Piecing the puzzle together: enhancing the quality of road trauma surveillance through linkage of police and health data

Angela Watson

BA (Hons Psych), MHS

A thesis submitted as fulfilment for the degree of Doctor of Philosophy
Queensland University of Technology
Centre for Accident Research and Road Safety - Queensland
School of Psychology and Counselling
Brisbane, Australia

2014

Keywords

Data linkage, road crash injuries, data quality, hospitalisations, road safety, crash data, hospital data, ambulance data

Acknowledgements

I would like to firstly thank my supervisors Professor Barry Watson[1] and Associate Professor Kirsten Vallmuur. I want to thank Barry for agreeing to be my principal supervisor despite being one of the busiest men alive. I don't believe I would have ever done a PhD if you didn't agree to be part of it. You have given me so much support, advice, and encouragement both with the PhD and beyond. Thanks for tolerating my 'data nerdiness' and always keeping an eye on the 'big picture'. Kirsten, thank you for 'getting it'….for knowing that data is exciting and that data nerds are the 'new black'. You didn't really know me when we approached you to be my associate supervisor, but you took a chance and I really appreciate it. You have been an outstanding mentor and I couldn't have done this without you.

To the data providers for my PhD, thank you to the Department of Transport and Main Roads, Queensland Health, Queensland Ambulance Service, and the National Coronial Information System. I would particularly like to thank Dr. Nerida Leal of TMR for her undying efforts to get the MOU together and helping make all of this possible. I would also like to thank Ben Wilkinson (QHAPDC), Dr. Ruth Barker (QISU), Emma Bosley and Jamie Quinn (QAS), Jean Sloan (EDIS), and Joanna Cotsonis (NCIS) for helping me with applications and other advice relating to the data collections and for helping me obtain the data I needed for my PhD. I would also like to particularly thank Catherine Taylor of the data linkage unit in Queensland Health for her tireless work in completing my linkage and addressing my potentially annoying queries (including "are we there yet, are we there yet" emails). On a similar note, I would like to thank all of the participants in the interviews, the data custodians, data users, and data linkage experts. Thank you for taking to time to talk to me and for providing such rich and thoughtful responses. Thanks also go to the NHMRC for funding my PhD. Thanks for the generous support. I would also like to acknowledge the feedback I received from my Final Seminar review panel, Professor Narelle Haworth, Professor James Harrison, and Mr. Mike Stapleton. I would also like to acknowledge my examiners for their feedback. Thank you all for your time and expertise.

I would also like to thank Emeritus Professor Mary Sheehan, Ms. Cynthia Schonfeld, and Barry (again) for giving me the opportunity many years ago to work at CARRS-Q and for always providing me with boundless opportunities and support. CARRS-Q is a wonderful place to work and study and that is due in a large part to Mary and Barry.

Thank you to the CARRS-Q family, past and present. If I mentioned you all, I would go over the word limit, but you know who you are. Thanks for the drinks, the laughs, the 'lunch table', the checking in on me, the advice, and so much more. CARRS-Q has

[1] No relation

attracted some amazing people over the years, many of which have become lifelong dear friends, cheers!

I would also like to thank my oldest and dearest friend Robyn; you are like a sister to me. You have always been there for me in thick and thin (and boy have there been a lot of both). We have shared all of the important milestones together. We are family and I feel so privileged to have such a wonderful person be my friend for over 25 years. Here’s to another 25! Another dear friend I would like to thank is Hollie, my former roommate and general partner in crime. We just clicked right from the start. Some of the most fun nights of my life have been spent with you. Thank you for being there from the beginning of this PhD and for always listening. You are always encouraging and you are one of the awesomest (that’s a word!) people I have ever had the privilege of knowing. I look forward to us both finishing and having our ‘big night in’; we should try to make it a Tuesday.

To my Mum and Dad, for loving me and for always encouraging me to ‘do my best’ and seek out knowledge and understanding. Dad, thanks for listening to me for countless hours about my PhD and chatting about life, the universe, and everything (and sport). You have always been my rock and during this PhD has been no exception. Thanks to my brother Marty for our weekly chats again about life, the universe, and everything (and sport and movies and TV shows…). I can’t imagine a better big brother. Thank you to Dad and Marty for providing me with a strong sense of family and for teaching me to keep a sense of humour, always.

Last but certainly not least to my partner Dale, where do I begin? A friend from the start and a partner for life. Thank you for listening, advising, helping, distracting, and understanding. I’d like to promise that I won’t bore you with endless data discussions once the PhD is submitted, but we still do what we do for a living, so I guess that’s a promise I can’t make. I look forward to being Drs together and carving out our own little place in the world. I can’t think of anyone I’d rather do that with.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made.

Signed: QUT Verified Signature

Date: 5/9/14

Abstract

Injuries resulting from road crashes are a significant public health problem world-wide. In order to reduce the burden of road crash injuries, there is a need to better understand the nature of and contributing circumstances to road crashes and the resulting injuries. The National Road Safety Strategy 2011-2020 (Australian Transport Council, 2011) highlights that a key aspect in reducing the burden of road trauma is the availability of comprehensive data on the issue. The use of data is essential to undertake more in-depth epidemiologic studies of risk as well as effective evaluation of road safety interventions and programs.

Police reported crash data are the primary source of crash information in most jurisdictions. However, the definition of serious injury within police-reported data is not consistent across jurisdictions and may not be accurate, which could lead to misleading estimates of the impact and cost of crashes. In light of the National Road Safety Strategy (Australian Transport Council, 2011) emphasising serious injury reduction targets as well as fatalities, which has not previously been the case, there is a need to assess the current serious injury definition in each jurisdiction. It is possible that linking police-reported crash data with health-related data may provide a more accurate measure of injury severity. Also, data linkage can result in other improvements to data quality by including road crash injuries not reported to police and increasing the accuracy of existing data through the detection and correction of errors.

It has not, however, been established whether data linkage of this nature is feasible in Queensland. It is also necessary to establish whether linked data provide advantage over non-linked data, both qualitatively and quantitatively. The overall goal of the program of research is to examine the extent to which data linkage provides a more comprehensive picture of road crashes and resulting injuries in Queensland. In doing so, this research will have important implications for data linkage and the measurement of serious injuries in other Australian and international jurisdictions.

Study 1a of the research program involved a review of legislation and documentation relating to the following road crash injury data collections: Queensland Road Crash Database (QRCD); Queensland Hospital Admitted Patients Data Collection (QHAPDC); Queensland Injury Surveillance Unit (QISU); Emergency Department Information System (EDIS); electronic Ambulance Recording Form (eARF); and the National Coronial Information System (NCIS). This information was supplemented by interviews with relevant data custodians. The study explored the characteristics of the data collections to provide some insights into the quality of these data collections in terms of completeness, consistency, validity, representativeness, timeliness, and accessibility. The results indicate that there are limitations associated issues with the police collected Queensland Road Crash Database (QRCD) in terms of severity definitions and under-reporting. In this regard, the other data collections explored in this study appear to offer potential to add information to the police data in terms of both scope and content. These data collections include cases that may not be reported to police that should have been as

well as including variable fields that may provide more reliable information about other factors of importance including injury nature and severity.

Study 1b involved a qualitative analysis of semi-structured interviews with data custodians, data users, and data linkage experts. It explored issues relating to data quality characteristics, including: relevance, completeness, and consistency. It also examined the perceptions of the potential benefits and barriers of using data linkage for enhancing road safety monitoring, planning, and evaluation. The results confirmed concerns about the police collected Queensland Road Crash Database (QRCD), which is relied on for reporting and research in road safety, in terms of severity definitions and under-reporting. The results also indicated that there are many perceived benefits of data linkage including efficiency, increased sample sizes, and the ability to conduct research on issues that would not be possible with only one data collection. Specifically, it was suggested that the major potential benefit of data linkage for road safety research would be the ability to gain a more complete picture of both the circumstances and outcomes relating to road crash injury. There were also some barriers to data linkage highlighted relating to lack of resourcing, skills, and information, as well as potential reluctance among the relevant custodians to share the data required for linkage to occur. Overall, however, most participants were keen to see linkage trialled with road crash injury data in Queensland, in order to better assess the potential benefits it offers.

Study 2 involved the secondary data analysis of the six data collections (i.e., QRCD, QHAPDC, QISU, EDIS, eARF, NCIS) which include road crash injury information. It included analyses regarding the quality of the data collections in terms of completeness of variables, consistency, validity of coding, and representativeness. It also examines these issues specifically in terms of injury severity coding. The results indicated that there are limitations associated with the police collected Queensland Road Crash Database (QRCD), in terms of the broadness of the severity definitions and potential under-reporting. Also, the under-reporting, particularly for some road user groups, is problematic for road safety investigation, intervention development, and evaluation and could impact on the allocation of resources. The results suggest that a more precise measure of serious injury would be preferred over current practice as it is more closely related to threat to life and therefore more directly corresponding to the outcomes being measured when cost and impact is determined. Unfortunately, due to the large amount of missing information in police data, and the questionable accuracy of what is there, relying on police data to determine the prevalence and nature of serious injury crashes could be misleading. The inclusion of other data sources, such as hospital data, in the determination of serious injury crash impact has the potential to address the shortcomings of current approaches.

Study 3 involved the secondary data analysis of linked data from five road crash injury data sources (i.e., QRCD, QHAPDC, QISU, EDIS, eARF). The fatality only data (NCIS) was not included in this study since the focus was on serious non-fatal injury. This study included analyses relating to linkage rates, discordance, validity, and profiles of different combinations of linked data sources. It specifically examined the potential for linked data to enhance the quantification of serious injury and explores issues such as under-reporting of crashes to police. The validity analysis in this study demonstrated that using the police defined measure for the counting of serious injuries results in an inaccurate, or at least incomplete, picture of the serious road crash injury problem.

Study 3 showed that a benefit of using linked data, that has previously not been explored, is the potential for obtaining additional information about cases in the police data (i.e., QRCD) from other data sources. More particularly, this study examined linkage rates of police-reported cases to hospital data collections (with police-reported road crash injuries as the denominator), rather than just focussing on the discordance (or under-reporting) in the police data (with the hospital data as the denominator). This study showed that for more than half of all police-reported cases and around 80% of 'hospitalised' cases, linking with hospital data provides important information about the nature and severity of road crash injuries. It also showed that, while there are some biases in using linked police-reported cases as opposed to all police-reported cases, the profiles are very similar. The implication of this is that, while not all QRCD cases would have injury nature or severity information added, a significant subset of cases would. It could also be argued that this subset would include the most serious non-fatal cases and therefore be most useful in the reporting of serious injury. At the very least, data linkage would allow for the confirmation of the status of injured persons (i.e., attended or admitted to hospital), which would still be an improvement to the current practice.

Another implication of this study was to confirm that there are a number of road crash injuries that are not reported to police as shown in studies elsewhere in Australia (Boufous, Finch, Hayen, & Williamson, 2008; Rosman & Knuiman, 1994) and overseas (Alsop & Langley, 2001; Amoros, Martin, & Laumon, 2006; Langley, Dow, Stephenson, & Kypri, 2003). It has also confirmed the pattern of under-reporting found elsewhere in terms of bias towards certain types of road users (i.e., cyclists and motorcyclists). This bias could greatly impact on road safety research and policy. As discussed elsewhere, an accurate representation of the prevalence of road crash injuries is essential for: prioritising funding and resources; targeting road safety interventions into areas of higher risk; and calculating the cost of road crash injuries in order to estimate the burden of road crash injuries.

An additional implication of Study 3 relates to the validity of the health data sources in identifying road crash injuries. Combined with the results of Study 2, there are some limitations relating to the reliable identification of relevant cases. More particularly, the results suggest that the current method for selecting road crash injury cases could lead to an overestimation of road crash cases. In addition, it was shown that the classification of road users, particularly for some data collections (i.e., EDIS and eARF) was also problematic. Specifically, it was found that motorcyclists and cyclists may be easier to identify in text, suggesting that some of the bias in under-reporting may be somewhat exaggerated. As a result, it is possible that any estimates of under-reporting of crashes to police, both overall and for particular road user groups, may be over-estimated. This needs to be taken into account in future research using these health data sources and when developing strategies to enhance reporting practices.

It is important to note that a considerable amount of time (approximately 2 years) was required to undertake the data linkage informing Study 3, since this impacts on the conclusions drawn about feasibility of the approach. Also, the time and effort required to facilitate the data linkage needs to be weighed against the benefits. While it did take a considerable time to gain approval and for the data linkage to be completed, many of these issues were due to this being the first study of its kind in Queensland. Now that agreements are in place and the method has been established, it would be arguably easier and less time consuming to conduct a similar data linkage in the future. However, it still may not be feasible to conduct linkage frequently or at least often enough to be part of annual reporting practices, as some aspects of the time required would still apply in the future (e.g., ethics, custodian approval, manual reviews).

This program of research has demonstrated that data linkage is possible in the Queensland context and that there are likely to be benefits for road safety research and policy making arising from conducting periodic linkage. It has shown how data linkage can be used to highlight issues of data quality particularly in relation to defining serious injury and the under-reporting of road crash injuries to police. In addition, it has been shown that by linking other data sources with QRCD, improvements to reporting and the classification of serious injury can be achieved. Specifically, QRCD could be linked to QHAPDC to confirm the hospitalisation status of a case, AIS and SRR could be mapped to QRCD cases using hospital data to provide a more precise and/or objective measure of injury severity, and adjustments could be made to reporting on the basis of cases not captured in QRCD to better represent certain groups such as cyclists and motorcyclists.

While the program of research has demonstrated the potential of data linkage for enhancing our knowledge of road crash injury, some caution is needed in assuming that the health data collections include all relevant cases and that these cases are always accurately identified. Further research on this issue is required, including the refinement of the methods used to identify cases and classify road users in these data. It is also possible that data linkage in the future could restrict the data collections linked with QRCD to those that are relevant to the purpose of use and have the most accurate information. For example, the current program of research suggests that linking the QRCD with the QHAPDC and EDIS should be given the highest priority in the future, particularly in terms of better quantifying serious injury outcomes.

Overall, the program of research has shown how data linkage could be utilised (with refinements appropriate to the context) in other jurisdictions. It has also demonstrated how it could improve our understanding of the road safety problem, particularly in relation to the scale and nature of serious injury. Even if linkage was not performed routinely, further research could be conducted to develop adjustments based on linked data, which could then be applied routinely to current reporting, for a more accurate representation of the road trauma problem.

Table of Contents

Listing of Tables

Table of Figures

List of Acronyms

A&E	Accident & Emergency Department
ABS	Australian Bureau of Statistics
ACT	Australian Capital Territory
AIHW	Australian Institute of Health and Welfare
AIS	Abbreviated Injury Scale
ATC	Australian Transport Council
BAC	Blood Alcohol Content
BITRE	Bureau of Infrastructure, Transport and Regional Economics
CDL	Centre for Data Linkage
CHeReL	Centre for Health Record Linkage
CODES	Crash Outcome Data Evaluation System
Core MDS	Core Minimum Data Set
Core ODS	Core Optional Data Set
DLK	Data linkage key
DLU	Data Linkage Unit
eARF	Electronic Ambulance Reporting Form
ED	Emergency Department
EDIS	Emergency Department Information System
ICD	International Classification of Diseases
ICD-10-AM	International Classification of Diseases, 10th edition, Australian Modification
ICISS	International Classification of Diseases–based Injury Severity Score
NCIS	National Coronial Information System
NCRIS	National Collaborative Research Infrastructure Strategy
NHTSA	National Highway Traffic Safety Administration
NSW	New South Wales
OECD	Organisation for Economic Co-operation and Development
PHRN	Population Health Research Network
QAS	Queensland Ambulance Service
QH	Queensland Health

QHAPDC	Queensland Hospital Admitted Patients Data Collection
QISU	Queensland Injury Surveillance Unit
QLD	Queensland
QPS	Queensland Police Service
QRCD	Queensland Road Crash Database
SA	South Australia
SAIL Databank	Secure Anonymised Information Linkage Databank
SA-NT Link	South Australia and Northern Territory Link
SRR	Survival Risk Ratio
TMR	Queensland Department of Transport and Main Roads
WA	Western Australia
WADLS	Western Australian Data Linkage System
WHO	World Health Organisation
VDL	Victorian Data Linkage
VIC	Victoria

Chapter One: Introduction

1.1 Introductory Comments

Injuries resulting from road crashes are a significant public health problem world-wide (WHO, 2004). It is predicted, that unless substantial gains are made in the prevention of crashes, these injuries will become the third ranked global burden of disease and injury by 2020. In Australia, approximately 1,400 people are killed on our roads each year. On average, the economic cost of a single road related fatality is $2.7 million, with a hospitalisation injury costing $266,000 per individual (BITRE, 2010). In order to reduce the burden of road crash injuries, there is a need to fully understand the nature and contributing circumstances of crashes and the resulting injuries. The National Road Safety Strategy 2011-2020 (Australian Transport Council, 2011) outlines plans to reduce the burden of road trauma via improvements and interventions relating to safe roads, safe speeds, safe vehicles, and safe people. It also highlights that a key aspect in achieving these goals is the availability of comprehensive data on the issue. The use of data is essential so that more in-depth epidemiologic studies of risk can be conducted as well as effective evaluation of road safety interventions and programs.

There are a variety of data sources in which road crash-related incidents and resulting injuries are recorded, which are collected for a defined purpose. These include police reports, transport safety databases, emergency department data, hospital morbidity data and mortality data. However, as these data are collected for specific purposes, each of these data sources suffers from some limitations when seeking to gain a complete picture of the problem. It is generally considered that no single data source is sufficient to examine the issue effectively and as a result, there is increasing interest in data linkage as a possible solution to enable a more complete understanding of the issues surrounding transport incidents and the injuries resulting from them. The Queensland Trauma Plan states that:

> "Integrating the existing information will result in a more comprehensive characterisation and monitoring of the public health problem of injury and create a valid and balanced picture on which appropriate policy development and program implementation can be based." (Queensland Trauma Plan, 2006, p.38)

However, each agency and jurisdiction has different data systems with unique considerations for linkage and use. If the ultimate aim is to create an integrated national data linkage system as researchers in the area suggest (Holman et al., 2008; Turner, 2008) then it is important to understand the nature of each jurisdiction's information systems and data linkage capabilities. Given the lack of standardisation of data sources, legislation, and data linkage progress, work needs to first be undertaken at an individual jurisdiction level before informing a national (and potentially international) approach.

1.2 Rationale for the Research

Police reported crash data are the primary source of crash information in most jurisdictions (International Traffic Safety Data and Analysis Group (IRTAD), 2011). Over the years there have been significant reductions in fatalities in Australia (The

Parliament of Victoria Road Safety Committee, 2014) as there has been in many other highly motorised countries (International Traffic Safety Data and Analysis Group (IRTAD), 2011). However, there has been less of a reduction (and in some cases an increase) in the number of serious non-fatal road crash injuries in many of these jurisdictions. This in combination with the substantial burden of serious non-fatal road crash injuries has meant that nationally and internationally, the focus in road safety has shifted towards a greater understanding of road crash serious injuries in addition to fatalities (International Traffic Safety Data and Analysis Group (IRTAD), 2011; The Parliament of Victoria Road Safety Committee, 2014). Unfortunately, however, the definition of serious injury within police-reported data is not consistent across jurisdictions and may not be accurately operationalised, which could lead to misleading estimates of the impact and cost of crashes. Furthermore, the current National Road Safety Strategy (Australian Transport Council, 2011) features a strong emphasis on serious injuries, as well as fatalities, which has not previously been the case. Specifically, it includes the setting of a 30% reduction target for serious injuries during the life of the strategy. Together, these developments highlight the need to assess the current serious injury definition in each jurisdiction. It is possible that linking police-reported crash data with health-related data may provide:

- a more accurate measure of severity of injury; and
- a more accurate estimate of the cost of crashes.

Also, data linkage can result in other improvements to data quality by:

- including road crash injuries not reported to police;
- including more information about crashes and injuries of interest to road safety researchers and policy makers; and
- increasing the accuracy of existing data through the detection and correction of errors.

A report by Austroads (1997) suggests that investment in linked data systems for road safety would greatly increase the value of data sets by allowing the use of data for a wider range of purposes. It is also suggested that data linkage will lead to more efficient day-to-day operations, easier access to data, and a greater ability to effectively evaluate road safety policy. It has not however been established whether data linkage of this nature is feasible in Queensland. It is also necessary to establish whether linked data provide advantage over non-linked data, both qualitatively and quantitatively.

1.3 Defining Road Crash Injury

Throughout this thesis, the injuries of interest will be referred to as road crash injuries. A road crash injury is defined according to what is considered a reportable road crash in the Queensland Road Crash Database, which is as follows:

> *"a crash which resulted from the movement of at least one road vehicle on a public road and involving death or injury to any person*."

In some of the other data collections that include road crash injuries, these cases are identified as 'traffic injuries'. In the literature also, this term is often used particularly when the research involved the use of health data collections or coding. When appropriate to the data or the research literature reviewed in this thesis the term 'traffic injury' will be used, but this term should be treated as synonymous with 'road crash injuries'.

1.4 Research Aims

The overall goal of this program of research is to provide a more comprehensive picture of road crashes and the resulting injuries in Queensland and to assess data linkage possibilities for road crash injury data. It is expected that this research will have both national and international implications.

More specifically, the program of research aims to:

1. Scope existing data sets relating to road crash incidents and injury in order to assess the data quality characteristics of these data sets.

2. Determine the linkage opportunities to enhance the value of the relevant data collections in terms of road safety investigation, intervention development and evaluation.

3. Develop a possible linkage/matching methodology appropriate for existing road crash injury data sets in Queensland.

4. Provide a more comprehensive assessment and profile of road crash injuries in Queensland, including the nature and contributing circumstances of the incidents using both linked and non-linked data sources.

5. Assess the discordance and concordance of the different road crash injury data sources.

6. Assess the feasibility of conducting data linkage with road crash injury data collections for road safety investigation, intervention development, and evaluation.

7. Assess whether linked data provide qualitative and quantitative advantage over unlinked data, both overall and for specific road user groups.

1.5 Demarcation of Scope

This thesis is examining road crash injuries as per the definition described previously (Section 1.3). While it is acknowledged that transport injuries occurring off-road are a significant burden, the main focus of this research is to examine the use of data linkage in relation to official reporting of crash injuries, which in Queensland and most other jurisdictions is restricted to those injuries that occur on a public road. Therefore, it was considered that off-road transport injuries were out of scope for this research program. Despite the fact that up to 40% of all road crashes reported to police do not involve an injury (property damage only), these incidents were also considered out of scope as the focus of this program of research was on identifying strategies to improve the quality of road crash injury reporting, particularly serious injuries.

The next issue relates to the data linkage itself. It was not considered part of the research program to compare data linkage methods or models. While some aspects of best-practice approaches for linkage will be discussed, it was beyond the scope of this research program to use multiple linkage methods and test the feasibility via comparison. The focus of this research was to examine the feasibility of conducting data linkage within the current methods and models available in Queensland. As the researcher was not able to directly conduct the data linkage due to legislative restraints and thus did not have the authority to insist on any methodology outside of the current practice, any comparisons of methods was deemed beyond the scope of the research program. Similarly, it was beyond the scope of the research program to assess data linkage software and/or infrastructure.

Data linkage can be conducted for a range of reasons including: linkage with pre-cursor data to examine the predictors of crashes (e.g., traffic offence histories) and linkage with population data to assess incidence rates. While these are interesting applications of data linkage and are important issues relating to the prevention of road trauma, the main focus of this research was on the reporting of crash and injury incidence as they relate to the National Road Safety Strategy and to give an indication of the crash event itself and the injury outcomes, not to examine the antecedents of these crashes or profile other characteristics that lead to individuals being involved in crashes.

It was also beyond the scope of this PhD to conduct a cost-benefit analysis for the conduct of data linkage for road crash injuries in Queensland. As this is the first research of its kind in Queensland, any cost calculations would not be representative of the cost of conducting data linkage in road safety in the future. Also, while it is the aim of this research to examine potential improvements to data quality that may be seen as benefits, it would require more detailed and specific work to quantify all the benefits in terms of monetary value.

Finally, it is beyond of the scope of this thesis to compare the validity of different injury severity classification systems (e.g., Abbreviated Injury Scale, Survival Risk Ratios). While they are discussed in relation to what is possible from the data that is provided, it is not part of this thesis to select a preferred system.

1.6 Structure of the Research and Outline of Thesis

The studies and chapters in this thesis are related to each other, addressing the two key and interrelated themes of data quality and data linkage.

Figure 1.1 shows how the studies and chapters are related to each other in the formation of this thesis.

