2674 (10), 739–748. <https://doi.org/10.1177/0361198120935439>.
- Santoro, S., Estay, H., Avci, A.H., Pugliese, L., Ruby-Figueroa, R., Garcia, A., Curcio, E., 2021. Membrane technology for a sustainable copper mining industry: the Chilean paradigm. *Clean. Eng. Technol.* 2, 100091.
- Sasser, S.M., Hunt, R.C., Faul, M., Sugerman, D., Pearson, W.S., Dulski, T., 2012. Guidelines for field triage of injured patients: recommendations of the National Expert Panel on Field Triage, 2011. *Morb. Mortal. Wkly Rep.* 61 (1), 1–20.
- Schneider, R.J., Willman, J., Hargarten, S., 2023. Linking police and EMS records: an approach to strengthen bicyclist injury reporting. *Transp. Res. Rec.* <https://doi.org/10.1177/03611981221151073>.
- Sciortino, S., Vassar, M., Radetsky, M., Knudson, M.M., 2005. San Francisco pedestrian injury surveillance: mapping, under-reporting, and injury severity in police and hospital records. *Accid. Anal. Prev.* 37 (6), 1102–1113. <https://doi.org/10.1016/j.aap.2005.06.010>.
- Shlomo, N., 2019. Overview of data linkage methods for policy design and evaluation. In: Crato, N., Paruolo, P. (Eds.), *Data-Driven Policy Impact Evaluation*. Springer, pp. 43–65. https://doi.org/10.1007/978-3-319-78461-8_4.
- Short, J., Caulfield, B., 2016. Record linkage for road traffic injuries in Ireland using police hospital and injury claims data. *J. Saf. Res.* 58, 1–14. <https://doi.org/10.1016/j.jsr.2016.05.002>.

- Singleton, M., Qin, H., Luan, J., 2004. Factors associated with higher levels of injury severity in occupants of motor vehicles that were severely damaged in traffic crashes in Kentucky, 2000–2001. *Traffic Inj. Prev.* 5 (2), 144–150.
- Smith, M., Flack, F., 2021. Data linkage in Australia: the first 50 years. *Int. J. Environ. Res. Public Health* 18 (21), 11339. <https://doi.org/10.3390/ijerph182111339>.
- Soltani, A., Askari, S., 2014. Analysis of intra-urban traffic accidents using spatiotemporal visualization techniques. *Transp. Telecommun.* 15 (3), 227–232. <https://doi.org/10.2478/ttj-2014-0020>.
- Soltani, A., Askari, S., 2017. Exploring spatial autocorrelation of traffic crashes based on severity. *Injury* 48 (3), 637–647. <https://doi.org/10.1016/j.injury.2017.01.032>.
- Soltani, S., Schwarcz, L., Morris, D., Plevin, R., Dicker, R., Juillard, C., Wier, M., 2022. What is counted counts: An innovative linkage of police, hospital, and spatial data for transportation injury prevention. *J. Saf. Res.* 83, 35–44. <https://doi.org/10.1016/j.jsr.2022.08.002>.
- Soneson, E., Das, S., Burn, A.M., Van Melle, M., Anderson, J.K., Fazel, M., Moore, A., 2023. Leveraging administrative data to better understand and address child maltreatment: a scoping review of data linkage studies. *Child Maltreat.* 28 (1), 176–195.
- Stutts, J.C., Hunter, W.W., 1999. Motor vehicle and roadway factors in pedestrian and bicyclist injuries: an examination based on emergency department data. *Accid. Anal. Prev.* 31 (5), 505–514. [https://doi.org/10.1016/S0001-4575\(99\)00007-X](https://doi.org/10.1016/S0001-4575(99)00007-X).
- Suriyawongpaisal, P., Kanchanasut, S., 2003. Road traffic injuries in Thailand: trends, selected underlying determinants and status of intervention. *Inj. Control Saf. Promot.* 10 (1–2), 95–104.
- Tainter, F., Fitzpatrick, C., Gazillo, J., Riessman, R., Knodler Jr, M., 2020. Using a novel data linkage approach to investigate potential reductions in motor vehicle crash severity - An evaluation of strategic highway safety plan emphasis areas. *J. Saf. Res.* 74, 9–15. <https://doi.org/10.1016/j.jsr.2020.04.012>.