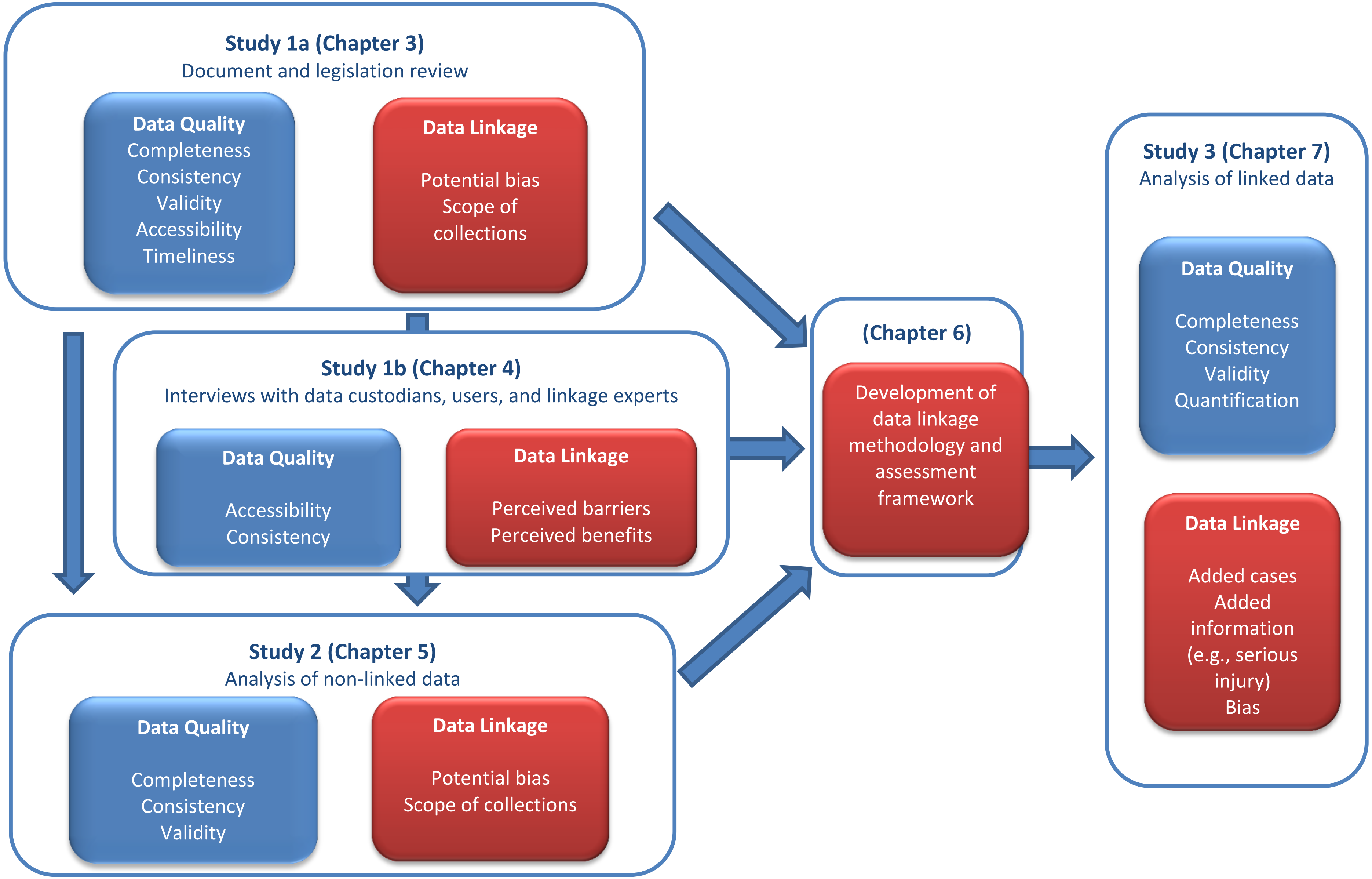


Figure 1.1 Flow of research program

Chapter Two presents a review of the literature relating to data quality and data linkage both generally and specifically for road safety monitoring. The literature review covers topics such as: data quality frameworks, methods of data linkage, benefits of data linkage, potential barriers of data linkage, and data linkage in road safety. At the conclusion of this chapter, a number of research questions are identified for the research program. Aspects of this literature review were included in a peer-reviewed conference paper presented at the Australasian Road Safety Research, Policing & Education Conference 2011:

> Watson, Angela, McKenzie, Kirsten, & Watson, Barry C. (2011) Priorities for developing and evaluating data quality characteristics of road crash data in Australia. In *Proceedings of Australasian Road Safety Research, Policing and Education Conference 2011*, Perth Convention and Exhibition Centre, Perth, WA (see Appendix A, for full paper).

Chapter Three outlines a review of data collections identified as including road crash injury cases. This was based on the results of Study 1a of the research program which included a review of legislation and documentation relating to the data collections as well as discussion with relevant data custodians. It outlines the scope, purpose, governance, data collection procedures, content, access, and timeliness of each of the relevant data collections. It also discusses the data quality implications of these findings and the potential for linkage of these data collections.

Chapter Four examines the perceptions of data custodians, expert data users, and data linkage experts of data quality and data linkage. It reports on the results of Study 1b, which was a qualitative analysis of semi-structured interviews with these groups. It explores issues relating to data quality characteristics of the data collections, including: relevance, completeness, and consistency. It also examines the perceptions of the potential benefits and barriers of using data linkage for road safety monitoring, planning, and evaluation.

Chapter Five includes the results of Study 2 within the research program, which involved the secondary data analysis of the road crash injury data collections. It includes analyses regarding the quality of the data collections in terms of completeness of variables, consistency, validity of coding, and representativeness. It also examines these issues specifically in terms of severity coding. Elements of this chapter have been included in a peer-reviewed conference paper:

> Watson, Angela, Watson, Barry C., & Vallmuur, Kirsten (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 & Education Conference*, Australasian College of Road Safety (ACRS), Brisbane Convention and Exhibition Centre, Brisbane, Australia (see Appendix A for full paper).

Chapter 6 is based on the results of Studies 1a, 1b, and 2 and involves the development of the data linkage approach that was taken for this program of research. It also outlines a framework by which an assessment of the success of the data linkage will be conducted, including assessments of completeness, validity, representativeness, and the issues associated with data linkage in this area.

Chapter 7 presents the results of Study 3, which involved the secondary data analysis of linked data from five data sources. It includes analysis relating to linkage rates, discordance, validity, and profiles of different combinations of linked data sources. It specifically examines the potential for linked data to enhance the quantification of serious injury and explores issues such as under-reporting to police.

Chapter 8 brings the findings of Studies 1a, 1b, 2, and 3 together and discusses the implications of these findings for road safety research and practice, as well as for road crash injury surveillance. The limitations of the research are discussed, as well as suggestions for future research.

Chapter Two: Literature Review

2.1 Introductory Comments

The current chapter reviews the available literature relating to data quality and data linkage, with a particular focus on road safety research, policy and practice. The key issues explored are: the importance of data in road safety; data quality evaluation frameworks; and the history, nature, potential benefits, and potential barriers of data linkage, particularly in the road safety context. This chapter also focusses on consolidating the available research literature and identifying gaps in current knowledge. In doing this, it lays a foundation for the program of research reported in this thesis.

2.2 Literature Search Strategy

Sources of information for this review included empirical journal articles and websites found using databases such as the Australian Transport Index (ATRI), PsychINFO, ScienceDirect, and TRIS Online (Transportation Research Information Services), and web based searches. A variety of search terms were used in combination such as: data linkage, road safety, injury, health data, administrative data, data quality, traffic data, and crash data. A review of data linkage centres around the world was also conducted to determine the nature of their programs (methods and framework). This review was completed using materials from websites relating to the centres as well as other reviews found in the literature.

2.3 The Importance of Data

High quality data are needed in road safety to: monitor trends; identify risk groups and locations; and make regional, interstate and international comparisons (Elsenaar & Abouraad, 2005). These data also make it possible to: design and apply appropriate interventions; and monitor the results and assess the impacts of interventions (Holder et al., 2001). Quality data can also be used to determine the cost implications of road trauma (Austroads, 1997). As stated in a World Health Organisation report:

> “Reliable, accurate data can also help build political will to prioritise road safety…….The use of reliable data to identify problems and target resources more effectively is a key element of the Safe System approach to road safety – an approach increasingly recognized as the most effective way to make road transport systems safer for all users” (World Health Organization, 2010).

In terms of particular government agencies or sectors and their priorities, there is a need for good quality data. For example, police rely on crash data to allow for intelligence-based enforcement, including the identification of speed camera and alcohol enforcement locations and timing (World Health Organization, 2010). Another example is transport authorities. They require information about crashes and their locations to inform policy, legislation, and develop interventions related to road infrastructure, vehicles, and driver behaviour. (World Health Organization, 2010). Health-related agencies also require quality information to inform health promotion programmes and evaluate their effectiveness. Data can also allow them to effectively plan trauma care and rehabilitation services (World Health Organization, 2010).

In order to perform all of these functions, data not only have to be available, but also must be of a high quality. To determine if a data source is capable of providing good quality information, examination is required to identify any limitations of the collection in relation to its capacity to report on road crash injury which may affect the accuracy and validity of conclusions that are able to be drawn from the data (Holder et al., 2001; Horan & Mallonee, 2003). This information could be obtained through an evaluation of a data collection and its capacity to perform injury surveillance both generally and within the specific road safety context (Mitchell, Williamson, & O'Connor, 2009).

2.4 Framework for Assessing Data Quality

There are a number of suggested criteria against which the quality of data related to injury can be assessed (Australian Bureau of Statistics, 2009; Mitchell et al., 2009). The Australian Bureau of Statistics (2009) outlines a data quality framework that consists of relevance, timeliness, accuracy, coherence, interpretability, and accessibility. Mitchell and colleagues (2009) developed a framework for evaluating injury surveillance systems based on both literature and expert opinion. It built on the existing frameworks discussed above and suggested that data need to be assessed on quality, operational and practical characteristics. Details of these characteristics are presented in Table 2.1.

Table 2.1: Criteria for the characteristics of the evaluation framework for injury surveillance systems

Data quality	Completeness	The amount of missing or unknown data for key characteristics of the injured population
	Sensitivity	Ability to correctly detect all cases of true injury events that the collection intended to detect
	Specificity	Ability to detect all non-injury cases that the data collection should not have detected
	Representativeness	Ability of the collection to provide an accurate representation of the distribution of key characteristics of the injured population
Operational	Purpose and objectives	The purpose, objectives and use of the injury surveillance system should be described
	Data collection process	The method of data collection for an injury surveillance system
	Case definition	The injury case definition adopted by an injury surveillance system to identify cases should be described
	Uniform classification systems	The classification system(s) used to record information in the injury surveillance system should be identified.
	Quality control measures	The quality control measures regularly utilised by the agency responsible for the injury surveillance system should be identified
	Confidentiality and privacy	The methods by which an individual's information in the injury surveillance system is safe guarded against disclosure should be described.
	System security	The data access requirements that safe guard against the disclosure of confidential information should be described.
Practical	Accessibility	The method by which potential data users access data from the injury surveillance system should be reported.
	Usefulness	Usefulness refers to the ability to contribute to the identification of potential key areas for preventive action in terms of the ability to: (a) identify new and/or emerging injury mechanisms; (b) monitor injury trends over time; and (c) describe key characteristics of the injured population (i.e. WHO's core minimum data set for injury surveillance).
	Data analysis	The routine data analyses conducted using data from the injury surveillance system by the agency responsible for the surveillance system should be described.
	Guidance material	The availability of guidance material on the interpretation of data from the injury surveillance system should be described.

Source: (Mitchell et al., 2009)

In terms of road safety in particular, a report by National Highway Traffic Safety Administration (1998b) lists six indicators of quality (timeliness, consistency, completeness, accuracy, accessibility and data integration with other information) which overlap to differing degrees with those outlined by Mitchell and colleagues (2009). Similarly, a number of other reports cover various aspects of these criteria, though in a much less structured way (Austroads, 1997; Turner, 2008) or in a quite specific context (e.g., spatial data) (Chapman & Rosman, 2008; Strauss & Geadelmann, 2009; Strauss & Lentz, 2009).

There are a variety of terms used to describe the key characteristics and quality of data and data systems. For the purposes of this review, data will be discussed in terms of six core quality characteristics: relevance; completeness; validity; consistency; timeliness; and accessibility. The concept of relevance was chosen because of its inclusion in the ABS Data Quality Framework (Australian Bureau of Statistics, 2009) and potential overlap with the concepts of usefulness, purpose, and representativeness outlined by Mitchell and colleagues (2009). The concepts of completeness and accuracy were chosen due to their inclusion in several of the guidelines (Australian Bureau of Statistics, 2009; Austroads, 1997; Mitchell et al., 2009; National Highway Traffic Safety Administration, 1998b). The concept of consistency is used in the NHTSA guidelines (1998b) and overlaps with the concepts outlined by Mitchell and colleagues (2009) such as usefulness and representativeness. Timeliness and accessibility are included as they are mentioned by all the guidelines reviewed (Australian Bureau of Statistics, 2009; Austroads, 1997; Mitchell et al., 2009; National Highway Traffic Safety Administration, 1998b). The six key data quality characteristics or concepts selected to underpin this program of research are described below.

2.4.1 *Relevance*

Relevance is defined as how well the data meet the needs of users in terms of what is measured, and which population is represented. Relevance is important in order to assess whether the data meets the needs of policy-makers and researchers and must be useful for planning and evaluation purposes (Australian Bureau of Statistics, 2009; Australian Transport Council, 2011). The needs of different data users are diverse, and what one considers ‘relevant’ may differ from another view. This means that within each record, a wide range of data items is usually needed.

Mitchell and colleagues (2009) discuss the term usefulness, which is a characteristic which also relates to the relevance of a data collection. As shown in Table 2, usefulness refers to the ability to: (a) identify new and/or emerging injury mechanisms; (b) monitor injury trends over time; and (c) describe key characteristics of the injured population (i.e. WHO's core minimum data set for injury surveillance).

In order to address the issue of relevance, the World Health Organisation’s Injury Surveillance Guidelines (Holder et al., 2001) recommend dividing injury surveillance data into two main categories (core and supplementary) with each of these then subdivided into ‘minimum’ and ‘optional’ data. The core minimum data set (core MDS) contains the least amount of data a viable surveillance system can collect on all injuries and usually includes:

- a unique person identifier;
- age of the injured person;
- sex of the injured person;
- intent (e.g. unintentional or resulted from violence or self-harm);
- location the injury occurred;
- nature of the activity being undertaken when the injury occurred
- mechanism or cause; and
- nature of the injury (Holder et al., 2001).

The core optional data set (core ODS) involves information that is not necessary to collect but may be collected, if it is seen as useful and feasible to collect. Optional data may include:

- race or ethnicity of the injured person;
- external cause of injury;
- date of injury;
- time of injury;
- residence of the injured person; and
- severity of injury (Holder et al., 2001).

It is also suggested that the core ODS include a narrative or a summary of the incident.

Supplementary data includes any additional data that a surveillance system wishes to collect on specific types of injury, such as those that are road crash related. In the case where an injury surveillance system focusses on a particular type of injury, it would be suggested that more than just core information would need to be collected. It can be divided into the supplementary minimum data set (supplementary MDS) and supplementary optional data set (supplementary ODS) (Holder et al., 2001). The supplementary MDS is the least amount of additional data a surveillance system collects on a particular type of injury and supplements the data collected as part of the core data set.

In the case of road crash injuries, relevant information may include details about the circumstances of an incident (e.g., speeding, fatigue) or about other people involved (even if not injured). The National Highway Traffic Safety Administration (NHTSA) in the United States outline that data should include information about the roadway, vehicle, and driver (National Highway Traffic Safety Administration, 1998b). An Austroads report (1997) emphasises the importance of information relating to the precise geographical location as well as the inclusion of speed limit, road design, lighting conditions, weather conditions, road surface, traffic control, crash type, vehicle type, road user, severity, licence type (including unlicensed), alcohol and/or drug involvement, work-relatedness, restraint use, helmet wearing, and seating position. A WHO report on minimum crash data set (World Health Organization, 2010) also outline these variables as important, if not necessary, inclusions.

In 1970, William Haddon Jr. designed a tool for analysing an injury event. Haddon's Matrix allows simultaneous consideration of factors and the stages, over time, of an event. As shown in Table 2.2, the matrix involves three stages: pre-event; event; and post-event and four factors: host; agent; physical environment; and social environment. In this table, the examples provided relate specifically to a road crash event.

Table 2.2: Haddon's Matrix

	Host (human)	Agent (vehicle)	Physical environment	Social environment
Pre-event	Pre-disposed or over-exposed to risk (e.g., substance abuse, lack of driving skill)	Hazardous vehicle (e.g., faulty brakes)	Hazardous environment (e.g., slippery road)	Environment encourages risk-taking or hazards (e.g., social acceptability of speeding)
Event	Lack of tolerance force (e.g., not wearing a seatbelt)	Un-protective vehicle (e.g., no airbags)	Environment contributes to injury (e.g., roadside hazards)	Environment contributes to injury (e.g., lack of speeding enforcement)
Post-event	Severity of trauma (e.g., older driver)	Vehicle contributes to trauma	Environment adds to trauma (e.g., slow emergency response)	Environment contributes to recovery (e.g., lack of rehabilitation support)

Source: Injury Surveillance Guidelines (Holder et al., 2001)

The matrix can be used to analyse an injury event in order to identify interventions that may prevent the event from happening, or reduce the harm arising if it does occur. Therefore, in order for interventions to be developed and evaluated, it is important that data include information on each of these stages and factors.

In light of these suggested data elements and their potential relevance and usefulness to road safety researchers, practitioners, and policy-makers, it is possible that a data collection could be considered relevant or useful if any of these fields is present.

2.4.2 *Completeness*

Strongly related to the issue of relevance is completeness. Completeness refers to the extent to which all relevant cases, all relevant variables, and all data on a relevant variable are included in the data collection (Mitchell et al., 2009). Firstly, data collections would be considered complete if they detect all cases of road crash injury they intend to detect by definition (sensitivity) and unlikely to detect those injury events they do not intend to detect (specificity). This relates to the issue of representativeness. In other words, to what extent the data collection represents the population of all road crash injuries or incidents (Mitchell, et al., 2009). In order to draw conclusions on the incidence and distribution of road crash injury, the data collection would need to include all of these injuries regardless of the type of injury, where the injury occurred, or who was injured. Non-representative data may focus prevention efforts on populations that are not truly at risk and could result in a misdirection of resources (Mitchell et al., 2009).

Most data collections do not include all road crash injuries, instead only including those that fit a particular definition that is relevant for the collection's purpose. For example, hospital admissions data would only include those road crash injuries that were serious enough to involve admission to hospital. Therefore, hospital admissions data would only be representative of serious road crash injuries rather than of all road crash injuries. Data collections based on police reported incidents would also not be representative of the entire injury population, as certain road crash injuries do not fit the definition for inclusion in these collections (e.g., if the injury does not occur on a public road). It is important to understand that a data source may only be relevant for the understanding of a particular sub-population of transport-related injuries, and that possibly no single source of data will provide the complete picture of the road crash injury problem. Another issue besides the definition of inclusion in a data collection is that not all road crash injuries may be included in a data source due to a failure to report. For example, the reliance on police data for the counting of road crash injuries could be problematic, as it well known that not all road crash injuries are reported to police (Alsop & Langley, 2001; Amoros et al., 2006; Boufous et al., 2008; Langley, Dow, et al., 2003). This under-reporting can have impacts not only on the overall measure of the impact of road crash injuries, but also that this under-reporting could potentially be biased towards particular groups of road users (see Section 2.4.4 on consistency).

The completeness of a data collection in terms of whether it includes all the relevant cases is often difficult to determine. This is due to that fact that often the 'true' population is unknown. In other words, there is no 'gold standard' to which data collections can be compared. A possible method to address this problem is the capture-recapture method. This method uses two or more sources of data that contain relevant cases to estimate the population (Corrao, Bagnardi, Vittadini, & Favilli, 2000; Hook & Regal, 1995, 2000). Once this estimate has been calculated, researchers can explore how well each data collection, or combination of data collections, best represents the total number of road crash injuries. This method, along with comparisons of profiles of road crash injuries for each data collection, can also provide information about any bias or inconsistency in the data collections in terms of capture (see Section 2.4.4) which could affect their representativeness.

Another issue relating to completeness is that the data collection would need to include all relevant variables. As a benchmark, Mitchell and colleagues (2009) suggest that if between 76% and 100% of the Core MDS and ODS (see section 2.4.1) were included in a data collection, it would rate as ‘very high’ on completeness of variables. In a road crash injury specific context, there are other data elements, as mentioned previously (section 2.4.1), that would also be required to consider a data collection as having high completeness.

Also, not only should the collection include variables relating to the Core MDS and/or Core ODS, these variables should have minimal missing and/or unknown data for them to be considered complete. Mitchell and colleagues (2009) suggest that a ‘high’ level of completeness would exist if less than 5% of data within a specific field is missing. In addition to missing or unknown data, a data collection can lack completeness if there are a large number of unspecified or ‘other’ specified classifications (Mitchell et al., 2009). Incomplete data can be due to a lack of detailed information required to assign a code or classification, a lack of appropriate codes or classifications, lack of time, or lack of skilled coders (Mitchell et al., 2009; National Highway Traffic Safety Administration, 1998b). The impact of incomplete data is that the data collection may not provide enough information to allow for adequate data interpretation and could lead to flawed or biased results and therefore poor decision making.

2.4.3 *Accuracy*

Accuracy in this context refers to the degree to which data correctly describe the events or persons they were designed to measure (Australian Bureau of Statistics, 2009). Location information for engineering purposes demands a very high degree of accuracy (within metres), which is frequently not met (Austroads, 1997; Strauss & Lentz, 2009). If location information is not accurate, a problem location might go undetected, and the nature of a location-specific problem might be difficult to determine due to incomplete data. Roadside objects may contribute to occurrence and severity; and thus must be identified, along with their role (e.g. struck first or as a result of a collision between vehicles). This is important both for specific locations, and across the road system.

One of the main indicators of the safety and operation of the road system is the occurrence of road crashes at different levels of severity. Accurate severity information is important for prioritisation of locations, understanding road crash mechanisms, and for evaluating the effectiveness of interventions or countermeasures. Both in Australia and around the world, police data have been the primary source of this information. However, the definitions of severity do differ across jurisdictions. The definition of a fatality is relatively consistent and usually fits the Organisation for Economic Co-operation and Development (OECD) definition of a death within 30 days of a road crash. In terms of other severity levels, particularly in defining a serious injury, the definitions are much more variable. Many of the countries in the OECD define a serious injury as a person who is admitted to hospital for 24 hours or more as a result of a road crash (World Health Organization, 2010). However, with a reliance on the police to classify this definition of severity in most cases based on the person being transported to hospital, and with a reported lack of liaising between police and hospitals on the length of admission, a serious injury category using this definition could range from cuts and bruises to severe head injuries. Also, the varying admission policies across jurisdictions could impact on this measure (World Health Organization, 2010).

As a result of this broad, and likely inconsistent, serious injury classification, more objective and precise measures of severity have been proposed (International Traffic Safety Data and Analysis Group (IRTAD), 2011) which rely on either police to assign a nature of injury code or rely on the use of hospital discharge diagnoses (e.g., Abbreviated Injury Scale, ICISS). The Abbreviated Injury Scale (AIS) is a body-region based coding system developed by the Association for the Advancement of Automotive Medicine (Association for the Advancement of Automotive Medicine, 2008). A single injury is classified on a scale from 1-6 (1 = minor; 2 = moderate; 3 = serious; 4 = severe; 5 = critical; and 6 = maximum). Another example of a more precise measure of severity is the International Classification of Diseases–based Injury Severity Score (ICISS) (Osler, Rutledge, Deis, & Bedrick, 1996). ICISS involves using ICD diagnoses to calculate threat-to-life. Survival Risk Ratios (SRR), which are the proportion of cases with that diagnosis code which did not die, are calculated for each ICD diagnosis code. Cases are then assigned an ICISS, which is the multiplication of SRRs of all their diagnoses. It should be noted that there is some debate surrounding the most appropriate injury severity classifications, however these two measures are widely accepted and often used in injury research as reasonably reliable measures of the probability of death (Langley & Cryer, 2012; Stephenson, Langley, Henley, & Harrison, 2003).

It could be suggested, however, that even if more detailed information was collected in order to assign these more objective and/or precise measures, the police are not necessarily in the best position to collect this information. Police do not have the training or expertise to record information on the nature of an injury, or injuries, with the required level of accuracy. Also, even if they were trained to assess this, classifying injury at the scene of a crash could be problematic, as not all injuries are apparent at the scene and the police have many competing priorities in these situations (e.g., traffic control). Also, it is argued that the consistency of the recorded information from case to case could be questionable (Amoros, Martin, Chiron, & Laumon, 2007; Chapman & Rosman, 2008; Farmer, 2003; McDonald, Davie, & Langley, 2009; Ward et al., 2010). The World Health Organisation (2010) suggests some possible strategies for addressing the issue of serious road crash injuries, including data linkage between police and hospital databases either periodically to check the accuracy of the police data or to be routinely included in reporting; and/or the following up of cases by police (or reported by the hospital) to determine the length of the hospital stay.

The accuracy of a data collection, and the variable fields within it, is difficult to assess as there is often no real comprehensive or objective data by which to compare the data to a gold standard. However, the literature does suggest that accuracy in part may be assessed by determining if certain aspects known to enhance the accuracy of data, such as: standardised coding and/or classification (e.g., ICD, AIS); quality control procedures; and the use of technology (GPS), are present (Mitchell et al., 2009; National Highway Traffic Safety Administration, 1998b). It should be noted, however, that even when these coded variables are present, they are not always accurately recorded. For example, previous research has shown that external cause ICD coding is not always accurate with anywhere between 8% and 26% of external cause codes being incorrect (Davie, Langley, Samaranayaka, & Wetherspoon, 2008; Hunt et al., 2007; Langley, Stephenson, Thorpe, & Davie, 2006). This has implications for the ability to identify relevant road crash injury cases and their related circumstances in data that rely on this coding.

While the accuracy of data may be difficult to determine, it is possible to get an indication of the accuracy of a data collection by assessing validity. Researchers have applied what is known as criterion validity to assess a data collection, by comparing the data collection with a criterion or a 'truth', often referred to as a 'proxy gold-standard'. Using this approach, the data collection is compared to detailed text records, such as medical records (Butchart et al., 2001; Johnson et al., 1997; Langley et al., 2006), with self-report surveys (Yorkston, Turner, Schluter, & McClure, 2005), or other data collections that include the same variable and/or population (Fox, Stahlsmith, Remington, Tymus, & Hargarten, 1998). However, to use these methods of validity assessment, an assumption about the 'truth' of the proxy gold-standard has to be made. While in some cases this may be possible, it is not always. For example, the assumption that a patient's self-reporting of an injury will be more accurate than the record in the police or hospital data is not always appropriate. As the self-report data is only able to be gained after the event (sometimes a long period of time after, for serious cases), issues with patient recall can be a problem. Also, there would be some cases where the self-reported information about the event would be unavailable due to the injured person dying, being unable to communicate, or experiencing memory deficits as a result of their injury. As an alternative, validity assessments could rely more on what is known as convergent validity, in which the validity is measured as the degree of correspondence between two or more measures, or in this case variables and/or cases within two or more data collections, designed to measure the same thing. Data linkage can be used to compare the data collections and examine the concordance between them in terms of both the cases they capture and the categorisation or coding of the variable elements. Using this method, there is no assumption about which, if any, of the data sources are the 'truth', it simply makes the assertion that if all the data collections in the comparison include the same case and coded or classified that case in the same way, and are relatively independent of one another, you could be confident that they are measuring the same thing and therefore be valid.

It should be noted that there are trade-offs between accuracy and other criteria. Databases which provide a high degree of accuracy and detail, e.g. the National Coronial

Information System, tend to have much more restricted coverage (Young & Grzebieta, 2008). Also, it is not always the case that accurate information is reliably or consistently recorded.

2.4.4 *Consistency*

Consistency of data refers to their ability to reliably monitor road crash injuries over time, and compare between characteristics within a data set as well as across other relevant data (Australian Bureau of Statistics, 2009). Ideally, the quality of the data should not vary over time, nor should they vary in quality, by the nature of the event/injury, where or when the event/injury occurred, or who was injured or involved. Essentially, users of the data need to be confident that any changes over time or differences between events/individuals are due to actual changes or differences, not simply due to inconsistencies in the data (Holder et al., 2001; National Highway Traffic Safety Administration, 1998b).

An apparent increase or decrease in the number of road crash incidents or injuries over time could be caused by a number of factors and may not reflect any actual change in these incidents. Changes in reporting criteria, work policy or practice (e.g. hospital admission policy, responsibility changes), or coding/classification systems, could result in the incidents or injuries that used to be recorded no longer being recorded or vice versa.

Inconsistencies within the data based on the characteristics of the incident or injury can also occur for a variety of reasons. Firstly, reporting, work policy or practice, or coding/classification systems may vary by the location of the incident/injury. An incident occurring in a remote location may not be reported, or a lack of resources in some hospitals may lead to less detailed classification. Besides the location of the incident, certain types of incidents/injuries may be less likely to be reported or coded/classified validly or adequately. For example, a road crash incident involving illegal behaviour (e.g., unlicensed, alcohol) may not be reported to police to avoid prosecution. If there are inconsistencies in reporting in a particular data collection, their representativeness can also be affected (see Section 2.4.2).

One suggested way of enhancing the consistency of a data collection is the use of uniform classification systems (Holder et al., 2001; Mitchell et al., 2009; National Highway Traffic Safety Administration, 1998b). These systems should include a comprehensive set of standard coding/classification guidelines which should be readily available to personnel assigned the duty of recording, classifying or coding data collections. These personnel should also be specifically trained in the procedures and should refer to the guidelines often. Without this training and available material, personnel could base their coding or classification decisions on their own intuitions, opinions, or preconceived notions (German et al., 2001). It is also necessary that any changes to reporting, classification, and recording should be documented in detail (National Highway Traffic Safety Administration, 1998b).

2.4.5 *Timeliness*

Timeliness refers to the delay between the date an event occurs and the date at which the data become available (Australian Bureau of Statistics, 2009). It is suggested that data should become available for use quickly, however the definition of what is ‘quick’ may vary between agencies (Austroads, 1997). It is crucial that agencies are able to respond rapidly to emerging problems, so that the rapid processing of road crash incident data to make it available is a key concern. For example, Logan and McShane (2006) noted that clusters of crashes could develop quickly, in just a couple of years. Unless the data become available quickly, techniques aimed at detecting emerging clusters will not be effective. Data also needs to be timely for effective evaluations of countermeasures and interventions (National Highway Traffic Safety Administration, 1998b). Mitchell and colleagues (2009) rates the timeliness of the collection, availability, analysis and dissemination as being of high importance for injury data collections. Specifically, they suggest that if data are disseminated within a month the data collection would rate as ‘very high’; one to two years as ‘high’, and more than two years as ‘low’. The NHTSA (1998b) suggest that it is preferable for data to be available within 90 days. However, they highlight that some supplemental information could wait longer.

The nature of some sources of data means that not all data items can be entered into the database at once; if the data items that have been completed are withheld until each crash record is complete, timeliness will be affected. For example, blood alcohol concentration (BAC) data cannot be entered until results of the toxicology analysis are made available.

Another factor that could influence the timeliness of data availability is related to resourcing. Specifically, an insufficient number of trained personnel to input, code, analyse and/or interpret the data will likely have a negative impact on the timeliness of the data. It is also the case that the roles of the personnel involved, particularly relating to inputting and coding data, are quite diverse (i.e., police officers, nurses), with their priorities directed toward other, arguably more important, tasks (e.g., patient care). This demand on resources can increase the time taken for data to become available.

There are also often trade-offs between the timeliness of the data collected and the level of detail recorded regarding a case, as well as the accuracy, completeness and consistency of the data. While the processes that may be in place for coding, recoding, checking, and cleaning of data improve the consistency and accuracy, it may also then increase the time taken for the data to become available, therefore reducing timeliness.

2.4.6 *Accessibility*

Accessibility relates to the ease with which data can be accessed, which includes ascertaining its availability and suitability for the purpose at hand (Austroads, 1997). The NHTSA (1998b) suggests that data should be readily and easily accessible to policy makers, law enforcement, and for use in road safety research and analysis. The NHTSA

(1998b) further suggest that data should be available electronically, at a unit record level, provided that safeguards are in place to protect confidentiality and privacy. Mitchell and colleagues (2009) suggest that if data is accessible to users in a unit record format from an internet-based interface or data warehouse, it would rate as ‘very high’ on accessibility. While it may be ideal to have free and easily accessible data, there are a number of issues that can limit accessibility.

Major barriers to accessing data relate to confidentiality and privacy. Even when names and addresses are removed, there is still concern that variables such as age and gender in combination with location and temporal variables can lead to the identification of the person/s involved. It is important to understand and comply with the legislation and policy that relates to the particular data collection so that ethical research can be conducted including the protection of privacy of the individual.

However, legislation, policy, and guidelines can often be open to interpretation which can complicate the process of negotiating access with different agencies. Therefore, it is often the case that these processes can include intense negotiations which can go back and forth over a long period of time. Even if the process is straightforward, completing the required documentation and having it considered by the relevant authorities can still be quite time consuming.

Another potential barrier to accessing data involves concern that data will be misinterpreted or misreported. This is particularly a concern when data custodians are not confident that end-users of the data are aware of the data constraints, limitations and coding conventions. This issue may potentially be overcome by end-users and data custodians communicating better about the nature of the data, including coding information, scope and limitations, as well as by discussing the reporting of data prior to its release or publication.

A third possible barrier to access lies with the data systems themselves. Some data sets do not have relevant information in a format that is quantifiable. Instead they have long text descriptions or reports, making extraction of specific information about an incident or its location difficult and time consuming. Even in the case of data being held in a suitable format, the software used may be difficult to navigate, except for those who are specifically trained, and may not be easily extracted and exported into a format conventionally used by those who work with data (i.e. Excel, text delimited, SPSS, or Access).

2.5 Data Linkage

Data linkage involves the bringing together of two or more different data sources that relate to the same individual or event (National Collaborative Research Infrastructure Strategy, 2008). In principle, any datasets that contain information about individuals has the potential to be linked. Data linkage is used for a variety of reasons including data quality improvements and gaining information that a single data source cannot provide. Details of the potential uses and benefits of data linkage as well as the different methods and frameworks are described in subsequent sections.

2.5.1 *Data linkage centres*

There are a variety of data linkage centres in operation around the world. The longest-running and perhaps the most successful example of data linkage within Australia is the Western Australian Data Linkage System (WADLS). The system was established in 1995 and is a multi-set system for the creation, storage, update and retrieval of links between health-related data. It uses probabilistic matching (see Section 2.5.2) to create a Master Linkage Key between over 30 population-based research and administrative health data collections in Western Australia (WA). Up to 2008, the linkage program has contributed to over 400 research projects (Holman et al., 2008).

The Centre for Health Record Linkage (CHeReL) was established in 2006 to create and maintain a record linkage system for health and human services in New South Wales (NSW) and the Australian Capital Territory (ACT). It involves collaboration between ACT Health, the NSW Clinical Excellence Commission, the NSW Department of Health, the Cancer Institute NSW, The Sax Institute, the University of New South Wales, the University of Newcastle, the University of Sydney, and the University of Western Sydney. By June 2009, 57 linkage projects had been completed with a further 30 underway. In total, 42 million records have had a Master Linkage Key attached to them.

In 2009, the South Australia and Northern Territory Link (SA-NT Link) was established, linking records from a variety of government sources (14 data collections as at the end of 2012). Victorian Data Linkage (VDL) has also been established and will routinely link a number of core data collections (including hospital and death data). As with WADLS and CHeReL, SA-NT Link and VDL use probabilistic linkage for the creation of Master Linkage Keys. A data linkage unit is also being established in Tasmania (Tasmanian Data Linkage Unit).

In Queensland, the Data Linkage Unit within Queensland Health has been linking health data since 2008. However, there is limited published information to describe whether the unit has the capacity or infrastructure required to link external data sets, or whether external agencies are willing or able to release identifying information to the unit for linkage purposes. This program of research was informed by discussions undertaken with this group as part of Study 1 (see Chapter 4, Section 4.4).

Each of the state and territory data linkage centres in Australia is a node of the Population Health Research Network (PHRN). The PHRN was established to facilitate data linkage in Australia. The PHRN has also established the Centre for Data Linkage (CDL) at Curtin University. This centre is tasked with establishing a secure facility to link data between the jurisdictions around Australia. Two other centres for linkage within Australia, that conduct linkage of national data, are the Australian Institute of Health and Welfare (AIHW) and the Australian Bureau of Statistics (ABS).

Overseas, there are a number of data linkage centres. In Canada, there is Population Health British Columbia; Manitoba Centre for Health Policy; and Statistics Canada. Each of these centres includes the linkage of administrative, registry, and survey data. They were generally developed to provide profiles of health and illness, as well as facilitate multi-sectoral research in areas including health, education, and social services. In Great Britain, there is the Oxford Record Linkage System with the linked dataset used for health services statistics and for epidemiological and health services research. Also in Great Britain, is the Secure Anonymised Information Linkage Databank (SAIL Databank) at the College of Medicine, Swansea University, Wales. The SAIL Databank was established in 2006 to conduct data linkage for health related research and currently holds 500 million records. In the United States a road safety specific data linkage system is the Crash Outcome Data Evaluation System (CODES). CODES allows for the tracking of those injured in road crashes from the crash through the health system to provide information on the outcomes of injuries. It also allows for the linkage of crash data with licensing, registration and traffic histories to gain information about the antecedents to and associations with road crash injuries. Data collections routinely linked in CODES includes: traffic records, crash data, emergency services, emergency department, insurance, admitted patients, rehabilitation services, and death records. According to a report released in 2009 (NHTSA, 2009), CODES is operating in 20 states across the United States, although the status of these states systems is often changing.

2.5.2 *Methods of data linkage*

There are two main methods of data linkage: deterministic and probabilistic. Deterministic linkage requires an exact match with common identifier/s between the data collections. Probabilistic record linkage matches on multiple identifying data, but instead of an exact match, a match is created when the calculated statistical probability of a match exceeds a certain predetermined threshold (National Collaborative Research Infrastructure Strategy, 2008).

In the deterministic method, the unique identifiers need to be able to identify an individual across data collections and across time and be able to absolutely discriminate them from another person. It relies on these unique identifiers being accurately and reliably recorded as it must match exactly, with no room for error. This method relies on these data elements to be recorded accurately and reliably. The deterministic method often works best when there is a single unique identifier, like an ID number, that is shared between the data collections.

The probabilistic method was developed to deal with situations in which the identifying data elements shared between sources (or over time) are not always reliably or accurately recorded and therefore may not match exactly. Weights are assigned to matches in terms of their distance, or degree of difference, from each other. Each of the identifiers are assigned a weight and the total weight across the variables is used to determine if a record is linked, not linked, or possibly linked based on a predetermined threshold (Winkler, 1999). The probabilistic method was first introduced by Newcombe and colleagues (1959) and was then expanded upon by Fellegi and Sunter (1969). Their mathematical models are still the foundation of many probabilistic data linkage programs and the probabilistic method is the most commonly used technique in data linkage centres around the world.

2.5.3 *Data linkage framework*

Many data linkage centres apply what is known as the 'separation principle', by only using identifying information required for linkage without any content or clinical data (WADLS, CHeReL, VDL, SA-NT Link). The data linkers' task is to establish links using this identifying information and assign a linkage key to each match. This linkage key is then sent to the custodians for them to extract the relevant content data and provide these data with the linkage key to the researcher. Using this separation principle approach means that those performing the linkage will be unaware of the circumstances by which any individual is included in the data collection or any details relating to these circumstances. Also, the researcher will only have the data required for analysis without any identifying information. No entity, except the data custodian, ever has access to both the personal data and the content data. This approach is often used to preserve the privacy of the individual as well as to allow data custodians to maintain control over the data collections within their governance and is considered best-practice in Australia (Boyd et al., 2012).

The data collections that have historically been included in data linkage include population-based data collections (e.g., census, registry of births, deaths, marriages) and health data collections (hospital admissions, cancer registries) (e.g., WADLS, CHeReL, and VDL). In recent years, data linkage has gone beyond health data and included data sources such as education, child protection, corrections, and police data. However, these non-health data collections are usually not included in the routine linkage process, but instead done on an ad-hoc or project basis. Other, non-administrative or population data sources have also been used in data linkage, including survey data from cohort and longitudinal studies. In this case, participants in the studies would provide consent for their data to be linked to administrative data. This type of linkage would also be done on a project basis rather than routinely.

2.5.4 *Benefits of data linkage*

There are a number of suggested benefits of using linked data for research, monitoring and policy development (Glasson & Hussain, 2008; Goldacre & Glover, 2002; Holman et al., 2008; Productivity Commission, 2013). It is possible that data linkage can result in improvements to data quality by including more cases or variables and increasing accuracy through the detection and correction of errors. Another application of data linkage is the ability to use the capture-recapture method described earlier (Section 2.4.2).

The capture-recapture method requires data linkage in order to determine the number of cases data collections have in common which is a key part of the calculation. This method has been used in a variety of health settings (Corrao et al., 2000; Klevens et al., 2001), including exploring under-reporting of road crash injury (Meuleners, Lee, Cercarelli, & Legge, 2006; Miller et al., 2012; Thomas, Thygerson, Merrill, & Cook, 2012)

It is also argued that data linkage can be cost-effective. By linking pre-existing data to provide additional information and address research questions, there is less need to collect additional data on an ad-hoc basis which can be time consuming and expensive (Goldacre & Glover, 2002; Productivity Commission, 2013). Data linkage allows for longitudinal study of key health and social outcomes for the population, by tracking individuals through the various government systems. In the cases where data is linked to population-based data collections, it could improve the ability for researchers to better estimate the prevalence and incidence of certain conditions or events in the community. It has also been argued that in case-control type studies, data linkage can help identify control groups that are more representative and inclusive (Productivity Commission, 2013).

While the benefits of data linkage in other areas of health have been well established, it is less clear what the benefits of data linkage are for road safety. A report by Cairney (2005) suggests that investment in linked data systems for road safety would likely lead to more efficient day-to-day operations and easier access to data for decision makers. It was suggested that the linking of databases will greatly increase the value of data sets by allowing the use of data for a wider range of purposes (Cairney, 2005). One potential benefit relates to the identification of under-reporting by police. If the under-reporting and any related bias can be quantified in a jurisdiction, then adjustments could be made to the reporting of these cases and allow for better examination of the true impact of road crash injury in the community. Specifically, the capture-recapture method, described earlier (see Section 2.4.2), could be applied to make estimates about the true population of road crash injuries and draw conclusions about how well each data collection represents this population.

In addition, a possible benefit relates to serious injury classification in road crash injury. As stated earlier (see Section 2.4.3), the classification of serious injury by police has met with some criticism, and other data collections, if linked, could provide valuable insight into this issue. Specifically, the inclusion of health data, where the nature of an injury is more clearly defined and captured by qualified clinical personnel, could allow for more objective and precise measures of severity to be used to establish a more complete and accurate assessment of the impact and cost of road crash injuries to the community.

2.5.5 *Potential barriers to data linkage*

The first major barrier relates to issues of privacy and confidentiality that were mentioned previously (see Section 2.4.6). In order to conduct a record linkage project, a researcher needs to obtain approval from multiple data custodians and human research ethics committees. The time and effort involved in this process may discourage the frequent conduct of record linkage studies (Ferrante, 2008). It may also be necessary to involve an appropriate third party (or possibly one of the data custodians) in the data linkage process, as access to the identifying information required for data linkage is more restricted, if not prohibited, for researchers. It is important to note, however, that processes in order to provide linked data to researchers while safe-guarding privacy have been established in other Australian jurisdictions as well as overseas.

Another potential barrier is the linkage process itself. In the case of the data sources discussed previously, though information in different data sets may relate to the same incident, person or case, there is no system of unique identifiers across all data sets. Also, in the case of the police data often the unique identifier is assigned to an event (i.e., the crash), while the unique identifiers within health data sets are assigned to a patient. It is possible that the probabilistic method may be more useful in the absence of a shared unique identifier. However, this method relies on having specific and accurate information on the relevant variables in both data sets and requires that enough points of matching can be chosen so that no two events or individuals will be confused leading to a lack of specificity. Conversely, if the data matching criteria is too specific, there is a potential for an individual to not be matched despite them actually being present in both data sets (lack of sensitivity). So although this method has been utilised in the past in other jurisdictions, a limitation is that the formats used with different data sets may not be compatible, resulting in an inability for some of the data sets to communicate with each other or make errors in matching.

There is little research on the perceived barriers to data linkage, particularly in terms of establishing links with data outside of health (e.g., police data). However, there have been some discussions about the slow, sometimes lack of, uptake of this approach in some jurisdictions and/or sectors (Ferrante, 2008). Some of the perceived barriers in addition to those above for expanding data linkage to other jurisdictions and/or sectors include a lack of willingness from data custodians and ethics committees and resource limitations (Ferrante, 2008; Productivity Commission, 2013). One reason for the reluctance among some agencies may involve misunderstanding about the data linkage process. It is possible that some custodians believe that they need to supply all of their data to another agency, rather than just the identifying data, which could cause concerns about their governance. Some custodians may also believe that data linkage would require personal information to be provided to researchers for linkage and this would result in violations of their privacy obligations, particularly in situations where participant consent was not possible. It is also possible that some data collections by their nature can violate the ‘separation principle’. Specifically, some data collections are so defined by their scope that the data linkers would have information about the individuals within them simply by their inclusion on the data collection. As an example, if a corrections agency were to supply identifying information to a data linker, based on the scope of the data collection, the linker would know that the individuals within that collection had been involved in some sort of criminal activity. Another barrier may relate to data custodians and other relevant parties being unaware of the benefits of data linkage. Alternatively, they may be aware of the benefits to research, but do not see the benefit for their core business, or may believe that any benefits would not outweigh their concerns about data ownership and privacy.

2.5.6 *Data linkage in road safety*

In the area of road crash incidents and injuries, a variety of data linkage projects have been conducted (Alsop & Langley, 2001; Amoros et al., 2006; Aptel et al., 1999; Boufous et al., 2008; Cercarelli, Rosman, & Ryan, 1996; Langley, Dow, et al., 2003). Alsop and Langley (2001) used probabilistic linkage of police and hospital records in New Zealand. They found that less than two-thirds of all hospitalised road crash casualties were recorded in the police data. They also found that this varied based on the number of vehicles involved, the geographical location, age and injury severity. Amoros, Martin, and Laumon (2006) conducted a similar study looking at the under-reporting of road crash casualties in France. They used probabilistic methods to link police crash data with the road trauma registry in Rhone County. The results showed a police reporting rate of around 38%. However, this rate varied according to injury severity, the road user type, and the location of the crash (i.e., metropolitan vs. rural). Another French study conducted by Aptel and colleagues (1999) found that after linking police and hospital data, only 37% of non-fatal road crash injuries were recorded by police. Similar to other studies, they found that rate of reporting varied depending place of crash, the type of vehicle involved, and the injury severity. They also determined that police-reports tended to over-estimate the severity of the injury sustained. Langley and colleagues (2003) conducted probabilistic linkage between hospital records and police records to specifically examine the potential under-reporting of cyclist injuries in New Zealand. The results showed that only 22% of cyclists that crashed on a public road could be linked to the police records. Of the crashes that involved a motor vehicle 54% were recorded by police. They also found that age, ethnicity, and injury severity predicted whether a hospitalised cycle crash was more likely to be recorded in the police data. Within Australia, Cercarelli and colleagues (1996) linked police reports, hospital admissions and accident and emergency (A&E) department data. The researchers found that around 50% of attendances at the A&E were recorded by police, and that around 50% of cases recorded by police as being admitted to hospital were actually admitted. The researchers outline that while the discrepancy between the data sets does represent an under-reporting of cases, it also suggests that differences in coding systems may also lead to cases not being linked. Another Australian study conducted in NSW by Boufous and colleagues (2008) linked hospital admissions data (Inpatient Statistics Collection [ISC]) with the Traffic Accident Data System (TADS). Using probabilistic linkage, the researchers matched 56.2% of hospitalisations as a result of road crash with a record in TADS. The researchers also found that the linkage rate varied according to age (i.e., lower linkage rate for younger age groups), road user type (e.g., lower linkage rate for cyclists), severity (i.e., higher linkage rates with increased severity) and geographical location.

While these studies highlight the issues of under-reporting and bias within police data systems, the barriers and limitations of data linkage were not explored either at all, or in any depth. Also, many of these studies tended to limit data linkage to only two data sets (e.g., hospital and police data) rather than exploring the methods, issues and findings from linkage of several data sets to obtain a more complete picture of road crash injury. There is also an opportunity to examine the use of data linkage as a method for exploring the quality of the police data (that is relied upon so heavily in road safety) particularly expanding on the work relating to severity and the classification of serious injury (Amoros et al., 2007; Chapman & Rosman, 2008; Farmer, 2003; McDonald et al., 2009; Ward et al., 2010) as well as further exploring the under-reporting of cases to police using the capture-recapture method.

There has also been no research of this nature conducted in Queensland, with the majority of the studies conducted using New South Wales, Western Australian and international data sources. Each jurisdiction has different data systems with unique considerations for linkage and use. If the ultimate aim, as researchers in the area suggest (Cairney, 2005; Holman et al., 2008; Turner, 2008), is to create an integrated national data linkage system, then it is important to understand the nature of each state's (including Queensland's) information systems and data linkage capabilities.

2.6 Research Questions

In order to address the aims of this research and gaps in knowledge discussed in the literature review, the following research questions have been formulated for this program of research:

RQ1: *How well do data collections which collect road crash injury information in Queensland conform to the core/minimum requirements for road crash injury data?*

This research question will be addressed as part of Study 1, by reviewing the characteristics of the available data collections in Queensland that include information on road crash injury. As shown in the literature, it is important to consider the relevance of data collections to road safety related research, policy, and practice. The relevance of the data will be assessed based on their compliance with the minimum data requirements outlined by the WHO guidelines as well as other national and international guidelines relating to injury surveillance and road safety specifically (Austroads, 1997; Holder et al., 2001; National Highway Traffic Safety Administration, 1998a; World Health Organization, 2010).

RQ2: What are the strengths and weaknesses of each of the road crash injury data collections within the context of road safety investigation, intervention development, and evaluation?

This research question will be addressed as part of Studies 1, 2 and 3. In Study 1, by reviewing the characteristics of the available data collections and from interviews conducted with both researchers that use these data and the relevant custodians, the strengths and weaknesses of the data collections can be explored. Specifically, data quality characteristics of completeness, consistency, accuracy, accessibility, and timeliness will be examined. In Study 2, through the use of secondary data analysis, the data collections can be assessed more thoroughly in terms of their completeness by examining the amount of missing and unknown data. It will also examine the consistency of the data by examining changes in missing or unknown data over time and differences between key characteristics (e.g., road user types, age, gender, and location) in the amount of missing and unknown data. In Study 3 the accuracy, or more specifically validity, of the data collections will be examined, by comparing variable fields within and between the data collections. This will also include an assessment of the serious injury classification in each data collection. Examining and quantifying these data quality characteristics as part of these three studies will inform and expand on the understanding of the impact of using these data to inform policy and practice in road safety.

RQ3: To what extent are the road crash injury data collections consistent with one another in terms of scope, data classification, and epidemiological profile?

This research question will be examined in Studies 1, 2 and 3 of the research. Firstly, in Study 1, the scope and variable fields of the different data collections will be compared to assess their consistency with each other. In Study 2, a profile of road crash injuries for each data collection will be produced and compared against the road crash data (police data) that is currently relied upon in road safety research. This will allow the researcher to understand how using different data sources may provide a different picture of the road crash injury problem, and thus highlighting quality issues with relying on one source of information. This study will also provide some indication of the under-reporting of these incidents to police in Queensland that has been suggested by research in other jurisdictions (Alsop & Langley, 2001; Amoros et al., 2006; Boufous et al., 2008; Langley, Dow, et al., 2003). Study 3, by using linked data, will expand on the findings of Study 2 by more precisely assessing the concordance between the data collections to further explore the profile differences and under-reporting issues.

RQ4: What are the facilitators of and barriers to linking road crash injury data collections in Queensland and elsewhere?

This research question will be addressed as part of Study 1 and Study 3. In Study 1, interviews will be conducted with expert data users, data custodians, and data linkage experts to explore the perceived benefits and barriers of performing data linkage both generally and specifically in terms of road safety. While the benefits of data linkage, particularly of health data, has been well established, it is less clear why the uptake of data linkage in the road safety sector and in certain jurisdictions has been slow. As part of Study 3, the potential barriers to data linkage in road safety in Queensland will be examined, by assessing the issues in using this methodology for the current research program.

RQ5: What aspects of road crash injury data quality can be improved by using linked data for road safety investigation, intervention development, and evaluation?

This research question will be addressed in Study 3, by examining the profiles of road crash injuries using different linkage combinations of the data collections and comparing these to the unlinked study results of Study 2. Quality assessments will also be conducted with the linked data to determine any improvements to the quality of the information if linked data is used instead of non-linked data. These quality assessments will include completeness, representativeness, and validity (particularly of the classification of serious injury). The process of using linked data to address the issue of serious injury classification and the issue of under-reporting to police is in line with the recommended strategies of the World Health Organisation (2010).