- Tarko, A., Azam, M.S., 2011. Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data. *Accid. Anal. Prev.* 43 (5), 1689–1695. <https://doi.org/10.1016/j.aap.2011.03.027>.
- Tercero, F., Andersson, R., 2004. Measuring transport injuries in a developing country: an application of the capture–recapture method. *Crash Anal. Prevent.* 36 (1), 13–20.
- Tikkinen-Piri, C., Rohunen, A., Markkula, J., 2018. EU General Data protection regulation: changes and implications for personal data collecting companies. *Comput. Law Secur. Rev.* 34 (1), 134–153.
- Tin Tin, S., Woodward, A., Ameratunga, S., 2013. Completeness and accuracy of crash outcome data in a cohort of cyclists: a validation study. *BMC Public Health* 13 (1), 1–10. <https://doi.org/10.1186/1471-2458-13-420>.
- Torbaghan, M.E., Sasidharan, M., Reardon, L., Muchanga-Hvelplund, L.C., 2022. Understanding the potential of emerging digital technologies for improving road safety. *Accid. Anal. Prev.* 166, 106543.
- Tsui, K.L., So, F.L., Sze, N.N., Wong, S.C., Leung, T.F., 2009. Misclassification of injury severity among road casualties in police reports. *Accid. Anal. Prev.* 41 (1), 84–89. <https://doi.org/10.1016/j.aap.2008.09.005>.
- Tulu, G.S., Washington, S., King, M., Haque, M., 2013. In: Why Are Pedestrian Crashes so Different in Developing Countries? A Review of Relevant Factors in Relation to Their Impact in Ethiopia. *Australasian Transport Research Forum*, pp. 1–18.
- Van Eck, N. J., & Waltman, L. (2018). VOSviewer Manual https://www.vosviewer.com/documentation.Manual_VOSviewer_1, 6.
- Vassar, M., Atakpo, P., Kash, M.J., 2016. Manual search approaches used by systematic reviewers in dermatology. *J. Med. Library Assoc.: JMLA* 104 (4), 302.
- Vatsalan, D., Sehili, Z., Christen, P., Rahm, E., 2017. Privacy-preserving record linkage for big data: Current approaches and research challenges. *Handbook of Big Data Technologies* 851–895.
- Victor, T.W., Mera, R.M., 2001. Record linkage of health care insurance claims. *J. Am. Med. Inform. Assoc.* 8 (3), 281–288. <https://doi.org/10.1136/jamia.2001.0080281>.
- Wang, L., Ning, P., Yin, P., Cheng, P., Schwebel, D.C., Liu, J., Hu, G., 2019. Road traffic mortality in China: analysis of national surveillance data from 2006 to 2016. *Lancet Public Health* 4 (5), e245–e255. [https://doi.org/10.1016/S2468-2667\(19\)30057-X](https://doi.org/10.1016/S2468-2667(19)30057-X).
- Watson, A., Watson, B., Vallmuur, K., 2013. How accurate is the identification of serious traffic injuries by Police? The concordance between Police and hospital reported traffic injuries. In *Proceedings of the 2013 Australasian Road Safety Research, Policing and Education Conference* (pp. 1–10). Australasian College of Road Safety (ACRS).
- Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. *Accid. Anal. Prev.* 83, 18–25. <https://doi.org/10.1016/j.aap.2015.06.011>.
- Weijermars, W., Bos, N., Stipdonk, H.L., 2016. Serious road injuries in the Netherlands dissected. *Traffic Inj. Prev.* 17 (1), 73–79. <https://doi.org/10.1080/15389588.2015.1042577>.
- Wilson, S.J., Begg, D.J., Samaranyaka, A., 2012. Validity of using linked hospital and police traffic crash records to analyse motorcycle injury crash characteristics. *Accid. Anal. Prev.* 49, 30–35. <https://doi.org/10.1016/j.aap.2011.03.007>.
- World Bank. (2023). The World by Income and Region. World Bank Data Topics. Retrieved from <https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html>, accessed on: 19/04/2023.