2.7 Chapter Summary

Data is vital to informing policies and interventions designed to reduce the burden of road trauma. It is generally accepted that the relevant epidemiological information cannot be obtained from a single data collection and that linkage of key data collections has the potential to overcome the limitations of single data source and maximize the collective benefit of data relating to road trauma. However, particularly within the context of Queensland, it has not been established as to whether road safety data linkage is feasible and whether linked data provide advantage over non-linked data, both qualitatively and quantitatively. This project aims to assess the quality of current sources of road crash injury data and the linkage opportunities that exist within Queensland in order to provide a more comprehensive picture of road crashes and the resulting injuries. It will also aim to provide recommendations about the feasibility and benefit of data linkage for other jurisdictions within Australia and internationally that do not currently use this methodology.

Chapter Three: Review of Road Crash Injury Data Collections

3.1 Introductory Comments

This chapter outlines Study 1a conducted as part of the research program. It involved a review of legislation and other documentation relating to the relevant road crash injury data collections. It outlines the scope, purpose, governance, data collection procedures, content, access, and timeliness of each of the data collections. In doing so, this study, in combination with Study 1b (Chapter 4), will provide information on the quality of the reviewed data collections in terms of relevance, completeness, consistency, accessibility, and timeliness. It also outlines each data collection's potential for data linkage.

Study Aims and Research Questions

The aim of the current study was to address the research questions below.

RQ1: How well do data collections, which collect road crash injury information in Queensland, conform to the core/minimum requirements for road crash injury data?

> *RQ1a: What is the scope and representation of road crash injuries for the data collections which collect road crash injury information in Queensland?*
>
> *RQ1b: How well do the data collections comply with the core/minimum data elements as outlined by the guidelines discussed in the literature review (e.g., WHO, Austroads, NHTSA)*

RQ2: What are the strengths and weaknesses of each of the road crash injury data collections within the context of road safety investigation, intervention development, and evaluation?

> *RQ2a: What is the completeness of the data collections in terms of the inclusion of the core/minimum data set variables?*
>
> *RQ2b: How consistent are the data collections over time?*
>
> *RQ2c: What quality assurance and coding practices are used by the data collections and how does this impact on their accuracy?*
>
> *RQ2d: What are the protocols for gaining access to the data collections?*
>
> *Rq2e: What are the delays in data being available for research?*

RQ3: To what extent are the road crash injury data collections consistent with one another in terms of scope, data classification, and epidemiological profile?

> *RQ3a: How does the scope for included cases compare across data collections?*
>
> *RQ3b: What data fields do the data collections have in common with each other?*

3.2 Method

3.2.1 *Review of legislation and documentation*

Legislation and other documentation relating to the relevant data collections were identified by the data custodians during the interviews (see Chapter 4, Section 4.3.1) as well as through internet searches for "Queensland privacy" and each of the data collections names (full name and acronym). The website of the government agency responsible for the data collection was also searched for relevant documentation. Documentation found included manuals, data dictionaries, and web page text. Legislation was sourced from the relevant Queensland (https://www.legislation.qld.gov.au) and Australian Government (www.comlaw.gov.au*)* websites. These documents were reviewed to obtain information on the scope, purpose, collection and coding methods, and content of each data collection in terms of the WHO injury surveillance (Holder et al., 2001), Austroads (1997), and MMUCC (National Highway Traffic Safety Administration, 1998a) minimum data sets, as well as the related legislation, governance, and access protocol.

In order to gain a complete understanding of the collection, cleaning, and coding of the Queensland Road Crash Database, discussions took place with staff from the State Traffic Support Branch and Forensic Crash Unit, Queensland Police (QPS), and staff from the Office of Economic and Statistical Research (OESR). The discussions with staff of the QPS were covered by QUT HREC approval and QPS ethics approval.

3.3 Results

3.3.1 *Summary of data sources*

3.3.1.1 *Queensland Road Crash Database (QRCD)*

Scope

The QRCD stores information relating to all police reported crashes in Queensland since 1986. The definition of a crash that should be recorded in QRCD is:

> "*a crash that has been reported to the police which resulted from the movement of at least one road vehicle on a road and involving death or injury to any person, or property damage. Note also that to qualify as valid, crashes must meet the following criteria:*
>
> - *the crash occurs on a public road, and*
> - *a person is killed or injured, or*
> - *the value of the property damage is:*
>
> *(a) $2500 to property other than vehicles (after 1 December 1999)*
> *(b) $2500 damage to vehicle and property (after 1 December 1991 and prior to 1 December 1999)*
> *(c) value of property damage is greater than $1000 (prior to December 1991) or;*

- *at least one vehicle was towed away."* Department of Transport and Main Roads (2010)

The following major exclusions apply:

- The incident occurs in an area outside the road or road related area.
- There is no moving vehicle involved.
- The incident is not attributable to vehicle movement.

Also, in cases where a person was involved in a crash by attempting suicide or from a medical condition, the crash will only be included if there was subsequent involvement of another person. For example, if a driver has a heart-attack and collides with a pedestrian, the crash will be included, the pedestrian will be included, but the driver with the heart-attack will not be. It should be noted that for a police-reported crash it is implicit that the crash has been reported to police and recorded by police.

Purpose

The primary purpose of data collected in QRCD is to provide information to decision makers in order for them to develop treatments or countermeasures for particular crash types, road user groups, vehicle types, or road characteristics.

Data governance

The QRCD is housed within the Data Analysis Unit (DAU) at the Department of Transport and Main Roads (TMR). TMR fund and own the database itself, including the costs of data cleaning and coding performed by the Office of Economic and Statistical Research (OESR). They do not fund the collection of data by the Queensland Police Service (QPS).

Data collection

The data within the QRCD are collected by the Queensland Police Service (QPS) either at the location of the crash or reported at a police station. Details of the crash are recorded on the Traffic Crash Report Form (PT51) (see Appendix B).

The information on the PT51 is entered into the Queensland Police Service information management system (QPRIME) by the reporting officer. The data is entered usually at the end of the reporting officer's shift; however delays can occur for a number of reasons, including the performance of other duties. At the time of the initial data entry, there may be a significant proportion of information unavailable (e.g., witness reports, BAC). These data are modified by the reporting officer when available.

Most of the data are transferred weekly from QPRIME to the Queensland Road Crash Database. Notification of and selected details for fatal crashes is sent via email on a daily basis from QPS to TMR. When required, additional data can be obtained from CITEC Confirm (an online crash reporting system that is accessible by OESR), directly from

QPS reporting officers, or via TRAILS (for licensing and registration related information). There are some data that are not loaded directly into the database by police and must be entered manually following the guidelines set out in the Queensland Road Crash Database Manual.

The process by which crashes are recorded in the QRCD is detailed in Figure 3.1.

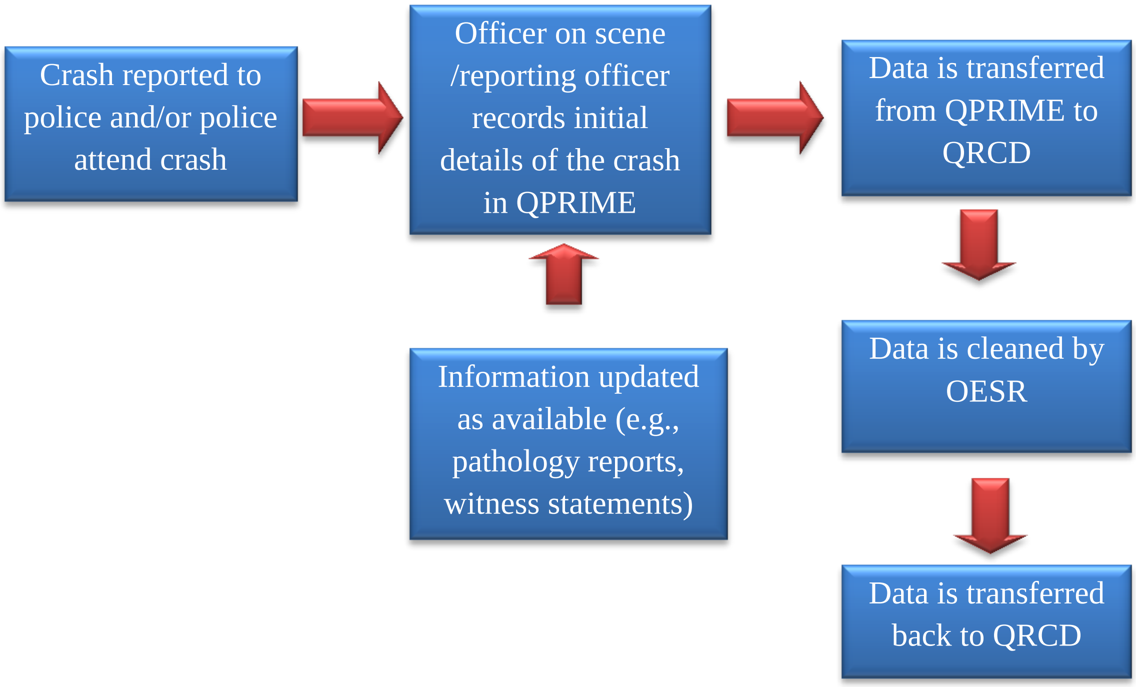

Crash reported to police and/or police attend crash
Officer on scene /reporting officer records initial details of the crash in QPRIME
Data is transferred from QPRIME to QRCD
Information updated as available (e.g., pathology reports, witness statements)
Data is cleaned by OESR
Data is transferred back to QRCD

Figure 3.1: Flow chart for recording of crash data in QRCD

Data cleaning and coding

The data are subjected to a series of validation checks conducted by the Office of Economic and Statistical Research. These checks are in the form of both clerical and computerised checks and are designed to check for completeness, accuracy and consistency of information that has been supplied by the Queensland Police Service. A report, giving details of those crash records that fail any validation checks is generated as required. Clerical intervention is then necessary to resolve discrepancies in relation to the crash to ensure data are 'clean' prior to finalisation and release. Some variables are coded by police by filling out the PT51 form. Other variables are coded by staff at OESR using information in text descriptions and diagrams. The coding of these data is based on a coding manual developed by OESR, TMR, and QPS. It should also be noted, that when requests for data are fulfilled by DAU, further coding or re-coding may occur to fit with the need of the requesting party or to comply with legislation. The details of the relevant coding conventions will be presented below in the content section.

Content of QRCD

The unique identifier in the QRCD is applied to the crash (crash number), so the database is essentially event-based. However, the crash number is also applied to all the controllers (drivers, riders, cyclists, and pedestrians) in a crash and casualties resulting from a crash (injured persons). Also, each controller and casualty involved in each crash are given a number from one through to however many controllers or casualties are involved. This allows for identification of all individuals injured in crashes and also allow for the connections between crash circumstances, controllers (drivers, riders, pedestrians), units (vehicles), and casualty characteristics to be explored.

As shown in Table 3.1, the QRCD includes all data outlined in the Core Minimum, Core Optional, and Supplemental datasets with the exception of nature of activity. Some elements of the Core Minimum dataset are not variable fields in QRCD, but are included by definition. Specifically, based on the scope of the data collection, intent (only unintentional), place (only road or road-related area), mechanism (all traffic injury), and broad external cause (motor-vehicle traffic accidents) are specified. It should be noted however, that some variables are either not available to researchers at all, or have limited availability. This is due to either privacy restrictions (CRN) or an established lack of reliability in the variable field (ethnicity and Indigenous status).

Table 3.1: QRCD compatibility with WHO guidelines core MDS, core ODS, and supplemental data sets

	WHO variable	QRCD variable/s
Core minimum	Unique person/event identifier	Customer Reference Number (CRN) /Crash number
	Age of injured person	Casualty age
	Sex of injured person	Casualty gender
	Intent	BY DEFINITION
	Place	BY DEFINITION
	Nature of activity	-
	Mechanism	BY DEFINITION
	Nature of injury	Injury Description
	External cause	BY DEFINITION, Unit type[1], road user group[1]
Core optional	Race or ethnicity of injured person	Ethnicity[2], Indigenous status[2]
	Date of injury	Crash date
	Time of injury	Crash time
	Residence of injured person	Unit origin town
	Severity of injured person	Casualty severity
	Alcohol use	Contributing factors (alcohol involvement)
	Other psychoactive substance use	Contributing factors (alcohol and drug involvement)
	Narrative	Text description[2]
Supplemental	Mode of transport	Unit type
	Road user	Road user group
	Counterpart	Unit type

[1] Not ICD-10AM coded

[2] Not generally available to researchers

Despite most of the variables required for a Core Minimum dataset being present in QRCD, some of the coding of these variables are either not coded to an international standard (e.g., ICD 10) or lack specific detail. For example, there are two forms of injury description in the database which could be used to determine *nature of injury*, one is a coded injury description completed by OESR and the other is an injury text description that comes directly from police. OESR code injuries using an ICD based coding system, when the police have mentioned an injury in the general text field but have not completed the injury text description, or at least one person in the crash died. For all other cases, the injury description variable is coded as ‘refer to text description’ (099). In these cases, the police injury text description, which is not coded, is the only source of information about injuries. As a result, there may be a large proportion of cases in which there is insufficient information to draw any conclusions about the nature of injury in this database. Another example is the injury severity variable field. This field complies with the international definition of a fatal injury in that it is an injury that results in death within 30 days of a crash (WHO, 2010). However, the QRCD does not currently comply with the international definition of a hospitalised injury, in that it does not just include cases in which an injured person is admitted to hospital for 24 hours or more (WHO, 2010), since it includes all cases where a person was transported to hospital, regardless of their admission status or length of stay.

For the purposes of linkage, the QRCD does not include a unique identifier that is shared with any other government agency, which would preclude a simple matching of data. It does however, include name, address, date of birth, and date of crash. These variables would allow probabilistic linkage with other data collections that also have this identifying information.

Beyond the WHO guidelines, the QRCD includes other variables that would be of importance to road safety research, policy and practice. It also complies with the minimum datasets outlined by WHO (2010), Austroads (1997), and the MMUCC (National Highway Traffic Safety Administration, 1998a). Specifically, QRCD includes: the exact location of a crash recorded as GPS co-ordinates; the posted speed limit; Blood/Breath Alcohol Content (BAC) of tested drivers; seating position; licence status; and the culpability (most at fault status) of an individual involved in a crash.

Legislation relating to QRCD

There are two key pieces of legislation relating to data held by the Department of Transport and Main Roads. The first of these is the Queensland's *Information Privacy Act, 2009*. This act applies to all data collected and held by government departments in Queensland and therefore is an act that applies to each of the data collections in this thesis. Within the *Information Privacy Act, 2009* are the Eleven Information Privacy Principles (IPPs). These principles allow for the sharing of this information with other government agencies or other external persons under certain circumstances. Information Principle 11 outlines the disclosing of information for research purposes and specifies that if it is necessary for research, does not involve the publication of identifying information, and obtaining consent is not practicable, then release of the data is permissible.

Section 77A of the Queensland *Transport Operation (Road Use Management) Act, 1995* allows for the provision of data to researchers if consent is provided (using an approved form) by the person to which the information relates. It also allows for the release of driver licence or traffic history information for approved research purposes without consent as long as the information does not identify a person in anyway.

While the Queensland *Transport Operation (Road Use Management) Act, 1995* makes no direct or specific reference to road crash data, it is possible that reference to 'traffic history' under s77A could be interpreted as including involvement in a crash.

Access to QRCD

While there is no specific requirement for ethics approval to request data from TMR, ethics is required by researchers within a university context to gain access to the data. Following ethics approval, an application can be made to the Data Analysis Unit, within TMR using the crash data request form (see Appendix C). The Data Analysis Unit assesses the request, and if approved, provides the data to the researchers in comma separated variable (.csv) files.

The release of data is based on compliance with both the *Information Privacy Act, 2009* and the *Transport Operation (Road Use Management) Act, 1995* described above. In complying with the *Transport Operation (Road Use Management) Act, 1995*, only the release of de-identified information is possible for research purposes. The data provided must not only be de-identified in the form of removal of names, addresses and date of birth, it must also be unable to potentially identify involved persons or their crash. Some variables in combination are considered potentially identifying and are not approved for release (e.g., postcode and age in years). As described above, data may be re-coded by DAU to prevent individuals from potentially being identified, such as collapsing categories or assigning higher level categorisations (e.g., assigning ARIA+ classification instead of postcode).

Another release mechanism for crash data held by TMR is through consent from the person to which the data relates. TMR have a consent form that participants in research projects and the chief investigator can complete to provide permission for TMR to release the participants' crash, licensing, and/or traffic offence histories for research purposes. Once these consent forms are completed and provided to TMR, the researcher can then make a request using the same procedure described above.

Another avenue for accessing elements of the data in QRCD is via Webcrash 2.3 platform. Webcrash is a subscription based online database. Access requires approval from the DAU at TMR, with approved users being provided with a unique username and password to log on to the website. Not all information is available in Webcrash for privacy reasons; also unit record data is restricted to a limit of 500 cases. Aggregate or unit record reports are produced based on queries in the form of Excel, text, or pdf.

The release of identifying information to researchers for the purposes of linkage is not currently possible unless consent is provided by the individual (see *Information Privacy Act 2009*). The release of identifying information to other government agencies for the purposes of linkage is possible with a Memorandum of Understanding between the relevant agencies. As a result of negotiations for the completion of this research project, TMR and Queensland Health (QH) signed an MOU allowing for TMR to provide identifying information (name, address, date of birth, date of crash etc.) to QH for the purposes of linking with data QH hold (e.g., Emergency Department Information System). The MOU only allows for the release of the identifying information required for linkage and does not allow the sending of any 'content' (specific details of the crash) information to external agencies. The MOU extends beyond the current project to allow researchers in the future to also access linked data if prescribed conditions are met. The process for the sharing and linking of data is described in detail in Chapter 5.

Timeliness

As discussed previously, while there are limited delays in terms of the reporting of crashes from police to TMR, there can be delays for some of the information relating to the crash. It takes time to gather witness statements, alcohol/drug test results, and investigate the circumstances of a crash. Also, once this information becomes available it then needs to be cleaned and, for some variables, coded by OESR. This process involves following up with police or ‘CITEC Confirm’ when variables are incomplete and/or inconsistent with other variables.

The availability of ‘complete’ data varies depending on the severity of the crash. The cleaning and finalising of fatal crashes are given the highest priority, with hospitalisations second. As a result, the reporting, cleaning, coding of fatal crashes can currently take up to 9 months, ‘hospitalised’ up to12 months, and approximately 2 years for the lower severity crashes (i.e., medically treated, minor injury, and property damage only).

Metadata

QRCD has a publicly available glossary that includes data definitions and scope information (Transport and Main Roads, 2012). It also has information within its publications about data quality issues (e.g., 2009 Road Traffic Crashes in Queensland, Transport and Main Roads, 2012).

3.3.1.2 *Queensland Hospital Admitted Patients Data Collection (QHAPDC)*

Scope

QHAPDC contains data on all patients discharged, statistically separated, died, or transferred from a Queensland hospital permitted to admit patients (including public hospitals, licensed private hospitals, and day surgery units). According to the QHAPDC manual, generally “a patient can be admitted if one or more of the following apply:

- The patient’s condition requires clinical management and/or facilities are not available in their usual residential environment.
- The patient requires observation in order to be assessed or diagnosed.
- The patient requires at least daily assessment of their medication needs.
- The patient requires a procedure(s) that cannot be performed in a stand-alone facility, such as a doctor’s room, without specialised support facilities and/or expertise being available.
- There is a legal requirement for admission (eg. under child protection legislation).
- The patient is aged nine days or less.” (Queensland Health, 2012, p. 32)

Purpose

Under the National Healthcare Agreement (NHA) between the Australian government and the State of Queensland, hospitals permitted to admit patients must provide information about admissions to QHAPDC. These data are used for a number of purposes including monitoring funding arrangements, requesting additional funding, epidemiological study (morbidity and mortality), education of students of medicine, nursing, and allied health.

Data governance

QHAPDC is housed on a secure server within the Health Statistics Centre (HSC), under the governance of Queensland Health.

Data collection

Data is collected in each of the facilities included in the collection. Data is collected in two ways depending on the hospital, either the Hospital Based Corporate Information System (HBCIS) or a paper based system (Identification and Diagnosis Sheets and Patient Activity Form). HBCIS data are extracted and mapped to the Data Collections Unit requirements, the translation of which is outlined in the QHAPDC manual. Data is collected monthly in unit record form. If forms are used, they are sent to the Area Health Service to be converted into approved electronic format and then forwarded to the Data Collections Unit (HSC). HBCIS data is sent directly to the Data Collections Unit (HSC).

Different elements of the data are collected by different staff. Admitting staff collect the following:

- Unique Record ID
- Facility name and number
- Queensland Ambulance number (eARF number)
- Admission date
- Admission time
- Date of birth
- Sex
- Patient family and given names
- Patient address
- Compensable status
- Country of birth
- Indigenous status
- Nature of injury

Discharge staff complete the following:

- Separation date
- Separation time

- Mode of separation

Medical practitioners complete the following:

- Principal diagnosis
- External cause; place of occurrence

Data cleaning and coding

Data is coded at the facility as well as at the Data Collections Unit. At the facility, trained data coders code clinical details using the current version of the ICD-10-AM. At HSC, data may be coded in different ways for the release of data to external parties (e.g., collapsing categories to prevent possible identification, assigning ARIA+ classifications).

The HSC checks for errors including valid values, logical consistency, and historical consistency. Validation reports are produced for the hospital, in which the hospital will make corrections and resubmit to HSC. A record of these procedures conforms to the Australian Classification of Health Interventions. Data can be modified by the hospital up to September of the year after the financial year to which the data relates.

It should also be noted, that when requests for data are fulfilled by HSC, further coding or re-coding may occur to fit with the need of the requesting party or to comply with legislation. The details of the relevant coding conventions will be presented below in the content section.

Content of QHAPDC

A facility unique ID (FUR number) is assigned to each episode of care (within each facility). The data collection is episode based rather than based on individuals. However, in each facility a patient will also be assigned a unique ID (UR number) that they keep for that facility. This allows within a facility for an episode and a person to be tracked through the system. However, as the UR is only unique for one facility, it is not possible to track an individual across hospitals using any unique ID. Within HSC, probabilistic data linkage is performed to identify individuals across different episodes and facilities. Generally however, this form of the data (individually linked) is not provided for external use and counts are based on episodes not patients.

QHAPDC includes almost all of the Core Minimum, Core Optional, and Supplemental data as outlined by WHO (Holder et al., 2001), with the exception of a narrative variable. The external cause, activity, place, and diagnosis strings are ICD-10-AM coded. Variables relating to location (i.e., Statistical Local Area and ARIA+) are coded using the Australian Bureau of Statistics Australian Statistical Geography Standard (ABS, 2001). While some variables are not generally made available to researchers due to privacy restrictions, some variables can be recoded to a higher level for release to reduce the potential identification of a person (e.g., address of usual residence coded into ARIA+, date of admission coded into day of week, month of year, and year). It should be noted that the time and date is for admission rather than injury. It is possible that admission could occur substantially later than when the injury occurred (e.g., emergency response, hospital waiting, delay in presenting by the injured person). The details of the correspondence between QHAPDC and the WHO guidelines are shown in Table 3.2.

Table 3.2: QHAPDC compatibility with WHO guidelines core MDS, core ODS, and supplemental data sets

	WHO variable	QHAPDC variable/s
Core minimum data set	Unique person identifier	UR number[1]
	Age of injured person	Age
	Sex of injured person	Sex
	Intent	External cause string
	Place	Place string
	Nature of activity	Activity string
	Mechanism	External cause string
	Nature of injury	Diagnosis string
	External cause	External cause string
Core optional data set	Race or ethnicity of injured person	Indigenous status[1], Country of birth, South-Sea Islander status[1]
	Date of injury	Date of admission[1]
	Time of injury	Time of admission[1]
	Residence of injured person	Address of usual residence[1], Statistical Local Area[1], ARIA+
	Severity of injured person	Diagnosis string, length of stay
	Alcohol use	External cause string
	Other psychoactive substance use	External cause string
	Narrative	-
Supplemental data	Mode of transport	External cause string
	Road user	External cause string
	Counterpart	External cause string

[1] Not generally available to researchers

While QHAPDC includes a place variable, this is restricted to a broad classification that would, at most, be able to identify cases as fitting the definition of a road crash. It is not specific enough to give an indication of the location the incident occurred. It may be possible to use the ARIA+ of the hospital or the usual residence to make some claims about location at a broader level (i.e., rural and remote factors).

There is no variable field or code that can be used to determine fault within QHAPDC and no other items (except those previously mentioned) that comply with the minimum data requirements outlined by WHO (2010), Austroads (1997), and the MMUCC (2012).

For the purposes of linkage, the QHAPDC does not include a unique identifier that is shared with any other government agency, which would preclude a simple matching of data. As mentioned previously, the UR number is also not common across facilities within the collection. It may be possible however, to link the UR number within a facility from the emergency department (EDIS and/or QISU). Also, QHAPDC, since 2009, has included the eARF number which relates to the Queensland Ambulance data (although it is not known how consistently this is recorded). Despite these similarities in unique IDs, probabilistic linkage would still be required, in combination with other identifying variables (e.g., name, address, DOB etc.) as it may not always be recorded well enough for direct matching.

Legislation relating to QHAPDC

Legislation covering the confidentiality of the QHAPDC is covered by Part 7 of the *Health and Hospitals Network Act, 2011* (Qld) and the *Private Health Facilities Act, 1999* (Qld) s. 147. Release of information from QHAPDC is also governed by the *Public Health Act, 2005* (Qld) and the *Information Privacy Act, 2009* (Qld).

Under the *Health and Hospitals Network Act, 2011* (Qld) s. 144, the release of confidential information is allowed for provided there is consent from the person to which the information relates. However, the *Health and Hospitals Network Act, 2011* (Qld) does not exclude the release of information as required by another Act or law. For the purposes of this Act, the definition of confidential information is as follows:

> *"confidential information means information, acquired by a person in the person's capacity as a designated person, from which a person who is receiving or has received a public sector health service could be identified."* (*Health and Hospitals Network Act, 2011* (Qld) s. 139)

Under the *Private Health Facilities Act, 1999* (Qld) s. 147, personal information may not be disclosed unless, consent is obtained, or the Chief Executive is satisfied that the release of data is in the public interest.

For the purposes of this act, personal health information means:

> *"information about a person's health that identifies, or is likely to identify, the person."* (*Private Health Facilities Act, 1999* (Qld) s. 147)

The *Public Health Act, 2005* (Qld) s. 283 allows for the application for the release of information using a Public Health Act Application. In order to receive approval, the *Public Health Act, 2005* (Qld) s.282 states that the research must be in public interest (balanced against the privacy of individuals) and identification of individuals is necessary.

Access to QHAPDC

The first step in gaining access to QHAPDC data is to apply for Human Research Ethics Committee approval. This approval can be from a university committee or the Queensland Health Human Research Ethics Committee. If the nature of the request does not require access to identifying information or any data that will specifically target at-risk populations (e.g., illegal behaviour, Aboriginal or Torres Strait Islanders, or people with mental illness), then a low risk ethics application would usually apply.

Following ethics approval, it is necessary to discuss the request with the data custodian so that they can advise on the requests suitability and feasibility. During this process, the researcher is required to also complete a Public Health Act (PHA) application (to comply with legislation) outlining the aims, benefits of the research, methods, and requested data (specifications of included cases and variable fields). It should be noted, that while under the legislation, a PHA is not required for access to de-identified data, it is often still required as it facilitates the data request and allows the custodians to have a direct role in the approval process.

Once the PHA is complete following discussions with the custodian, the custodian signs the PHA and the PHA is sent to the Director-General of Queensland Health for approval. When approval is received, the researcher notifies the custodian and the data is prepared for release. The data is released in text (.txt) format in a password protected zip folder that is put on CD to be collected in person by the researcher.

For the purposes of data linkage, access to the information required could be gained using a similar procedure as that described above for access to de-identified data. The exception to this is that a National Ethics Application Form (NEAF) would need to be completed as the research would not be considered low risk with identifying information included. This process would only be required if a third party (i.e., the researcher or someone other than a QH data custodian) was conducting the linkage. Currently in Queensland, QH has a dedicated data linkage unit to perform the linkage, so this process is not necessary. More detail about the data linkage process will be described in Chapter 6, Section 6.4.

Timeliness

Data is required to be sent to HSC within 35 days after the month of separation. Data is subject to validation checks by HSC and reports are sent back to hospitals for correction. This process of submission, validation, correction and re-submission can take up to 8 weeks. The data is not considered final, and therefore able to be released to external parties, for several months after the end of the financial year in which the episode occurred. Once the data has been submitted to the Commonwealth it is considered final.

Metadata

QHAPDC has a publicly available coding manual (Queensland Health, 2012) and reports on its quality (Queensland Health, 2012). Information on ICD-10-AM coding (on which many of the variables within QHAPDC are based) is also available (National Centre for Classification in Health, 2012).

3.3.1.3 *Emergency Department Information System (EDIS)*

Scope

The Emergency Department Information System (EDIS) includes all emergency department presentations in the following 29 hospitals across Queensland:

- Beaudesert Hospital
- Bundaberg Hospital
- Caboolture Hospital
- Cairns Base Hospital
- Caloundra Hospital
- Gladstone Hospital
- Gold Coast Hospital
- Gympie Hospital
- Hervey Bay Hospital
- Innisfail Hospital
- Ipswich Hospital
- Logan Hospital
- Mackay Base Hospital
- Maryborough Hospital
- Mt Isa Base Hospital
- Nambour Hospital
- Prince Charles Hospital
- Princess Alexandra Hospital
- QEII Jubilee Hospital
- Redcliffe Hospital
- Redlands Hospital
- Robina Hospital
- Rockhampton Hospital
- Royal Brisbane Hospital
- Royal Children's Hospital
- Toowoomba Base Hospital
- Townsville Hospital
- Wynnum Hospital
- Yeppoon Hospital

Purpose

The system is used to monitor a patient's progress through the ED system. It provides alerts and records treatment details.

Data governance

The database is held and governed within Health Service and Clinical Innovation Division in Queensland Health.

Data collection

The triage nurse enters information into EDIS for each patient that presents at a participating emergency department. Information is added and updated by ED clerical staff, ED nurses and ED doctors, throughout a patient's episode of care in the ED.

Data cleaning and coding

Some data fields are coded to NDS-IS and ICD-10-AM standards (more details below) and selected using drop down menus within the system. Data managers check the patient's written record against their record in EDIS for any discrepancies and if there are any, they are updated in EDIS. There is no additional coding or cleaning conducted.

Content of EDIS

As with QHAPDC and QISU, a facility unique ID (FUR number) is assigned to each episode of care (within each facility) included in the EDIS collection. The data collection is episode based rather than based on individuals. However, in each facility a patient will also be assigned a unique ID (UR number) that they keep for that facility. This allows within a facility for an episode and a person to be tracked through the system. However, as the UR is only unique for one facility, it is not possible to track an individual across hospitals using any unique ID.

In terms of the core MDS, there is information on age, sex, and nature of injury. There is however, no variables directly related to intent, place, activity, or mechanism. For core ODS, there are variables for date and time, residence, severity, and a narrative. There is no supplemental data set variables included in EDIS (see Table 3.3). It should also be noted that, as with QISU, the time and date is for presentation rather than injury. It is possible that presentation could occur substantially later than when the injury occurred (e.g., emergency response, delay in presenting to hospital by the injured person).

Table 3.3: EDIS compatibility with WHO guidelines core MDS, core ODS, and supplemental data sets

	WHO variable	EDIS variable/s
Core minimum data set	Unique person identifier	UR number[1]
	Age of injured person	Age
	Sex of injured person	Sex
	Intent	-
	Place	-
	Nature of activity	-
	Mechanism	-
	Nature of injury	Diagnosis ICD code
Core optional data set	Race or ethnicity of injured person	-
	External cause	-
	Date of injury	Arrival date
	Time of injury	Arrival time
	Residence of injured person	Address of usual residence, Postcode
	Severity of injured person	Triage priority/departure status/diagnosis ICD code
	Alcohol use	-
	Other psychoactive substance use	-
	Narrative	Presenting problem
Supplemental data	Mode of transport	-
	Road user	-
	Counterpart	-

[1] Not generally available to researchers

EDIS contains no information on the exact location of the injury. However, as with QISU, it may be possible to use the location of the hospital (hospital name) or the postcode of usual residence to make some claims about location at a broader level (i.e., rural and remote factors). EDIS includes no other items (except those previously mentioned) that comply with the minimum data requirements outlined by WHO (2010), Austroads (1997), and the MMUCC (2012).

For the purposes of linkage, the EDIS does not include a unique identifier that is shared with any other government agency, which would preclude a simple matching of data. As mentioned previously, the UR number is also not common across facilities within the collection. It may be possible however, to link the UR number within a facility from QHAPDC and the emergency department (EDIS). Also, since 2009, EDIS has included the eARF number which relates to the Queensland Ambulance data (although it is not known how consistently this is recorded). Despite these similarities in unique IDs, probabilistic linkage would still be required, in combination with other identifying variables (e.g., name, address, DOB etc.), as it may not always be recorded well enough for direct matching.

Legislation relating to EDIS

The release of information from EDIS is covered by the *Public Health Act, 2005* (Qld) and the *Information Privacy Act*, *2009* (Qld). The sections of the legislation that are relevant have already been described in a previous section (3.4.1.2).

Access to EDIS

As with the other QH based data collections, the first step in gaining access to EDIS data is to apply for Human Research Ethics Committee approval. This approval can be from a university committee or the Queensland Health Human Research Ethics Committee. As with QISU and QHAPDC, if the nature of the request does not require access to identifying information or any data that will specifically target (e.g., illegal behaviour, Aboriginal or Torres Strait Islanders, or people with mental illness), then a low risk ethics application would usually apply.

Following ethics approval, it is necessary to discuss the request with the data custodian so that they can advise on the requests suitability and feasibility. During this process, the researcher is required to also complete a Public Health Act application (to comply with legislation) outlining the aims, benefits of the research, methods, and requested data (specifications of included cases and variable fields) (see above for legislative requirements).

Once the PHA is complete following discussions with the custodian, the custodian signs the PHA and the PHA is sent to the Director-General of Queensland Health for approval. When approval is received, the researcher notifies the custodian and the data is released in excel (.xlsx) format in a password protected zip folder that is collected by the researcher.

For the purposes of data linkage, access to the information required could be gained using a similar procedure as that described above for access to de-identified data. However, as with QHAPDC and QISU, if the data linkage is not conducted within QH, a NEAF would be required for the researcher or other party to access identifying information.

Timeliness

Following the data being recorded in the database, coded, and cleaned, data are generally available 3 to 6 months from a person presenting at the ED.

Metadata

There is very little publicly available information on EDIS. Information on the ICD-10-AM coded diagnosis variable can be accessed (National Centre for Classification in Health, 2012).

3.3.1.4 *Queensland Injury Surveillance Unit (QISU)*

Scope

The Queensland Injury Surveillance Unit collects data on injuries presenting at Queensland emergency departments. It currently collects information from the following 17 hospitals:

- Bundaberg Hospital
- Cherbourg Hospital
- Clermont Hospital
- Collinsville Hospital
- Dysart Hospital
- Hughenden Hospital
- Innisfail Hospital
- Mackay Hospital
- Maryborough Hospital
- Mater Children’s Public Hospital
- Mater Hospital Mackay
- Moranbah Hospital
- Mount Isa Hospital
- Proserpine Hospital
- Royal Children’s Hospital
- Sarina Hospital
- Yeppoon Hospital

Purpose

The primary purpose of QISU is to monitor injuries of all types and for all ages in Queensland through data collection in a sample of hospitals.

Data governance

The database is held and governed within the Queensland Injury Surveillance Unit, in the Healthy Living Branch of Queensland Health.

Data collection

There are three ways in which data is collected for QISU. The first of these is through the Emergency Department Information System (EDIS). In participating hospitals, an injury surveillance screen is activated in EDIS when either the triage nurse indicates that an injury has occurred or when an ICD-10-AM diagnosis code for injury (S00-T98) is entered. Another way the data is collected is via the Hospital Based Clinical Information System (HBCIS). There is a facility within participating hospitals to collect additional text information when triggered by the entry of ICD-10-AM diagnosis code for injury (S00-T98). Finally, data is also collected using a paper-based system to collect additional injury information required for the database. Regardless of the collection method, demographic information and Level 2 National Data Standards for Injury Surveillance (NDS-IS, National Injury Surveillance Unit, 1998) is included. In whatever form the data is collected it is sent through to QISU after being entered or imported into the InjuryEzy database.

Data cleaning and coding

The data are cleaned and coded (for text descriptions) in accordance with the NDS-IS standards by trained coders within QISU. The data are then exported to an SQL database for interrogation and/or extraction.

Content of QISU

As with QHAPDC, a facility unique ID (FUR number) is assigned to each episode of care (within each facility) included in the QISU collection. The data collection is episode based rather than based on individuals. However, in each facility a patient will also be assigned a unique ID (UR number) that they keep for that facility. This allows within a facility for an episode and a person to be tracked through the system. However, as the UR is only unique for one facility, it is not possible to track an individual across hospitals using any unique ID.

All of the Core minimum data set and Supplemental data set variables are included in QISU. With the exception of race, alcohol use, and other psychoactive substance use, all variables from the Core optional data set are also included (see Table 3.4). All of the included variables, with the exception of those relating to the severity and nature of injury, are not ICD-10-AM coded although they are coded according to NDS-IS standards (NISU, 1998). It should be noted that the time and date is for presentation rather than injury. It is possible that presentation could occur substantially later than when the injury occurred (e.g., emergency response, delay in presenting to hospital by the injured person).

Table 3.4: QISU compatibility with WHO guidelines core MDS, core ODS, and supplemental data sets

	WHO variable	QISU variable/s
Core minimum data set	Unique person identifier	UR number[1]
	Age of injured person	Age
	Sex of injured person	Sex
	Intent	Intent
	Place	Place
	Nature of activity	Activity
	Mechanism	Mechanism
	Nature of injury	ICD code; ICD description
	External cause	External definition
Core optional data set	Race or ethnicity of injured person	-
	Date of injury	Arrival date
	Time of injury	Arrival time
	Residence of injured person	Postcode
	Severity of injured person	Triage score, ICD diagnosis code
	Alcohol use	-
	Other psychoactive substance use	-
	Narrative	Injury text description
Supplemental data	Mode of transport	External cause
	Road user	External cause
	Counterpart	Major injury factor

[1] Not generally available to researchers

While QISU includes a place variable, this is restricted to a broad classification that would, at most, be able to identify cases as fitting the definition of a road crash. It is not specific enough to give an indication of the location the incident occurred. It may be possible to use the location of the hospital (hospital name) or the postcode of usual residence to make some claims about location at a broader level (i.e., rural and remote factors).

There is no variable field or code that can be used to determine fault within QISU. However, some of these characteristics may be able to be identified in the narrative text field. QISU includes no other items (except those previously mentioned) that comply with the minimum data requirements outlined by WHO (2010), Austroads (1997), and the MMUCC (2012).

For the purposes of linkage, the QISU does not include a unique identifier that is shared with any other government agency, which precludes a simple matching of data. As mentioned previously, the UR number is also not common across facilities within the collection. It may be possible however, to link the UR number within a facility from QHAPDC and the emergency department (EDIS). Also, since 2009, QISU has included the eARF number which relates to the Queensland Ambulance data (although it is not known how consistently this is recorded). Despite these similarities in unique IDs, probabilistic linkage would still be required, in combination with other identifying variables (e.g., DOB etc.) as it may not always be recorded well enough for direct matching. It should be noted that QISU does not include identifiers such as name and address.

Legislation relating to QISU

The release of information from QISU is covered by the *Public Health Act, 2005* (Qld) and the *Information Privacy Act, 2009* (Qld). The sections of the legislation that are relevant have already been described in a previous section (3.4.1.2).

Access to QISU

The first step in gaining access to QISU data is to apply for Human Research Ethics Committee approval. This approval can be from a university committee or the Queensland Health Human Research Ethics Committee. As with QHAPDC, if the nature of the request does not require access to identifying information or any data that will specifically target (e.g., illegal behaviour, Aboriginal or Torres Strait Islanders, or people with mental illness), then a low risk ethics application would usually apply.

Following ethics approval, it is necessary to discuss the request with the data custodian so that they can advise on the requests suitability and feasibility. During this process, the researcher is required to also complete a Public Health Act application (to comply with legislation) outlining the aims, benefits of the research, methods, and requested data (specifications of included cases and variable fields) (see above for legislative requirements).

Once the PHA is complete following discussions with the custodian, the custodian signs the PHA and the PHA is sent to the Director-General of Queensland Health for approval. When approval is received, the researcher notifies the custodian and completes an online data request form. The data is then prepared for release. The data is released in excel (.xlsx) format in a password protected zip folder that is downloaded from a secure web-based file share.

For the purposes of data linkage, access to the information required could be gained using a similar procedure as that described above for access to de-identified data. However, as with QHAPDC, if the data linkage is not conducted within QH, a NEAF would be required for the researcher or other party to access identifying information.

Timeliness

Taking into account the time taken to receive the data from the EDs, code any text data, clean, and finalise for inclusion and release, data is usually available between 3 and 6 months from the date of a case presenting at the ED.

Metadata

There is some information about the content and coding of QISU on their website (http://www.qisu.org.au), including the scope of the data and the included hospitals. Also, there is extensive coding information for the data collection in the NDS-IS (NISU, 1998) on which the coding in QISU is based, as well as the ICD-10-AM coding manual for the diagnosis code (National Centre for Classification in Health, 2012).

3.3.1.5 *Electronic Ambulance Report Form (eARF)*

Scope

The data covers all Queensland Ambulance call-outs across Queensland from 2007.

Purpose

The primary purpose of the eARF is to assist with patient care and quality assurance.

Data governance

The data is held within the Information Support Unit (ISU) of the Queensland Ambulance Service (QAS). The Emergency Services Commissioner provides approval for access.

Data collection

The eARF is completed by QAS officers for all ambulance responses. Some data is collected at dispatch (e.g., place, some patient details). The remaining data is collected at the scene by ambulance officers using a dedicated electronic tablet. This data is then uploaded into the database at the end of the shift. The collection form includes both coded selections and free text. It should be noted that not all fields are mandatory to complete.

Data cleaning and coding

Some basic data cleaning for errors and inconsistencies are run by the ISU both at receipt of the data and prior to release of the data to external parties. There are no ICD-10-AM coding or other international standards in coding. However, the data collection is consistent with other Australian jurisdictions in terms of the data fields.

Content of eARF

Each patient in the data is assigned an eARF number when attended to by an ambulance. It is possible however, that multiple eARF numbers could be assigned to an individual over time.

In terms of the Core minimum, Core optional, and supplemental data outlined by WHO (Holder et al., 2001), eARF includes all but intent and nature of activity from the Core minimum data set. It only includes external cause, date and time of injury, severity of injury, and a narrative from the Core optional data set and mode of transport from the supplemental. It is possible that some of the information relating to the missing variables may be able to be identified in the narrative variable, however the validity and reliability of this field is not known. It should be noted that while there is information on injury severity and injury nature, these are not coded to international standards (e.g., ICD-10-AM coding) and therefore their validity and reliability is unclear. The correspondence between eARF and the WHO minimum dataset is presented in Table 3.5.

Table 3.5: eARF compatibility with WHO guidelines core MDS, core ODS, and supplemental data sets

	WHO variable	eARF variable/s
Core minimum data set	Unique person identifier	eARF number
	Age of injured person	Date of birth
	Sex of injured person	Gender
	Intent	-
	Place	Event location
	Nature of activity	-
	Mechanism	Case nature
	Nature of injury	Final Assessment
Core optional data set	Race or ethnicity of injured person	-
	External cause	Case Nature
	Date of injury	Date Case
	Time of injury	Time Case
	Residence of injured person	-
	Severity of injured person	Transport Criticality
	Alcohol use	-
	Other psychoactive substance use	-
	Narrative	Comments
Supplemental data	Mode of transport	Vehicle Type
	Road user	-
	Counterpart	-

The exact location is recorded in eARF; however this variable was not made available to the researcher, so the reliability and nature of these data is not known. The location variable field that is included only broadly classifies into location types (e.g., street, private residence etc.). This variable also only relates to the pick-up location which may not be the same location where the injury actually occurred.

For the purposes of linkage, the eARF does not include a unique identifier that is shared with any other government agency, which would preclude a simple matching of data. However, some of the health data collections (QHAPDC, EDIS, QISU), have included the eARF number since 2009. However, probabilistic linkage would likely still be required, in combination with other identifying variables (e.g., name, address, DOB etc.) as it is not always recorded well enough for direct matching. Also, it is not known how reliably eARF number is recorded in the QH data collections.

Legislation relating to eARF

The release of information from QAS is covered by the *Public Health Act, 2005* (Qld) and the *Information Privacy Act, 2009* (Qld). The sections of the legislation that are relevant have already been described in the previous section (3.4.1.2).

Access to eARF

For any data requests relating to eARF, ethics approval from a Human Research Ethics Committee is required. Once ethics approval is gained, it is advised that the data requirements are discussed with the staff at the Queensland Clinical Performance and Services Improvement Unit (QCPSI). Following these discussions a letter to Commissioner is required outlining the purpose of the research, proposed methodology, ethics clearance, and the nature of the data required. The Commissioner will then forward the request to the QCPSI for advice on the methodology and the ISU will provide advice on the availability of the data. Once the Commissioner approves the research, the researcher must sign an Agreement for the Provision of Queensland Ambulance Service (QAS) data. The QCPSI manager will also sign this document once they have received the Commissioner’s approval letter and any relevant ethics approval. Once all approvals have been gained and the agreement signed by both parties, the data is provided to the researcher in comma separated variable (.csv) format.

Timeliness

Data is generally available internally the day it occurs. Access to external bodies is generally possible approximately one month after the event to ensure the data is cleaned and coded correctly.

Metadata

There is little publicly available information on the eARF data collection in terms of its content or coding.

3.3.1.6 *National Coronial Information System (NCIS)*

Scope

NCIS includes all deaths reported to the coroner since 2000. Reportable deaths in Queensland are those where:

- the identity of the person is unknown;
- the death was violent or unnatural;
- the death was suspicious;
- the death was a health care related death;
- the death occurred in custody; or
- the death occurred as a result of police operations.

By definition, the NCIS data should include all deaths resulting from road crashes.

Purpose

The purpose of the NCIS is to provide access to coronial information for coroners, government agencies, and researchers to inform death and injury prevention activities.

Data governance

The NCIS Board of Management (which includes a representative from the Coroner, Justice Department, and public health sector in each state/ territory) oversees the operation of NCIS. The Victorian Department of Justice manages the operation on behalf of the Board. There is also an Advisory Committee that provides technical and methodological advice to the Board. This Committee has representation from the Australian Bureau of Statistics (ABS), epidemiologists, and coronial organisations.

Data collection

Staffs within the state/territory coroner's offices are responsible for the entry and coding of data into NCIS. This process is started when a case is reported to the coroner and continues until the coroner's case is closed.

Data cleaning and coding

Within NCIS, there are validation rules applied to ensure all mandatory fields are completed before a case is closed. Also, the NCIS team conducts quality reviews on all closed cases for errors and consistency. Some data are coded by the NCIS team after the data is entered into the system, including the application of geocoding and ICD-10-AM cause of death.

Content of NCIS

NCIS complies with the entire WHO guidelines core MDS, core ODS, and supplemental data sets shown in Table 3.6, with the exception of alcohol and drug use. It is possible however, that while there are no variable fields for these factors, the interrogation of toxicology reports could identify them. It should be noted, however, that these reports may not always be present for the case, and are only available when a researcher obtains Level 1 access (see Section 2.4.6) and the case has been closed.

Table 3.6: NCIS compatibility with WHO guidelines core MDS, core ODS, and supplemental data sets

	WHO variable	NCIS variable/s
Core Minimum Data set	Unique person identifier	NCIS number
	Age of injured person	Age
	Sex of injured person	Sex
	Intent	Intent
	Place	Location of incident
	Nature of activity	Activity
	Mechanism	Mechanism
	Nature of injury	ICD cause of death code
Core Optional Data set	Race or ethnicity of injured person	Indigenous identification/country of birth[1]
	External cause	Case type
	Date of injury	Date of incident
	Time of injury	Time of incident
	Residence of injured person	Address of usual residence[1], Postcode[1]
	Severity of injured person	BY DEFINITION/ICD cause of death code
	Alcohol use	-
	Other psychoactive substance use	-
	Narrative	Police report/finding/pathology report/toxicology report[1]
Supplemental data	Mode of transport	Mode of transport
	Road user	User
	Counterpart	Counterpart

[1] Only available to researchers with Level 1 access

The location of the incident is recorded as an address in NCIS, and for cases since 2006 this address has been geocoded. The geocoding however is applied approximately 3 years following the case.

NCIS includes no other items (except those previously mentioned) that comply with the minimum data requirements outlined by WHO (2010), Austroads (1997), and the MMUCC (2012). However, as noted previously, the inclusion of police reports, findings, and pathology may make it possible to identify these factors via manual review of the text. It should be noted that it is not clear how often these documents are included for a case and the level of detail may vary.

NCIS does not include a unique identifier that is common with any other data collection. However, if Level 1 access is granted, name, address, and date of birth are available to link either manually or probabilistically for closed cases.

Legislation relating to NCIS

NCIS and external parties applying for access to NCIS must comply with two Victorian acts, the *Information Privacy Act, 2000* (Vic) and the *Health Records Act, 2001* (Vic).

The *Information Privacy Act, 2000* (Vic) s.2 states that access to personal information for research purposes, without consent, is possible as long as it is impracticable to gain consent, is in the public interest, and does not involve the publishing of identifying information. The *Health Records Act, 2001* (Vic) s. 2 also outlines the release of personal information for research purposes. For research purposes it also outlines the conditions of this release are the same as for the *Information Privacy Act, 2000* (Vic) s.2.

Access to NCIS

In order to gain access to NCIS, an ethics application to the Victorian Department of Justice Human Research Ethics Committee must be approved. This application must first be forwarded to the NCIS Research Committee for consideration. Following the ethics approval, an NCIS Access Agreement must be signed between the applicant and the Victorian Department of Justice (NCIS). Once this agreement has been signed and any relevant fees paid[2], a user name and password will be issued to the approved user. This user name and password will allow access to the secure web-based NCIS platform. Data can then be queried and viewed online, or downloaded in Excel format (.xlsx). Attached documentation (i.e., police reports, findings, pathology reports, and toxicology reports) can also be viewed online or downloaded in Portable Document Format (.pdf) (if available).

Timeliness

Cases are regularly added to the system as they are reported to the coroner. However, much of the information will not be available until the case is closed by the coroner. The longer ago a case, the more likely it is to have been closed. Generally, more than 90% of cases are closed for the period 2 years before the date of access. For example, more than 90% of cases for 2010 will be closed (and have all relevant information included) by the end of 2012. It should be noted that some additional data such as geocoding and ICD cause of death coding may take a further year to be available. The time between a case being included and it being closed is not able to be precisely measured, as there are a variety of reasons for a delay (e.g., police investigations, coronial enquiries).

[2] NCIS charges an annual access fee unless an exemption has been approved (e.g., fulltime student)

Metadata

There is information about the content and coding of NCIS on their website (http://www.ncis.org.au), including the scope of the data and data quality statements. There is also a NCIS Data Dictionary and NCIS Coding Manual and User Guide.

3.4 Discussion

3.4.1 *Relevance*

With the exception of the Queensland Road Crash Database (QRCD), road safety research and reporting was not the primary purpose of the identified data collections reviewed in this chapter. However, the primary purpose of the Queensland Injury Surveillance System (QISU) could be seen as very closely relating to this purpose as their primary purpose is for the surveillance of injuries of which road crash injuries are a subset. For the Emergency Department Information System (EDIS), Queensland Hospital Admitted Patients Data Collection (QHAPDC), and the electronic Ambulance Reporting Form data (eARF), their primary purpose is administrative and they are designed for performance and quality assurance measures. It should be noted that for at least the eARF and QHAPDC, secondary purposes include surveillance and research (although not specifically injury or road crash injury). Despite the primary purpose of some of the data collections not directly aligning with that of road safety research and reporting, all of the identified data collections contain cases as well as variable fields that may be seen as relevant to road safety investigation, intervention development, and evaluation.

Each of the data collections includes road crash cases. QRCD includes all the road crash injury cases that are reported to police, QHAPDC includes all the cases admitted to hospital, EDIS and QISU include all the cases that present at the included emergency departments, eARF includes all the cases in which an ambulance was in attendance, and NCIS includes all the cases reported to the coroner. While each of the data collections includes some road crash cases, it is arguable as to whether any of them represent the entire population (see section 3.4.2 for more discussion of this issue).

All of the included data collections include information about these cases that is considered relevant. Each of the data collections includes elements of the Core MDS, Core ODS, and supplemental data sets. They also include information recommended by Austroads, WHO, and NHTSA. However, their compliance with these recommended data fields is varied and not necessarily complete.

3.4.2 *Completeness*

In this study, completeness was examined in terms of cases included (representativeness) and variables included (WHO, Austroads, NHTSA etc.). As mentioned previously, each of the data collections includes cases that are considered relevant, however by definition; some of them would not include all road crash injury cases.

QHAPDC only includes road crash injuries that were admitted to hospital and NCIS only those in which the injured person died. As a result, these data collections would only include the most serious cases and would not include the possible vast majority of injuries sustained in road crashes. For EDIS and QISU, not only do they include only cases where the injured person presented to hospital, each of the collections does not have reporting from every emergency department in Queensland. EDIS has cases from many of the large emergency departments and is arguably representative, however, QISU includes some facilities that EDIS does not, but overall has fewer included EDs and does not include some of the larger facilities (e.g., Royal Brisbane Hospital).

On the face of it, there is no specific reason to suspect that eARF would not include all road crash injuries, however it is possible that not all injuries require an ambulance and that some injured persons may transport themselves to hospital. It could be expected that QRCD includes all road crash injuries, as by definition these cases are legally required to be reported. However, it is conceivable that despite this requirement, not every injury would be reported. This is consistent with research reported from other jurisdictions (Alsop & Langley, 2001; Amoros, Martin, & Laumon, 2006; Boufous, Finch, Hayen, & Williamson, 2008; Langley, Dow, Stephenson, & Kypri, 2003).

Overall, none of the data collections included in this study would be expected to include all road crash injuries in Queensland, either by definition or due to under-reporting. It is possible however, that these data collections in combination could capture, if not all, many more cases than any of them on their own. Study 2 and 3 will explore this further by attempting to quantify the representativeness of each of the individual collections and explore the possible additional scope provided by linking these data collections together.

In terms of the completeness of each data collection, in their level of compliance with the Core MDS, Core ODS, supplemental data sets, as well as other recommended data elements, results were varied. Arguably, QRCD included the most data elements recommended by the guidelines. This is perhaps not surprising considering its primary purpose is for road safety reporting and research. While many of the data requirements are present in these data, questions relating to the reliability, specificity and validity of their recording remain. Specifically, the precision and reliability of the variables relating to injury nature and injury severity are in doubt. This issue will be further explored in Section 3.4.3 and in Studies 2 and 3 (Chapters 5 and 7).

QHAPDC includes all of the Core MDS, Core ODS, and supplemental data set variables with the exception of a narrative field. It does not however, include a specific location of where the injury took place, or any information on specific circumstances (e.g., speed, fatigue), or other crash or road user characteristics (e.g., road environment, seating position, licence status) outlined in the minimum road crash data requirements (Austroads, 1997; MMUCC, 2012; WHO, 2010).

NCIS also includes all the Core MDS, Core ODS, and Supplemental data set variables with the only exception being coded alcohol or drug use variables. Similar to QRCD, NCIS includes the exact location of the injury. It does not however, include any other coded variables recommended by Austroads (1997), WHO (2010), or MMUCC (2012).

QISU includes the vast majority of Core MDS, Core ODS, and supplemental data set variables. The only variables within these recommended data sets that are not included are race or ethnicity, alcohol, or drug use. As with QHAPDC, QISU does not include the exact location of the injury, or any other crash or road user characteristics beyond the WHO injury surveillance guidelines (Holder, et al., 2001).

The ambulance data (eARF) has some of the variables outlined by the WHO injury surveillance guidelines. However, it does not include intent, activity, race or ethnicity, residence of the injured person, alcohol or drug use, road user, or counterpart. While eARF includes the exact location of the ambulance call-out, this may not always correspond to the exact location of where the injury took place. Like QHAPDC and QISU, eARF does not include any other variables, beyond those in the WHO injury surveillance guidelines, recommended by Austroads (1997), WHO (2010), or the MMUCC (2012).

EDIS has the least included data elements of all the data collections. It does not include coded intent, place, activity, mechanism, race or ethnicity, external cause, alcohol or drug use, mode of transport, road user, or counterpart. It also has no information on the exact location of the injury or any other variables recommended by Austroads (1997), WHO (2010), or MMUCC (2012). It should be noted at this point that the narratives included in EDIS, QISU, eARF, and NCIS could provide information about other aspects of the injury or incident that are not coded, however the completeness, validity, and reliability of this variable in each of the collections would need to be explored.

Overall, QRCD, QHAPDC, NCIS and QISU have a high level of completeness of the Core MDS, Core ODS, and Supplemental data sets. eARF and EDIS, however, have only half of these variables at best. In terms of the other recommended variables, QRCD is clearly the most complete, with the other data collections lacking coded variables on many of these factors. Also, while variable fields that could represent an injury surveillance variable may be present, the completeness, in terms of data within these variables, as well the consistency and accuracy of these fields would still need to be determined (see Chapter 4).

3.4.3 *Accuracy*

As described in Chapter 2 (Section 2.4.3), one indication of the accuracy of a data collection and its variables is the existence of international coding conventions, data cleaning, and quality assurance practices. All of the data collections apply some level of data cleaning to their collection. However, the coding conventions applied to these data do vary. For QHAPDC, QISU, and EDIS, the presence of ICD-10-AM coding is an advantage, however, for EDIS and also QISU to some extent, not all variables are coded to this standard or are not coded at all.

Another aspect relating to accuracy is the location of the injury. With the exception of QRCD and NCIS, the data collections do not include an exact location of the injury. QRCD on the other hand includes GPS co-ordinates and NCIS applies geocoding to their data. The existence of these measures provides researchers greater confidence in the accuracy of the location information within these data collections.

In terms of the injuries themselves, the existence of IDC-10-AM coding of the diagnosis string within QHAPDC, QISU, EDIS, and NCIS allows for a more precise identification of the nature and severity of the injury compared to that from QRCD and eARF.

3.4.4 *Consistency*

There have been few changes to reporting practices or admission policies over the last ten years. For EDIS, there have also been few changes; however some emergency departments have only come into the system within the last ten years. This would impact on the data collection’s ability to monitor trends in emergency department presentations over time. However, it is not expected that this would impact on the consistency of the variable fields or their completeness over time. QISU too has had changes involving hospitals becoming part of the collection and others dropping out. Also, some of the hospitals that have consistently been included in the collection have dropped their ascertainment rate (i.e., the number of injuries presenting at hospital that they are including). Again, while this will impact on the monitoring of the number of cases over time, it is not expected to impact on the consistency of variables included.

eARF changed its collection system from paper based to electronic in 2006/07, which may affect the consistency of the data being collected in terms of the fields completed, however, unlike the issues for EDIS and QISU, there is no reason to suspect this change in system has impacted on the consistency of case inclusion. Finally, there is no evidence that NCIS has had any changes over time that may impact on either the consistency in case inclusion, the variables included, or the completeness of the data.

3.4.5 *Timeliness*

The lag between an injury and data availability varies between the data collections. Some data collections have data available as early as one month after an injury (e.g., eARF), while others can take up to two years (e.g., QRCD and NCIS). According to Mitchell and colleagues (2009) all of the data collections would rate at least ‘high’ on timeliness with these timeframes. However, using those data collections that are lagging by up to two years can impact on the ability to detect emerging issues in road safety and may not be seen as ‘high’ on timeliness for research or policy decisions that need to be made quickly (e.g., responding to an emerging ‘black spot’).

The delays described in this section do not include the time it takes to get access to the data for research. The process for accessing the data in each of collections can vary and in some cases can take considerable additional time. The process for access is described in the subsequent section.

3.4.6 *Accessibility*

All of the data collections allow access to some form of data by request. According to the NHTSA (1998b) in order for data to be considered accessible it should be available in an electronic unit record form as long as safeguards are in place to protect confidentiality and privacy. Mitchell and colleagues (2009) further suggest that data should be available via an internet-based platform for it to be considered 'very high' on accessibility. Using these criteria, only NCIS would be considered as 'very high'. However, based on the NHTSA (1998b) requirements each of the data collections would be considered accessible.

Regarding the process of gaining access, all of the data collections would require ethics approval. However, all research conducted within universities requires ethics approval so this is not considered an additional task for access to data per se. The rest of the process for gaining access is the same for all the hospital based data collections (i.e., EDIS, QHAPDC, and QISU). They all require a Public Health Act Application (PHA) and custodian approval for the release of data. In addition to ethics, the ambulance data requires Commissioner approval and police data requires custodian approval. NCIS requires an additional ethics approval from their dedicated ethics committee as well as custodian approval and a contract between parties. The entire process for each of the data collections can vary in length and this potentially has impact on the timeliness of data for research purposes (see Section 3.4.5). The impact of these processes on the current research and other research of this nature will further be explored in Chapter 4.

Another issue relating to accessibility relates to the available information about the data collections. Data may be available to researchers, however, the accompanying documentation and/or metadata may be lacking, making the interpretation and useability of the data more difficult. As presented above (section 3.3.1), each of the data collections included some information about their purpose, variables, coding etc. However, there were some cases in which this information was not easily accessible. There were no online resources or websites to gather information and direct questions to the data custodians were required. It is possible that researchers may not be aware of some of the collections scope or limitations and this could impact on their ability to use the data effectively. Not only would this make analysis and interpretation difficult, it could lead to inaccuracies being published that are not in the researcher or the custodians' interests.

In terms of the accessibility in the format of the data collections, as mentioned previously, all of the data collections are accessible in the recommended electronic unit record format. This allows for data to be analysed with all of the common statistical packages. However, some of the data collections have limited coded or quantitative fields, instead relying on text fields. These text fields are often not standardised and pose difficulties in terms of preparation for analysis. These fields need to be searched and coded so as to identify relevant cases and/or to apply a quantitative value for statistical analysis. This can be very time consuming particularly when there are a large number of cases (as would be the case with most hospital presentations). There are computing techniques (e.g., text mining) that can make the task easier, but these techniques require skills and infrastructure that not every researcher may have. There is also some question about how reliable these techniques are and the validity of the data in the text fields (see Section 3.4.3).

In determining how accessible data collections are for road safety research, more than whether the data itself is available needs to be considered. It is also important to ensure that sufficient information is available for use and interpretation and that the data are in a format that is useable for researchers and policy makers in the area. So while for each of the data collections summarised in this chapter are accessible in some form, some of the data collections (e.g., EDIS) are not as accessible when taking into account their ability to be easily analysed and interpreted by the user.

3.4.7 *Potential for linkage*

For the purposes of linkage, each of the data collections do not include a unique identifier that is shared with other government agencies. This would preclude a simple matching of data. Each collection, with the exception of QISU, does however, include name, address, and date of birth of the involved persons. These variables would allow probabilistic linkage between each of the data collections. For the data collections held within Queensland Health (i.e., QHAPDC, EDIS, and QISU) there is a Unit Record (UR) Number and a Facility number that could be used to link cases. These two fields would have to be used together as the UR number is not common across facilities within the collections. Also, these health data collections have included the eARF number which relates to the Queensland Ambulance data (although it is not known how consistently this is recorded). However, despite these similarities in unique IDs, probabilistic linkage would still be required, in combination with other identifying variables (e.g., name, address, DOB etc.) as it may not always be recorded well enough for direct matching. NCIS does not include a unique identifier that is common with any other data collection. However, if Level 1 access is granted, name, address, and date of birth of persons is available to link either manually or probabilistically.

It is also possible that the use of date of admission/presentation/injury/crash could be useful for probabilistic linkage to occur between the data collections. It is important in the context of road safety data linkage that not just individuals are matched correctly but that it is for the same transport-related injury, not some other ambulance callout, hospital attendance, or admission. Each of the data collections have a date that refers to the event in some capacity, so they each have the ability to be linked in this way. However, the date field in the data collections for health and ambulance do not necessarily correspond to the date that the injury occurred. It is possible that an individual is injured in a crash on one day, but does not seek treatment until a day later (or possibly even later). Based on this, the potential for the data collections to be linked in this manner may be more difficult.

Another issue relating to linkage involves accessibility. For example, currently the legislation surrounding the release of police data (QRCD) suggests that it may be difficult for identifying data to be released to an external agency. The mechanism required for enabling the sharing of data across agencies would need to be established for the linkage of data to occur in this area.

3.4.8 *Study limitations*

One of the limitations of the research was that some of the information about data collections was not available. While this was generally minimal, it could impact on the assessment of a data collections quality. Also, the exact nature of the quality issues surrounding completeness of fields; consistency over time, across incident types, and between data collections; validity issues; and representativeness have not been quantified.

3.4.9 *Future directions in research*

While this study has identified some potential data quality issues for the QRCD as well as other data collections, further analysis of the data collections is required to confirm and expand these findings. Study 2, using secondary data analysis, will provide information on the completeness of the data fields in terms of missing, unknown, and unspecified data. It will also allow for profiles of road crash injuries to be produced to highlight issues with the consistency between the data collections as well as the representativeness of each data collection and the possible under-reporting of road crash injuries to police. It will also explore the validity of some of the variables to identify cases, determine the severity of injuries, and other characteristics, as well as provide some insights into the utility of narrative variables in some data collections.

3.5 Chapter Summary

This chapter described Study 1a conducted as part of the research program. It explored the characteristics of the data collections relating to road crash injury and provided some insights into their quality in terms of completeness, consistency, validity, representativeness, timeliness, and accessibility.

The results indicate that there are limitations of the police collected Queensland Road Crash Database (QRCD), which is relied on for reporting and research in road safety, in terms of severity definitions and under-reporting. The other data collections explored in this chapter have the potential to add information to the police data in terms of both scope and content. These data collections include cases that may not be reported to police that should have been as well as including variable fields that may provide more reliable information about other factors of importance including injury nature and severity.

It should be noted however, that while many of the data fields required for road safety research are present in each of the data collections, this study did not explore the validity, completeness, or consistency of the data within these variable fields. Further examination of the data itself would be required to address these issues, which will be the focus of the next chapter.

Chapter Four: Perceptions of Data Quality and Data Linkage

4.1 Introductory Comments

This chapter outlines Study 1b conducted as part of the research program. It involved semi-structured interviews with data custodians of the relevant data collections and expert users of these data collections. It aimed to expand on the findings of Study 1a by further exploring issues relating to data quality characteristics of the road crash injury data collections, including: relevance, completeness, and consistency. It also examines the perceptions of the potential benefits and barriers of using data linkage for road safety monitoring, planning, and evaluation.

4.2 Study Aims and Research Questions

The aim of the current study was to address the research questions below.

RQ2: What are the strengths and weaknesses of each of the road crash injury data collections within the context of road safety investigation, intervention development, and evaluation?

> *RQ2f: What are the perceptions of data users and custodians on the quality of road crash injury data collections?*
>
> *RQ2g: What are the perceived areas of improvement to the quality of road crash injury data collections?*

RQ4: What are the facilitators of and barriers to linking road crash injury data collections in Queensland and elsewhere?

> *RQ4a: What are the perceived benefits of using data linkage in road safety?*
>
> *RQ4b: What are the perceived barriers to using data linkage in road safety?*

4.3 Method

4.3.1 *Interviews*

4.3.1.1 *Participants*

Three samples of participants were interviewed as part of Study 1: data custodians, expert data users, and data linkage experts. The data custodians were managers and/or analysts of the key data sources identified as potential sources of road crash incidents and/or injuries. An outline of the data managers/analysts in terms of the data source to which they were responsible and the organisation they were affiliated with is provided in Table 4.1. It should be noted that not all custodians who were approached to participate in the study were able to be interviewed. However, for the sake of anonymity, their agencies cannot be identified.

Table 4.1: Data custodians

Role	Data source	Organisation
Manager, Data Analysis Unit	Queensland Road Crash Database (QRCD)	Department of Transport and Main Roads (TMR)
Analyst, Data Analysis Unit	Queensland Road Crash Database (QRCD)	Department of Transport and Main Roads (TMR)
Director, Centre for Pre-Hospital Research	Queensland Ambulance Service Data (eARF)	Queensland Ambulance Service (QAS)
Manager, Statistical Output Unit	Queensland Hospital Admitted Patients Data Collection (QHAPDC)	Queensland Health (QH)
Analyst, Statistical Output Unit	Queensland Hospital Admitted Patients Data Collection (QHAPDC)	Queensland Health (QH)
Director	Queensland Injury Surveillance Unit (QISU)	Queensland Health (QH)

The expert data users were selected based on their involvement in research that utilises administrative and/or population-based injury data sets identified as potentially relevant to road crash incidents and/or injury in Queensland. All participants have had direct experience with at least one of the data sources described in Chapter 3, Section 3.2. In total eight expert data users were interviewed in order to cover each of the relevant data sets as well as a variety of relevant research topics. The participants were identified via the researcher's and supervisors' current networks and published materials in the area.

A total of twelve Australian and international data linkage experts were also interviewed. They represented both health data linkage generally and road safety data linkage specifically. Participants were identified based on their involvement in research utilising data linkage or employed at a key data linkage centre or unit. Contacts were determined via websites for data linkage centres and published materials in the area.

4.3.1.2 *Procedure*

A semi-structured interview schedule was developed based on the available literature and the review of relevant legislation and polices undertaken as part of Study 1a (Chapter 3). The full interview schedules for each of the participant groups are included in Appendix D. The interviews included questions relating to:

- Relevance, completeness, consistency, and timeliness of the data in terms of data quality

- How well the data identify new or emerging issues/problems and stable/consistent monitoring over time in injuries
- How well the data describe key characteristics of injuries and their external cause and what additional information is available in terms of identification of risk groups and factors
- What incidents/events are not included in the data collection and what is missing from those that are included
- Elements in the data collection, such as unique identifiers and/or other variables that would facilitate linkage to other sources of information on injury
- Who collects the data, where is it collected, when is it collected, how is it collected, cleaned, collated, coded, stored and what quality control processes exist
- Storing, reporting and access to data, timeliness, availability of glossaries/definitions/coding keys

While there was some overlap in the questions asked of data custodians and expert data users, some specific questions relating to their particular perspectives were also included. Data linkage experts were asked questions relating to their experiences with the linkage process and research conducted using linked data. Some of the information sourced from the interviews with the custodians was used, in conjunction with the document review, to ascertain the details of the data collections in terms of their scope, purpose, access, etc. These results are presented in Chapter 3, Section 3.3.1. The other questions for the custodians and those for the expert users were used to gather information on the perceptions of the quality of the data collections. Data linkage experts were asked questions relating to their experiences with the linkage process and research conducted using linked data. The linkage experts, data users and data custodians were also asked about their perceptions of the barriers and facilitators of data linkage in road safety. Also, data custodians were asked about their respective data collection's potential for linkage.

QUT ethical approval was obtained for the interviews with the expert data users and data custodians. Further ethics approval was granted from the Queensland Health HREC to interview Queensland Health employees. The Queensland Ambulance Service Commissioner approved an interview with the Director of the Australian Centre for Pre-Hospital Research. Participants were approached via email. The email outlined the nature of the study and contained an information sheet and full set of interview questions. Participants were asked, if they wished to participate, to contact the researcher to arrange a time and location for the interview. Before each interview commenced, verbal consent was obtained. Each interview took approximately one hour to complete. Following the interview, the participant was thanked for their time. To increase rigour and reliability, the interviews were tape recorded and transcribed. The transcription was double checked for accuracy. Any names or identifying information were removed from the transcription. For the data custodians, a transcript of their interview was sent to them for verification.

4.3.1.3 *Statistical analysis*

Qualitative analysis of the interviews was conducted to explore relationships between identified themes as well as to manage, summarise and find meaning in large semi-structured quantities of data. Themes were generated to index categories of information. Although this study is primarily exploratory in nature, there was a conceptual framework on which the interview questions were based. Specifically, the questions were based on the data quality characteristics of relevance, completeness, consistency, accuracy, timeliness, and accessibility. Therefore, themes were initially generated from this framework then confirmed by the data. However, other themes were also generated from the participants' responses.

4.4 Results

4.4.1 *Perceptions of data quality*

4.4.1.1 *Relevance*

<u>For what purpose/s do you use these data?</u>

The expert data users utilised the data collections for their research in a variety of ways, including exploring data quality, evaluation, trend analysis, and identifying risk groups.

Data source	Data custodian	Data expert
QRCD	Not Applicable	*"looking at the characteristics of different road user groups crash involvement"* *"For evaluation purposes"* *"Monitor crash trends over time"*
QHAPDC	Not applicable	*"Assessing the quality and completeness of the injury data"*

<u>How well do you think the data identifies new and emerging issues in road safety?</u>

The general view of all the expert data users and data custodians was that the major barrier to a collection identifying new or emerging issues was whether the relevant information was captured or coded in the first place. Also, even if it is captured, other factors can account for the change, such as coding or process changes. These changes may have implications for consistency (see Section 4.4.1.2).

Data source	Data custodian	Data expert
QRCD	*“.......it can be difficult at times because if you've actually got specific fields for scenarios that we know are an issue* *“....we wouldn't have a field for a particular instance that might be starting to occur.”* *“In the past, what was thought to be new and emerging issues that were incorporated into the database....things like bull bar, airbag, communication device, racial appearance, four wheel drives.”*	*“generally good”* *“data on some new issues are not collected, as historically they weren’t relevant, for example mobile phone use”* *“Some things if there's pre-existing items that will capture that, then fine. But otherwise no”* *“There are some things we've just been very bad at doing because it's a category of behaviour or phenomenon which is qualitatively different. Like when mobile phones came out there was nothing on crash forms about mobile phones, because they hadn’t existed.”*
QHAPDC	*“It’s only going to identify traffic injuries or traffic incidents if they’re coded in terms of ICD. So it needs to be recorded for it to be coded”* *“It will be dependent on how well the actual chart itself is written. Unless the treating physician or triage clearly specifies that it was this type of accident it’s not going to make its way through the coder after”*	*“The hospital data is really only good at tracking things it codes”* *“.......there may be things that are new that there may not be a code for so we can’t capture those”*
QISU	*“I don’t think it identifies particularly well in the transport-related area”* *“One of the limitations is that the data is coded at triage”*	*“No routine analysis to indicate emerging issues”*
eARF	*“Pretty well, however it does not capture all of the population, only those that call an ambulance. There may be certain risk groups or events in which people don’t call an ambulance......the count of incidents can be problematic”*	

How well do the data describe key characteristics of the road crash incidents and the injuries involved?

Generally, the data custodians and expert data users believed that some aspects of the data collections were described well. However, for the police data (QRCD) there were concerns about the level of detail relating to lower severity injuries. Also, the participants identified a number of factors that were not captured well, such as work-related road crash injuries and the injury type. In contrast the health data collections were seen as very good at capturing detail about injury nature, but not very good at capturing the circumstances of the injury.

Data source	Data custodian	Data expert
QRCD	*"It depends on who's collecting the data and what level of severity it would be."* *"As a general rule, yes, but in some cases if incidents are reported over the counter and there's a delay between that reporting. There have been some descriptions that might be recorded in a way that the exact date of birth is not known, for say a child who had a very minor injury."*	*"for the contributing factors it records well"* *"purpose of journey has historically not been collected"* *"ethnic status is now collected, but generally not collected well"* *"The crash data is obviously very good at location information and you can get some basic information on the kinds of vehicles involved and people."* *"The crash data is a little bit dubious about the level of injury apart from fatality."*
QHAPDC	*"The identification and break up of type of vehicle involved all those sorts of things - you're limited to the ICD-10 classification system."* *"Where a person previously had a suspended license for drink-driving. It's not something that we're going to know anything about or be in a position to find out about."*	*"whether they were the driver or the passenger"* *"good for the types of injuries"* *"demographics is relatively good"* *"not a lot of detail on the specifics of where it occurred"* *"not a lot about what went wrong......... whether there was alcohol or speed"* *"So the hospital data is very good at the injury side of things, but it's very poor at location. Often you don't know whether it was in fact a reportable crash or not because location is part of the criteria for whether or not a crash is reportable."*

QISU	*“So sometimes we're limited because of the urgency of the presentation……We just don't get all of the information, so we can't code it all.”* *“What we do tend to capture fairly well is usually whether it was a driver or passenger - that's not too bad.”* *“We tend to get that clinically relevant stuff but we don't always get a lot of the other mechanism stuff and particularly the safety stuff that we would like to get.”*	*“Only if it is documented in the first place”* *“They would describe the object that was associated with it quite well”*
eARF	*“Depends on what ambulance officer records, would identify that it is a traffic crash, may have vehicle information (sedan, truck), may describe the mechanism, demographic information, time, date, location”*	

4.4.1.2 *Completeness*

<u>What incidents/events are not included in the data collection?</u>

Participants reported that there would be a variety of road crash incidents and/or injuries that would not be captured by the data collections. While, they noted that some of these are not included by definition because they do not fit the collecting agencies purposes, they highlight that these incidents may still be of importance to the prevention of road trauma. In terms of the variables included in the collections, participants believed that the coverage was quite good, however they did note that some things are not recorded (e.g., work related incidents, indigenous status, specific location). The issue of missing data was not noted as being widespread; however, there was some suggestion that unspecified or unknown categories are used for some variables.

Data source	Data custodian	Data expert
QRCD	*"A crash wouldn't be included if it was to do with flooding or when there is a suicide or a medical condition"* *"There is also a threshold for damage or existence of injury"* *"It has to be on a road or road-related area"* *"It all comes back to the purpose of the road crash database which is implementation of policy. So we don't have jurisdiction over those areas, we can't prevent a suicide or a deliberate act, a medical condition or what happens on private property."* *"Those not reported to police obviously aren't in there, but we don't know to what extent that happens"*	*"Any that drivers choose not to report"* *"It's possible that some categories of road users are under-represented, such as a bicycle incident where no other person is involved"* *"Those not included by definition.....issues of community concern, such as what happens in driveways, car parks, and other off-road situations"* *"while these things may not be seen as the purpose of the data, they are seen as potential road safety issues that may be falling through the cracks"*
QHAPDC	*"Only if they weren't admitted to hospital"* *"It's only going to identify traffic injuries or traffic incidents if they're coded in terms of ICD"*	*"Those that don't seek treatment in an emergency department or in a hospital"* *"Those where the cause wasn't documented"*
QISU	*"Sometimes the triage nurses tick no to an injury because if they tick yes, the injury screen pops up and then they have to fill it out."* *"Not all hospitals are included in QISU"*	
eARF	*"Any not involving an ambulance"*	

Are the data able to identify risk groups and factors?

Data custodians and users stated that they thought the police and hospital collections were adequate at capturing risk groups and factors. However, users felt with both collections that there were factors missed, such as work-related driving in the case of the police data and alcohol-relatedness for the hospital data. It should be noted however, that it was pointed out by a hospital data custodian that it is not necessarily the purpose of health data to collect this information.

Data source	Data custodian	Data expert
QRCD	*"I think it's pretty good"* *"It's good in that it identifies the common high risk groups, like speeding and drink driving"* *"I think the implementation of policy over the last 15 years and more recently young driver, the road crash database was able to identify the key characteristics that young drivers were having problems facing. That was inexperience and the occupancy, how the risk changed with higher occupancy. You know, V8s, the high powered vehicles, identified young drivers of those were at great risk of higher severity collisions and things like that."* *"The reductions that we've achieved there have been outstanding. So I think those three examples, and there's many others, are based on evidence extracted from the road crash database. They were basic type characteristics that we evaluated."*	*"If they align with the mainstream road safety research, such as speeding and drink driving, it's generally good"* *"difficult to identify those who drive for work purposes, indigenous people"* *"There are also some issues of concern in road safety circles, such as aggressive driving, that aren't identified specifically in the database, although there are definitional issues as to what aggressive driving is"*
QHAPDC	*"Police obviously collect a range of information that's got nothing to do with the subsequent hospitalisation - that is appropriate for the police collect but not Queensland Health."*	*"I think broadly and at the more severe ends, such as age groups road user group"* *"Doesn't identify whether they were alcohol affected, what kinds of contributing factors there were (however not sure if this is health's role)"*

4.4.1.3 *Consistency*

How well do the data allow the monitoring of road crash incidents/injuries over time?

Participants reported that QRCD and QHAPDC generally allow the monitoring of incidents over time, however they did note that there are some factors that may influence the data collections consistency over time (e.g., changes in reporting practices and/or policies). Participants reported that some coding or at least categorisations are based on international or national standards and that there are some similarities in the nature of data from one collection to another. Some areas of improvement that are suggested include looking at other jurisdictions to establish best practice and improved training and resources.

Data source	Data custodian	Data expert
QRCD	*"this is what the crash data does best"* *There have been some changes in definitions with the property damage crashes, but since 1999 it's been the same"*	*"one of the strengths of the data collection is that it generally collects the same things over a long period of time"* *"it's a valuable tool for monitoring that which we know"* *"We know that the crash data under report crashes, but we have fairly good reason to believe that it under reports in a consistent way. So that you're generally able to pick things up."*
QHAPDC	*"The biggest change I think as far as the traffic accidents and transport accidents have been the shift from the precursor to ICD-10-AM had a differently structured set of traffic accidents……it's quite hard to, it's almost a break in series. It's really hard to go back beyond that step. So that's about 1999-2000. Time series is going to strike a bit of a glitch, if you go back further than that"*	*"Need to have some confidence, particularly when looking at trend data, that major peaks or troughs don't reflect coding changes"*
QISU	*"Adding new sites, other sites dropping out"* *"Losing their support person and then that goes down, so there's fluctuation in the ascertainment and fluctuation in the number of sites and the location of the different sites."*	
eARF	*"Going back to far may be problematic due to change in reporting systems, tend to not go back past 2007."*	

Does the nature and quality of information recorded vary depending on the type/nature of the incident/injury?

For both the health data collections and the police data there was some concern about the consistency in terms of the severity of the injury. Interestingly however, some highlighted that more severe incidents could have lower quality due to the higher demand at the scene or in hospital while others believed that the higher severity cases would have better quality information because of the impedance to collect detailed information. Other concerns for consistency were based on the inclusion of cases. Specifically, there was concern that some cases may be less likely to be reported to police (e.g., cyclists and motorcyclists), which would impact on the quality of the data in QRCD.

Data source	Data custodian	Data expert
QRCD	*“I think the police are fairly well trained in recording road crashes and so there is consistency across locations.”* *“there are sometimes issues depending on the severity of the crash, data is checked more carefully for the fatalities and the hospitalisations, so they may be better so to speak”*	*“Some external factors can impact. For example, changes to the law for making a CTP claim.....since late 90s you need to have a police report”* *“always assumed the more severe the crash the more reliable the data is”* *“We also know that the under reporting varies by road user type. So that cyclists in particular are radically under reported compared to other ones.”*
QHAPDC	*“If there were to be differences in the hospitals I imagine that it could be size of hospital - a tiny hospital out West or whatever. I’m not even sure that that’s going to be the case because they are going to be treating in the main much lesser sorts of serious acuity or whatever”* *“However, you’ve got to be trained as a coder, so it should be pretty consistent”*	*“might be more inclined to put someone in hospital if they are an elderly person or a child”*
QISU	*“The time of presentation….if it’s busy”*	*“I am aware that there are problems with things like QISU data because that has to be entered by emergency nurses on screen in the emergency ward. Come in on a Saturday night when you're deluged with bleeding drunks, then things get missed and it's not surprising.”*

eARF	*"Some situations demand on the ambulance officer to attend to patient care (first priority) or other distractions (other emergency personnel – police, fire department) may make filling out a lot of detail difficult"* *"Serious injury, multiple casualties, lots of activity, may make less comments or report less detail as priority is patient care, however more serious may increase detail because if the ambulance officer has time after hospital they may take great care to record as much detail as possible due to the injuries serious nature, minor injuries may not have a lot of detail"*	

<u>How could reliability and consistency both within and between data sets be improved?</u>

There were a couple of suggestions from the expert users as to how the data collections could be improved in terms of consistency. For the police data there was a suggestion of the inclusion of compulsory blood testing and for some alignment of definitions surrounding fatigue and severity. For the admitted patient hospital data, there were suggestions of better training, support networks, and increasing the awareness of the importance of the data to improve coding standards.

Data source	Data custodian	Data expert
QRCD		*"Compulsory blood testing"* *"I think there is a need to look at practices in other jurisdictions, particularly on the issue of severity"* *"look at how the fatigue definition aligns with the national one"*
QHAPDC		*"Better training for the people who are doing the coding"* *"A network for people to check up on things they're not sure of"* *"Emphasising that it is an important part of the data collection"*

4.4.2 *Perceptions of data linkage*

4.4.2.1 *Potential benefits of data linkage*

The participants identified a range of potential benefits associated with the use of data linkage in research including those relating to reductions in bias, increased sample size, and cost effectiveness.

> “*Often less selection bias. Administrative data systems are not normally subject to such systematic exclusions. Other types of selection effect from which cohort studies using population-based linked data are likely to be largely sheltered are those related to place of residence, language and propensity to volunteer.*” – Data linkage expert

> *“Potential for large cohorts and/or long follow-*up *at relatively low marginal cost. The relationship of cost to scale tends to be much more favourable in linkage based studies.”* – Data linkage expert

> *“Linked data is cost-effective for researchers as they can access large amounts of data at a fraction of the cost that would otherwise we required to collect the data via survey methodology.”* – Data linkage expert

> *“Linked data provides access to population level data which allows researchers to generalise the results to a broader population, or take into account any bias.”* – Data linkage expert

It was also noted that data linkage allowed for research that would not be able to be performed using only one data collection.

> *“They are able to answer more complex research questions. Fosters collaboration between disciplines. Clinicians can give insight into epidemiologic questions and vice versa.”* – Data linkage expert

> *“Linked data can provide additional information than what is otherwise information that is only retained within one data collection. For example, in road safety, police-reported data often contains detailed information regarding the circumstances of a crash, but little information regarding the injuries experienced and their treatment. The hospital separation data collection contains scant information regarding the circumstances of a crash, but detailed information regarding any injuries, treatment and care provided.”* – Data linkage expert

Data custodians too suggested potential benefits of data linkage for both their government agency and other groups. They did however; see more benefit for others than for themselves.

> *“It could be of value to us. It could be of value to medical practitioners. It could be added to our database, it could be added to their database. But I think a project would be best rather than doing it routinely”* – Data custodian

> *"Because, as far as like policy is concerned and what we're doing, we've got what we think we need, like it may be better for the medical practitioners to know more about the history of what occurred……It's going to be more value post-crash than working out prevention for us because we're looking at prevention whereas medical side is looking at treatment of injuries."* – Data custodian

4.4.2.2 *Perceived barriers to data linkage*

Many of the participants reported that a key barrier to data linkage was agencies lack of willingness to share the required data for linkage.

> *"I think the main thing that you'd have to get over is the data sharing, whether you can or you can't."* – Data custodian

> *"The other barrier is more of an institutional one. Getting agencies to cooperate in supplying the data and helping each other out. There's not really any interest in that because they've developed their own data systems for their own purposes."* – Data user

> *"Largely to meet the management purposes of that department and finally from a privacy point of view that there are those - the data's typically collected not for research purposes but for administrative purposes and hence, in recent years there's been growing concerns about using it for non-administrative purposes."* – Data user

> *"Some of the reluctance from some departments about releasing that data to someone else to do the linkage"* – Data user

> *"I think the main thing is common identifiers and whether they're MOUs or inter-departmental agreements about data sharing protocols and processes."* – Data user

Another issue related to the quality and nature of the data to be linked. There was some concern that inconsistent coding between data collections, the delay in data availability, and errors in the data could make linkage problematic.

> *"If it went one way or the other, if the hospitals wanted our data, when do they get it? In 18 months' time? Or do they want it now, whatever's there. Whether it was accurate, incomplete or whatever state it was in or do we get hospital data now and in 18 months incorporate it into our processing that would be a change of series."* – Data custodian

> *"Lack of consistent coding etc. same information but in a different form, starting to record ambulance unique identifier in QHAPDC which could assist in linkage"* – Data custodian

> *"There is a deep suspicion that there's an awful lot of mismatches where you could actually, with a bit of effort, match up with a letter wrong in a name. Or a*

> *digit wrong in a date and that sort of stuff. That probably accounts for a lot of the mismatches."* – Data user

> *"I think some of the barriers are the different kind of systems that the data are sitting in, that may not necessarily lend themselves well to producing a data set that can be linked."* – Data user

Resourcing was also an issue raised by almost all of the participants. There was a sense that linkage takes considerable amount of time and that many departments do not have the capacity to cope.

> *"It's normally a lot of effort involved, a lot of time involved. So we've got only a small capacity really for this kind of thing."* – Data custodian

> *"It's the size of it* - *especially if you are going outside four or five years."* – Data custodian

> *"If someone gives you a file with 10,000 names then it means you've got to go through your two million records 10,000 times. So you've got an awful lot of computer time chewed up in doing those sorts of comparisons."* – Data custodian

> *"Once you are dealing with more than a few thousand records….a lot of grey matches to do manually."* – Data custodian

> *"Not enough physical people there that are all skilled up to be able to do it"* – Data user

> *"Certainly it's going to be feasible to conduct linkage. It just needs to have the manpower and means to do it."* – Data custodian

There was also a concern expresses by some participants that it would be difficult to deal with the transient demand for linkage within a department.

> *"Can't get someone to come in just for two months just to work on someone's project…….It's not just a simple matter of knowing about oracle databases, you have to know all the table structures and data and data definitions, the history of the data collection* - *before you can really start to do that work."* – Data custodian

There were also some comments surrounding the capacity of the hardware to deal with large linkages.

> *"Sometimes the size of the data files outstrips the capacity of the hardware used to do the linkage."* – Data linkage expert

Many of the participants mentioned that the time required to undertake linkage takes currently is an issue particularly for researchers.

> *"In my experience, the time this entire process has taken has been approximately one year. Unfortunately, researchers have ended up being extremely disconcerted by the lengthy process."* – Data linkage expert

From a custodian and/or agency perspective there were concerns surrounding the impact of using linked data in their reporting practices. Specifically, they were concerned that it would cause a break in their data series and be difficult to explain the change.

> *"I think we've looked into that and we've looked at what possible impact it might have on us and the way we do things and whether it would improve or impede on what we're doing or whether - it could result in chopping and changing of casualty severity outcomes and we'd be reporting something one week and reporting - if we did a link - something different the next."* – Data custodian

Another issue, primarily raised by the data users, was the lack of information about the data linkage process. They believed this had impacts for the researchers in that they are unaware of the process for gaining access to the required data and/or the linkage of data. It was also noted that some custodians are not aware of what is involved in data linkage and/or the potential benefits of the methodology for research and policy.

> *"From the end user point of view, it's not clear how to get to the data linkage unit and how to get things to be done in a reasonable kind of timely manner."* – Data user

> *"It hasn't been made very apparent to people what the processes are."* – Data user

> *"Over and above that I think the potential benefits of linking have remained a bit nebulous so there perhaps hasn't been an impetus for it. Linking would need some kind of whole government impetus and a commitment to funding it for those reasons."* – Data user

Some of the concerns seemed to depend on the proposed nature of the linkage. In particular, data custodians were not supportive of the idea of data warehousing or consolidation of their data into one large linked data collection.

> *"To consolidate them together.....you're not going to get data to talk to each other. Even to get through the file of Queensland Health - the IT project involved in having four different government departments send their data through to match and put together - it's where we all back out."* – Data custodian

> *"You're talking about getting departmental agreement at head of executive level to engage in a project of research in an ongoing way as opposed to part of a research project. That's beyond the scope."* – Data custodian

The data custodians were however, more open to the idea of doing things on a project basis as a trial to see what the benefits, if any, are.

> *"I think maybe it should be based on historical data, not now data and do a trial for a particular year. We've already processed it. Maybe do a link for a certain period, as a trial"* – Data custodian

> *"Do a link for a certain period and find out where the benefits are in that, if there are any. What the accuracies and inaccuracies, what the differences were between the two."* – Data custodian

Both custodians and users stated that more advocacy and information about the potential benefit of data linkage could encourage more support for it among researchers and relevant custodians.

> *"We'd have to research it, we'd have to trial it before we did it and look at what effect it would have on us and what value would that have for the purpose. How would that help us?"* – Data custodian

> *"I think more advocacy across the board, not just in the health area but with outside agencies, to say here's what's happening."* – Data user

> *"Here are all these useful, interesting things that we can find out, so that you could get other custodians on board and other sectors on board, to see it as a good thing to do."* – Data user

4.5 Discussion

4.5.1 *Perceptions of data quality*

It was generally reported by the data users and custodians interviewed that QRCD and QHAPDC were consistent over time in terms of both case inclusion and the variable fields. However, some of the data users and custodians highlighted that some cases will not be recorded with the same level of detail as other cases. For example, the QRCD custodian and a number of users of these data suggested that more severe cases may have a greater level of detail associated with them and therefore may have more complete information relating to the characteristics and circumstances of these injuries. In terms of the scope of QRCD, there was also some suggestion that cases involving certain road users would be more likely to not be reported (e.g., cyclists and motorcyclists). This would have an impact not so much on the accuracy or completeness of the information about cases, but would bias the overall number of cases.

For QHAPDC, the only suggested threat to consistency was based on the characteristics of the injured person. It is possible that certain types of injured persons may be more likely to be admitted to hospital based on admission policies rather than the severity of the injury per se. Specifically, it was suggested that the scope of QHAPDC could be biased toward the very young and the very old. For example, if a child under 10 attends hospital with a possible head injury, they would certainly be admitted even if it was just for observation. This may also be true of older people, particularly if they have other medical conditions that could impact on their treatment or recovery. A person aged in their twenties that is otherwise healthy, may be less likely to be admitted and simply just treated at the emergency department. There were no real suggestions of any bias or inconsistency in terms of the variables being recorded or the accuracy or completeness of the data fields for QHAPDC.

4.5.2 *Perceptions of data linkage*

4.5.2.1 *Perceived benefits of data linkage*

There were many perceived benefits associated with the use of data linkage in research including those relating to reductions in bias, increased sample size, and cost effectiveness. It was also noted that data linkage facilitated research that would not be able to be performed using only one data collection. Particularly, in road safety research, it was suggested that police data would have a lot of information about the circumstances of a road crash injury, but very little information about the injury sustained. Conversely, hospital data would include this detail about the injury, but lack the information about the circumstances of the road crash injury. Therefore, a perceived advantage of linking data from these two sources is that one would gain information from both sources into a consolidated view of the incident, including a better defined serious injury profile. Data custodians too suggested potential benefits of data linkage for both their government agency and other groups. They did however; see more benefit for others than for themselves as they believed it would not add much to their prevention efforts, but rather be of benefit to those who deal with clinical outcomes.

4.5.2.2 *Perceived barriers to data linkage*

Many of the participants raised concerns about the potential unwillingness of agencies to share the required data for linkage. It was generally a view, particularly for custodians and/or agencies where linkage has not historically occurred, that there would be reluctance among them to share data with other agencies. Another issue related to the quality and nature of the data to be linked. There was some concern that inconsistent coding between data collections, the delay in data availability, and errors in the data could make linkage problematic.

Resourcing was also an issue raised by almost all of the participants. There was a perception that linkage takes considerable amount of time and that many departments do not have the capacity to cope. There was also a perception, amongst custodians, that it would be difficult to deal with the transient demand for linkage within a department. Particularly, they thought that there would not be enough linkage work to have permanent employees assigned to the task, and that when linkage projects come up, they would have trouble sourcing temporary staff with the required skills relating to data linkage and knowledge of the data collections. Another resourcing issue highlighted by some participants related to the capacity of the current hardware to deal with large linkages.

Relating to both the resourcing issue and the difficulties in gaining agreements to share data, many of the participants mentioned that the time required undertaking linkage currently is an issue for researchers. It was seen as difficult for researchers to meet the deadlines of their research within the current system.

From a custodian and/or agency perspective there were concerns surrounding the impact of using linked data in their reporting practices. Specifically, they were concerned that it would cause a break in series in their data and be difficult to explain the change to users.

Some of the concerns seemed to depend on the nature of the linkage. Data custodians were not supportive of the idea of data warehousing or consolidation of their data into one large linked data collection. They were however, more open to the idea of doing things on a project basis as a trial to see what the benefits, if any, are.

Another issue, primarily raised by the data users, was the lack of available information about the data linkage process. They believed this had impacts for the researchers in that they are not sure what the process is for gaining access to the required data and/or the linkage of data. It was also noted that some custodians are not aware of what is involved in data linkage and/or the potential benefits of the methodology for research and policy. As a result, both custodians and users stated that more advocacy and information about the potential benefit of data linkage could encourage more support for it.

4.5.3 *Study limitations*

One limitation of this study was that not every custodian agreed to be interviewed; again the lack of information about the data collection from all custodians could impact on the assessment of the data collections' quality.

It is also worth noting that only a selection of data users were chosen to be interviewed and it is possible that the current sample of data users was biased. Specifically, the data users were chosen because of their experience with these collections (based on published material) and hence they may be very knowledgeable about the data collections. While this was ideal in determining the exact nature of the collections from a research perspective, it may also be of interest to hear about others that do use data of this type, but are less knowledgeable about its strengths and limitations. This may have given a more rounded understanding of issues such as accessibility, including useability of the data.

Another potential relates to the perceptual nature of the data collected that limits some of the conclusions which can be drawn particularly about of the quality of the data. Also, the exact nature of the quality issues surrounding completeness of fields; consistency over time, across incident types, and between data collections; validity issues; and representativeness have not been quantified.

In terms of data linkage, there was also reliance in this study on the perceptions of barriers and benefits of data linkage. Despite this, many of the issues raised in this study are consistent with literature on the subject and reflect the experience of data linkers and users of linked data around the world. Also, while the perceived benefits of data linkage have been explored, the actual outcomes of this type of methodology have not been explored in the current context. This will be addressed in Study 3, which will involve the analysis of linked data to see what benefit, if any, it can provide over non-linked data in terms of data quality improvement and application to road safety research and policy.

4.6 Chapter Summary

This chapter described Study 1b conducted as part of the research program. It involved interviews with data custodians and users relating to the six data collections relevant to the recording of road crash injuries in Queensland. The results indicate that there are concerns about the police collected Queensland Road Crash Database (QRCD), which is relied on for reporting and research in road safety, in terms of severity definitions and under-reporting. However, to confirm the validity of these concerns it will be necessary to further explore the matters through direct analysis of the data collections (see Chapter 5).

Other data collections explored in this program of research have the potential to add information to the police data in terms of both scope and content. These data collections include cases that may not have been reported to police but should have, as well as variable fields that may provide more reliable information about other factors of importance including injury nature and severity. However, again the utility of these data collections and their data quality characteristics will need to be explored further (see Chapter 5).

The results also indicate that there is potential for data linkage to address issues of under-reporting and severity definitions. However, the exact nature of this linkage process will need to be explored as well as a quantification of any benefits to our understanding of the road safety problem. These two issues will be the topic of Chapter 6 and 7 respectively.

Chapter Five: Quality of Road Crash Injury Data Collections

5.1 Introductory Comments

This chapter outlines the second study conducted as part of the research program. It involved secondary data analysis of six data collections which include road crash injury information in Queensland:

- Queensland Road Crash Database;
- Queensland Hospital Admitted Patients Data Collection;
- Queensland Ambulance Service (eARF);
- Queensland Injury Surveillance Unit;
- Emergency Department Information System; and
- National Coronial Information System.

This study builds on the results of Studies 1a and 1b by examining the quality of the data collections in terms of completeness of variables, consistency, validity of coding, and representativeness. It also investigates these issues specifically in terms of injury severity coding. The results of this study will also form the basis for Study 3.

5.2 Aims and Research Questions

This section of the research aimed to address research questions three and four. Sub-questions for each of the broad research questions are outlined below.

RQ2: What are the strengths and weaknesses of each of the road crash injury data collections within the context of road safety investigation, intervention development, and evaluation?

> *RQ2h: What is the completeness of each of the core/minimum data set variables in each data collection?*
>
> *RQ2i: Is there any bias/inconsistency in the amount of incomplete data based on age, gender, road user, severity, or ARIA+?*
>
> *RQ2j: What is the validity of the coding/classification of the core variables?*

RQ3: To what extent are the road crash injury data collections consistent with one another in terms of scope, data classification, and epidemiological profile?

> *RQ3c: What is the prevalence of road crash injuries for each data collection?*
>
> *RQ3d: What is the profile (age, gender, road user, and ARIA+) of road crash injuries for each data collection?*

RQ3e: How does the profile of road crash injuries for each data collection compare to that of the Queensland Road Crash Database?

RQ3f: How do the different measures of severity relate to each other in terms of their classification of serious injury?

RQ3g: How do the data collections differ in terms of severity profile (classification of serious injury)?

5.3 Method

Ethics approval was obtained from the Queensland University of Technology's Human Research Ethics Committee (#1100001065). A Public Health Act agreement was completed by the researcher and signed by each of the Queensland Health data custodians (EDIS, QHAPDC, and QISU) and the Queensland Health Research Ethics and Governance Unit. Approval was also provided by the Queensland Ambulance Commissioner via mail correspondence. QRCD data was provided following approval (via designated form) from the Manager of the Data Analysis Unit at the Department of Transport and Main Roads. NCIS data was provided following ethics approval from the Victorian Department of Justice's Human Research Ethics Committee and a contract being signed between the researcher and the Victorian Department of Justice.

5.3.1 *Data characteristics*

Data was requested from the Queensland Road Crash Database, Queensland Hospital Admitted Patients Data Collection, Queensland Ambulance Service, Emergency Department Information System, Queensland Injury Surveillance Unit, and the National Coronial Information System. The characteristics of each of these data sets and the years examined are presented below. The time taken for data to be provided from request is also described for each data collection.

5.3.1.1 *Queensland Road Crash Data*

The data requested from the Queensland Road Crash Database included all police reported crashes, casualties resulting from crashes, and controllers (i.e., drivers, motorcycle riders, cyclists, and pedestrians) involved in crashes from 1st January 2005 until 31st December 2009. Data were provided in four separate comma separated variable (csv) files. These files were imported into SPSS 19 for data coding and analysis.

In total there were 114,749 casualties, 159,012 controllers, and 138,275 crashes (85,425 involved at least one casualty). For the purposes of this study casualties will be the countable unit of interest. The variable fields included information about the controllers involved in crashes (e.g., age, gender, licence type), temporal factors (e.g., time of day, day of week, month of year), location factors (e.g., ARIA+, police region), crash factors (e.g., nature, circumstances, number of units) and details about injured parties arising from the crash (e.g., age, gender, road user type, severity, injury description). For a detailed description of the variable fields included in the study data for QRCD refer to Appendix E (Table E.1).

The data was provided one month following the request.

5.3.1.2 *Queensland Hospital Admitted Patients Data Collection*

The data included all acute hospital admissions cases (episodes) in Queensland Hospitals (private and public) coded as a land transport injury (ICD-10-AM External Cause Codes from V00-V89) from 1st of January 2005 to 31st December 2010, totalling 75,495 cases. Data were provided in a comma delimited text (txt) file and was exported to SPSS 19 for data coding and analysis.

Variable fields provided for this study included demographic variables (e.g., age, gender), event information (e.g., external cause, place, activity), temporal information (e.g., day of week, month), and injury information (e.g., diagnosis, length of stay). A detailed outline of the variable fields is included in Appendix E (Table E.2).

The time taken from application (via PHA) for approval and the data being provided was 12 weeks.

The time taken from application for approval and the data being provided was 8 weeks.

5.3.1.3 *Emergency Department Information System*

All emergency presentations with an ICD discharge diagnosis with an ICD-10-AM code between S00-S99 and T00-T98 (Chapter 19: Injury, Poisoning, and Certain Other Consequences of External Causes) from 1st of January 2005 to the 31st of December 2010 were provided from the following hospitals:

- Beaudesert Hospital
- Bundaberg Hospital
- Caboolture Hospital
- Cairns Base Hospital
- Caloundra Hospital
- Gladstone Hospital
- Gold Coast Hospital
- Gympie Hospital
- Hervey Bay Hospital
- Innisfail Hospital
- Ipswich Hospital
- Logan Hospital
- Mackay Base Hospital
- Maryborough Hospital
- Mt Isa Base Hospital
- Nambour Hospital
- Prince Charles Hospital
- Princess Alexandra Hospital
- QEII Jubilee Hospital
- Redcliffe Hospital
- Redlands Hospital
- Robina Hospital
- Rockhampton Hospital
- Royal Brisbane Hospital
- Royal Children’s Hospital
- Toowoomba Base Hospital
- Townsville Hospital
- Wynnum Hospital
- Yeppoon Hospital

It should be noted that, with the exception of Townsville Hospital, data was not being collected in these hospitals for the entire study period.

In total, there were 1,296,204 cases. All injury cases were included in the request so the identification of transport injury could be assessed for validity. Data were provided in a comma separated text (txt) file. This file was exported to SPSS 19 for data coding and analysis.

Variable fields provided included event (e.g., date, presenting complaint), patient (e.g., age, gender), and injury information (e.g., diagnosis, triage priority). For more details, see Appendix E (Table E.4).

The time taken from application (via PHA) for approval and the data being provided was 5 weeks.

5.3.1.4 *Queensland Injury Surveillance Unit*

The QISU data include all patients presenting with an injury in 29 participating hospitals in Queensland. The following hospitals are included in the data set provided to the researcher as part of this project:

- Atherton Hospital[3]
- Bundaberg Hospital[1]
- Cherbourg Hospital[1]
- Clermont Hospital
- Collinsville Hospital[1]
- Dysart Hospital
- Gatton Hospital[1]
- Gold Coast Hospital[1]
- Hughenden Hospital[1]
- Innisfail Hospital[1]
- Logan Hospital[4]
- Mackay Hospital
- Mareeba Hospital[2]
- Maryborough Hospital[1]
- Mater Adult Public Hospital[2]
- Mater Children's Public Hospital
- Redland Hospital[2]
- Mater Hospital Mackay
- Moranbah Hospital
- Mount Isa Hospital
- Nanango Hospital[2]
- Princess Alexandra Hospital[2]
- Proserpine Hospital
- QEII Jubilee Hospital
- Richmond Hospital[1]
- Royal Children's Hospital
- Sarina Hospital
- Tully Hospital[2]
- Warwick Hospital[1]
- Yeppoon Hospital[1]

The data for this study included all cases from 1st of January 2005 to 31st December 2010, totalling 275,903 cases. All injury cases were included in the data request, so that the coding of transport injury could be examined for validity. Data were provided in a Microsoft Excel 2003 (xlsx) file. This file was exported to SPSS 19 for data coding and analysis.

[3] These hospitals joined the collection sometime after 1st of January 2005, so do not have data for the full study period

[4] These hospitals are now inactive, so do not have data for the full study period

Variable fields included patient demographics (e.g., age, gender), temporal (e.g., day of week, time of presentation, month), event information (e.g., external cause, place, activity), and injury information (e.g., diagnosis codes, triage score). A detailed description of the variable fields included is included in Appendix E (Table E.3).

The time taken from application (via PHA) for approval and the data being provided was 10 weeks.

5.3.1.5 *eARF (Queensland Ambulance Service)*

All cases attended by an ambulance in Queensland that involved a case nature coded as 'motor vehicle collision', 'motorcycle collision', 'bicycle collision', 'pedestrian collision', 'crush', and 'fall' between 1st January, 2007 and 31st December 2010 were provided, totalling 269,753 cases (the selection of these cases will be described in Section 5.3.2). The inclusion of 'crush' and 'fall' was based on advice from the data custodian as it was suggested that some transport cases may be coded in these categories. Data were provided in a comma separated variable (csv) file. This file was exported to SPSS 19 for data coding and analysis.

5.3.1.6 *National Coronial Information System*

Access to a secure web-based interface was provided to the researcher. The data collection includes all reported deaths in Queensland from 2001 (only accessed 2005-2010). The data include all injury deaths in Queensland as they are all reportable to the Coroner. Information includes: administrative; demographic; and incident information. Other information may be attached to each record including police reports, autopsy reports, toxicology reports, and coronial findings (access to these was only provided for closed cases). Variable fields include date, age, gender, work-relatedness, case type, intent, mechanism, object, activity, and ICD-10-AM code.

The time taken from application for approval, contract signing, and the data being provided was 20 weeks.

5.3.2 *Selection of road crash injuries and variables*

Cases for each data collection were selected based on their alignment with the Queensland Road Crash Data definition of a road crash injury (i.e., occurred on a public road and involved a moving vehicle). Where possible, other exclusions based on the definition in the Queensland Road Crash Data (see Chapter 3, Section 3.3.1.1) were also applied (e.g., intentional acts, pedestrian colliding with a railway train). For each data collection, a conservative approach was taken in the selection of cases. Only cases that were coded or directly identified in text were included. If a case was coded as unknown, unspecified, or other category it was not included even though it may be a road crash case. For the purposes of examining validity (Section 5.4.3) and completeness (Section 5.4.1), cases outside the selection of road crash (i.e., all transport injuries) were included in the analyses. This was done in order to be able to assess the validity and completeness of variables and selection criteria used in the selection of road crash cases (e.g., traffic coding).

For exploration of completeness (Section 5.4.1), all variables included in each data set that represents the Core Minimum, Core Optional, and Supplemental Data Set variables were examined.

In order to conduct analyses for Sections 5.4.3 (validity) and 5.4.4 (representativeness), the following variables were used (where possible) for each data set:

> *Age* was coded into 5 year age groups (with the exception 85+). It should be noted that data in some data collections were not provided in 5 year blocks after 85+ due to potentially small cell sizes that may lead to identification of cases.
>
> *Gender* (1 = Female; 2 = Male). Some data sets refer to sex rather than gender, however, gender will be the term used throughout the chapter
>
> *Severity of injury* was measured by three variables: *Broad severity*, *Abbreviated Injury Scale, and Survival Risk Ratios*.

1. *Broad severity* was coded into three levels (fatality; 'hospitalisation'; other injury). These categories were chosen as it was possible to code each of the data sets into these categories, or capture one of these categories entirely (e.g., NCIS – only fatalities, QHAPDC – only hospitalisations). These categories are also the basis for how severity is generally captured across jurisdictions. It should be noted that for the purposes of this categorisation, 'hospitalisation' will be treated as 'taken to hospital' as defined by the QRCD.
2. *The Abbreviated Injury Scale (AIS)* is body-region based coding system developed by the Association for the Advancement of Automotive Medicine (AAAM, 2008). A single injury is classified on a scale from 1-6 (1 = minor; 2 = moderate; 3 = serious; 4 = severe; 5 = critical; and 6 = maximum). If there is not enough information to assign a value, a code of 9 (not specified) is applied. For the purposes of this study, the AIS score was mapped to principal diagnosis ICD-10-AM codes in the data (when available). A tool for mapping ICD-10 codes to AIS score was sourced from the European Center for Injury Prevention. While this mapping is for ICD-10 to AIS, not ICD-10-AM, the principal diagnosis coding is compatible between the systems at a lower level of specificity (4th character).
3. *Principal diagnosis* SRRs were mapped to principal diagnosis ICD codes as used by Stephenson, Henley, Harrison, and Langley (2003). It should be noted that it was not possible to calculate ICISS, which is a more comprehensive assessment of injury severity than SRR alone. This was because, to calculate ICISS information on all the injuries a patient suffers requires the calculation of the multiplication of SRRs

for each injury, and each data set (apart from the hospitalisation data set) only provided the principle diagnosis.

In order to specifically explore issues of serious injury definitions, three classifications of *serious injuries* were also derived.

1. SRRs equal to or less than 0.941 were coded as serious with all other values coded as non-serious. This criterion was based on the work of Cryer and Langely (2006).
2. All those with an AIS of 3 or greater were classified as serious, the rest as non-serious. This classification is based on the designation described in the AIS manual (Association for the Advancement of Automotive Medicine, 2008)
3. All those coded as ‘hospitalised’ and ‘fatal’ were classified as serious, the rest as non-serious. This classification is consistent with the definitions used by many jurisdictions for police based crash data systems (D'Elia & Newstead, 2010)

Accessibility/Remoteness Index of Australia (ARIA+) broadly classifies geographic areas based on their distance from the five nearest major population centres (National Centre for Social Applications of GIS, 2009). ARIA+ is categorised into five groups (1 = Major Cities; 2 = Inner Regional; 3 = Outer Regional; 4 = Remote; 5 = Very Remote). Some of the data sets included ARIA+ classifications, while others provided postcode. In cases where postcode was provided without ARIA+, postcodes were mapped to ARIA+ using data from the Australian Bureau of Statistics (2013). Some postcodes map to multiple ARIA+ categories, so in these cases the postcode is assigned to the ARIA+ category that has the largest proportion of the population.

Road user was coded into five categories (1 = Driver, 2 = Motorcyclist, 3 = Cyclist, 4 = Pedestrian; 5 = Car passenger).

Year of event (2005; 2006; 2007; 2008; 2009; 2010)

These variables were chosen as a result of the literature review and Study 1, which indicated that these factors may differentially impact on the quality of data, and are key factors to explore in relation to establishing the nature of injuries in road safety research. An outline of the selection of cases and variables (including any coding or recoding of variables) for each data set are detailed below. A summary of the selection criteria for cases and the coding of variables are available in Appendix F as a pull-out A3 sheet for reference.

5.3.2.1 *Queensland Road Crash Database*

By definition, all cases in the QRCD for the study period were included in the analyses with the exception of comparisons with QHAPDC, EDIS, and QISU (in which only ‘hospitalised’ and fatality cases were included) and NCIS (in which only fatality cases were included).

The coding of variables was as follows:

> *Age* was provided in years, and was coded into 5 year age groups (with the exception of 85+).
>
> *Gender* was retained as coded (1 = Female; 2 = Male).
>
> *Broad severity* was coded from the variable *casualty severity* (1= fatality; 2 = hospitalisation; 3 = medical treatment; 4 = minor injury), with ‘medical treatment’ and ‘minor injury’ collapsed into the ‘other injury’ category.
>
> *AIS* and *SRR*, was coded using the *injury description* variable. This variable, while a text description, is recorded in a standard form that is the same as those of the ICD-10-AM principal diagnosis descriptions. This allowed a principal diagnosis ICD-10-AM code to be mapped to each injury description. These ICD codes were then mapped to the AIS and a SRR using processes described previously in Section 5.3.2.
>
> *ARIA+* was an already coded variable in the data, so was retained in its original form.
>
> *Road user* was categorised using the variable *casualty road user type*. The original variable coding was retained from this variable with the exception of ‘motorcycle pillions’ and ‘bicycle pillions’. These two classifications were collapsed into the ‘motorcyclist’ and ‘cyclist’ categories respectively.

5.3.2.2 *Queensland Hospital Admitted Patients Data Collection*

As stated earlier, there were 75,495 land-transport cases identified. Table 5.1 includes details of the different coding groups in this selection.

Table 5.1: Number of coded land transport incidents in QHAPDC 2005-2010

Transport accidents (V00-V99)	N	% of all cases
Pedestrian injured (V00-V09)	4,502	0.6
Pedal cyclist injured (V10-V19)	12,337	1.7
Motorcycle rider injured (V20-V29)	23,490	3.3
Occupant of three-wheeled motor vehicle injured (V30-V39)	67	0.0
Car occupant injured (V40-V49)	22,074	3.1
Occupant of pick-up truck or van injured (V50-V59)	547	0.1
Occupant of heavy transport vehicle injured (V60-V69)	1,348	0.2
Bus occupant injured (V70-V79)	511	0.1
Other land transport (V80-V89)	10,619	1.5
Total land transport	**75,495**	**10.6**

Using the fourth character in the ICD-10-AM external cause code to identify whether an incident was traffic or non-traffic, 43,991 (67.8%) of land transport cases were classified as traffic. As noted previously (see Section 5.3.2), while other cases could have arisen from road crashes, an approach was taken to only include those cases that were directly coded as a road crash case by using traffic status which has been used elsewhere (Henley & Harrison, 2011). It is noted that the *place* variable could also be used and this issue is discussed in later sections (see Section 5.4.3.1).

Other exclusions were also made due to cases not fitting the definition of a road-crash. Specifically, when the injury resulted from a pedestrian colliding with a pedestrian conveyance (V00) (*n* = 5) or a railway train (V05) (*n* = 6) it was not included. Also, all transfers, as identified by *separation mode* (*n* = 6,390) were excluded to partly eliminate multiple counts of cases (Berry, Harrison, & Bureau, 2008). The final number of road crash cases identified in QHAPDC was 37,480 (6.4% of total non-transfer cases).

Variables (as specified in Section 5.3.2) that were selected, created and/or recoded were as follows:

Age was provided in 5 year age groups (with the exception of 85+).

Gender was re-coded to be consistent with other data collections (1 = Female; 2 = Male).

Broad severity was defined using the *mode of separation* variable, with those coded as ‘died in hospital’ categorised as a fatality and all other cases categorised as ‘hospitalised’.

AIS and *SRR*, was coded using the *principal diagnosis* ICD-10-AM codes. These ICD codes were then mapped to the AIS and a SRR using processes described in Section 5.3.2.

ARIA+ was an already coded variable in the data, so was retained in its original form.

Road user was categorised using the second and fourth characters of the ICD-10-AM external cause code. The breakdown of this classification is presented in Table 5.2.

Table 5.2: ICD-10-AM external cause codes for road user categorisation for QHAPDC

ICD external cause code	Road user category
V4x5; V5x5; V6x5; V7x5	1 = Driver
V2x4; V2x5; V2x8; V2x9; V3x5; V3x6; V3x8; V3x9	2 = Motorcyclist
V1x4; V1x5; V1x8; V1x9	3 = Cyclist
V0x1	4 = Pedestrian
V4x6; V5x6; V6x6; V7x6	5 = Passenger

5.3.2.3 *Emergency Department Information System*

Transport cases were identified by applying a keyword search on the variable *presenting problem*. Relevant keywords were identified as those that were present in the text description for coded transport cases in QISU (e.g., car, motorbike, pedestrian). A full list of text terms are presented in Appendix G. In total 112,747 cases were identified as including these keywords. In order to identify road crash cases, exclusions keywords based on non-traffic locations or vehicle types that are used primarily for off-road use were identified (e.g., trail, off-road, path, quad bike). A full list of exclusion terms are presented in Appendix G. After these exclusions were applied, there were 90,640 road crash cases. Transfers were also excluded, using the variable *departure status* to reduce the chance of double-counting cases, leaving a total of 88,829 cases.

Age was provided in years, and was coded into 5 year age groups (with the exception 85+).

Gender was retained as coded (1 = Female; 2 = Male).

Broad severity was coded based on the variable *departure status* as presented in Table 5.3.

Table 5.3: Classification of broad severity based on departure status for EDIS

Departure status	Broad Severity
Died in ED	1 = Fatality
Admitted	2 = Hospitalisation
ED service completed – discharged; Left after treatment commenced[5]	2 = Hospitalisation

AIS and *SRR*, was coded using the principal diagnosis ICD-10-AM codes. These ICD codes were then mapped to the AIS and a SRR using processes described in Section 5.3.2.

ARIA+ was coded from postcode using the method specified in Section 5.3.2.

Road user was categorised using text identification of the *presenting complaint* variable. The text keywords relating to road users are presented in Table 5.4.

Table 5.4: Classification of road user from keywords in 'presenting complaint' for EDIS

Keyword examples	Road user
Driver	1 = Driver
Motorcycle, MCA, MBA, motorbike	2 = Motorcyclist
Bicycle, cyclist, PBC, PBA	3 = Cyclist
Pedestrian	4 = Pedestrian
Passenger	5 = Passenger
None of the above keywords	98 = Unspecified

5.3.2.4 *Queensland Injury Surveillance Unit*

Transport injuries were selected in the QISU data set by using the *external definition* field and included cases coded as:

- Motor vehicle – driver
- Motor vehicle – passenger
- Motorcycle – driver
- Motorcycle – passenger
- Pedal cyclist or pedal cyclist passenger
- Pedestrian

[5] These cases are in line with the definition of 'hospitalised' in QRCD which is 'taken to hospital'.

Table 5.5: Number of coded transport incidents in QISU 2005-2010

External Code	N	% of all cases
Motor vehicle - driver	4,844	1.8
Motor vehicle - passenger	3,438	1.2
Motorcycle - driver	6,610	2.4
Motorcycle – passenger	251	0.1
Pedal cyclist or pedal cyclist passenger	7,202	2.6
Pedestrian	982	0.4
Coded transport total	**23,327**	**8.5**

The variable *type of place* was used to identify road crash injuries. When *type of place* was coded as ‘Street or highway’ it was considered a road crash injury (n = 13,077). Further exclusions were applied based on the definition of a road crash injury (as specified in QRCD). Specifically, intentional cases and cases of pedestrians colliding with a railway train were excluded. Transfers were also excluded to reduce the chance of double-counting cases. The final number of road crash injuries for analysis was 12,509.

Age was provided in years, and was coded into 5 year age groups (with the exception of 85+).

Gender was retained as coded (1 = Female; 2 = Male).

Broad severity was coded based on the variable *mode of separation* as presented in Table 5.6.

Table 5.6: Classification of broad severity based on mode of separation for QISU

Mode of separation	Broad Severity
Died in ED; Dead on arrival	1 = Fatality
Admitted	2 = Hospitalisation
ED service completed – discharged; Left after treatment commenced[6]	2 = Hospitalisation

AIS and *SRR*, was coded using the principal diagnosis ICD-10-AM codes. These ICD codes were then mapped to the AIS and a SRR using processes described in Section 5.3.2.

ARIA+ was coded from *postcode* using the method specified in Section 5.3.2.

[6] These cases are in line with the definition of ‘hospitalised’ in QRCD which is ‘taken to hospital’.

Road user was categorised using the external code variable. The breakdown of this classification is presented in Table 5.7.

Table 5.7: Classification of road user based on external code for QISU

External code	Road user
Motor vehicle – driver	1 = Driver
Motorcycle – driver; Motorcycle passenger	2 = Motorcyclist
Pedal cyclist or pedal cyclist passenger	3 = Cyclist
Pedestrian	4 = Pedestrian
Motor vehicle – passenger	5 = Passenger

5.3.2.5 *eARF (Queensland Ambulance Service)*

For the eARF collection inclusion was based on cases with a *case nature* coded as:

- Bicycle Collision
- Motor Vehicle Collision
- Motorcycle Collision
- Pedestrian Collision

As mentioned earlier (see Section 5.3.2), while some other cases included in the data collection were potentially transport-related, only those directly coded as transport were included.

Table 5.8: Number of coded transport incidents in eARF 2007-2010

Case nature	N	% of all cases
Motor Vehicle Collision	45,731	18.2
Motorcycle Collision	5,832	2.3
Bicycle Collision	4,254	1.7
Pedestrian Collision	729	0.3
Coded transport total	**56,546**	**22.5**

In order to identify the cases that occurred on-road, the variable location type was used. Cases categorised with a *location type* of ‘street’, ‘public place’, or ‘vehicle’ were included (n = 40,070).

Age was calculated from date of birth and was coded into 5 year age groups (with the exception of 85+).

Gender was retained as coded (1 = Female; 2 = Male).

Broad severity was not able to be coded as there was no variable to determine it.

AIS and *SRR* were not able to coded, due to lack of specific information about the injury.

ARIA+ was coded from *postcode* using the method specified in section 5.3.2.

Road user was coded manually by reviewing a combination of *case nature*, *vehicle type* and keywords within the *comments* variable. This combination was used as it was not possible to identify passengers and drivers using case nature or vehicle type alone. It was also possible that some motor vehicle collisions also referred to motorcycle collisions. The details of this selection are presented in Table 5.9.

Table 5.9: Case nature and vehicle type for road user categorisation for eARF

Case nature	Vehicle type	Comment keyword	Road user
Motor vehicle collision		Driver	1 = Driver
Motor vehicle collision	Motorcycle		
Motorcycle collision			2 = Motorcyclist
Bicycle collision			3 = Cyclist
Pedestrian collision			4 = Pedestrian
Motor vehicle collision		Passenger	5 = Passenger
Motor vehicle collision			98 = Unspecified

5.3.2.6 *National Coronial Information System*

To select road crash injuries, the first step involved selecting cases that were coded as being transport-related. For the NCIS collection this included cases with a *primary mechanism* code of 'blunt force' and a *secondary mechanism* code of 'transport incident'. In total, there were 2,311 transport cases identified. In order to determine the cases that were land transport, the object variable was used to exclude water and air-related cases. After removal of these cases, there were 2,227 land-transport cases. The traffic status of the cases, used to determine a road crash injury, was determined by the variable *context*. Only those coded as 'Land Transport Traffic Injury Event' were included, leaving 2,009 cases. Other exclusions were also made due cases not fitting the definition of a road-crash. Specifically, only those with an *intent code* of 'unintentional' and a case type of 'Death due to External Cause(s)' were retained (n = 1,961).

Variables (as specified in Section 5.3.2) were selected, created and/or recoded as follows:

Age was provided in years and was classified into 5 year age groups (with the exception of 85+).

Gender was retained as coded (1 = Female; 2 = Male).

Broad severity, *AIS* and *SRR* were not determined for this data set as all cases are fatalities.

ARIA+ was coded from *postcode* using the method specified in Section 5.3.2.

Road user was categorised using a combination of the variables *user code* and *mode of transport* (see Table 5.10).

Table 5.10: User code and mode of transport for road user categorisation for NCIS

Mode of transport	User code	Road user
Light Transport Vehicle	Driver Rider or Operator	
Heavy Transport Vehicle	Driver Rider or Operator	
Special All-Terrain Vehicle	Driver Rider or Operator	1 = Driver
Two-wheeled motor vehicle		
Three-wheeled motor vehicle		2 = Motorcyclist
Pedal Cycle		3 = Cyclist
Pedestrian		4 = Pedestrian
Light Transport Vehicle	Passenger	
Heavy Transport Vehicle	Passenger	
Special All-Terrain Vehicle	Passenger	5 = Passenger
Unspecified and other specified mode of transport		98 = Unspecified

5.3.3 *Analysis*

5.3.3.1 *Assessing completeness*

Completeness in terms of the field completeness (i.e., the amount of missing and/or unspecified data) was examined for each data set, by identifying the proportion of: ‘missing’; ‘unknown’; ‘other specified’; and ‘unspecified’ values recorded for key variables outlined in the WHO guidelines for Core Minimum, Core Optional, and Supplemental Datasets using frequencies. The completeness of the information required for the identification of road crash injury cases in each data collection was also assessed using frequencies. It should be noted that variables in each data set relating to the date of injury are not able to be assessed for completeness as, based on the extraction timeframe criteria for each data set, any cases with missing or unknown injury dates would not be included by definition.

5.3.3.2 *Assessing consistency*

The consistency of: missing; unknown; other specified; and unspecified data was examined across a number of variables including: year, ARIA+, broad severity, gender, age, and road user group (where possible). The examination was restricted to Core Minimum, Core Optional, and Supplemental Dataset variables that were included in the data set and had more than 10% ‘missing’, ‘unspecified’, and/or ‘other’ coded cases. This threshold was based on recommendations from a number of sources that more than 10% missing data should be further explored for bias (e.g., Bennett, 2001) . For the QRCD, all cases were included for comparison. For all other data sets, the cases for comparison were those cases identified as transport-related cases. Comparisons were made using Chi-square tests of independence. Due to the large sample size, a more stringent alpha of .001 was adopted. Also, Cramer's V (ϕ_c) was calculated in order to provide an estimate of effect size to give a clearer idea of the meaningfulness of any statistical significance found. As suggested by Aron and Aron (1991), a Cramer's V of less than .10 was considered to be a small effect size, between .10 and .30 moderate, and more than 0.30 a large effect size. Post-hoc analyses were also undertaken using an adjusted standardised residual statistic. This statistic can be used to identify those cells with observed frequencies significantly higher or lower than expected. With an alpha level set at 0.001, adjusted standardised residuals outside -3.10 and +3.10 were considered significant (Haberman, 1973).

5.3.3.3 *Assessing validity*

As there is no gold-standard for the validity of the data collections, it is only possible to assess validity in broad terms, such as the coding of variables and the selection processes. For some data collections, it was possible to use other variables (e.g., text descriptions) within the data collection to illuminate validity issues in the selection processes and key variables. For the purposes of this process, the text description (or other variable) will be used as the proxy gold-standard or reference standard. In each case, the reference standard is presumed to be a more accurate way to identify the characteristic than the variable being evaluated.Validity in this instance will be discussed in terms of sensitivity and specificity. Sensitivity refers to the proportion of actual cases (as determined by the proxy) which are correctly identified. Specificity refers to the proportion of negatives which are correctly identified.

Sensitivity was reported using the following formula:

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives + number of false negatives}}$$

Specificity was reported using the following formula:

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives + number of false positives}}$$

The classification of true positives, false positives, true negatives, and false negatives are shown in Table 5.11.

Table 5.11: Characterisation of true positives, false negatives, false positives, and true negatives

		Reference standard	
		True	False
Coding classification	True	True Positive	False Positive
	False	False Negative	True Negative

The details of how specificity and sensitivity was assessed for each of the data collections, in which it was possible, are described below.

Queensland Hospital Admitted Patients Data Collection

It is not possible to assess the validity of any variable within QHAPDC directly, as there are no variables or fields that can be used as a benchmark for any other. However, it is possible to explore possible validity issues with the traffic status coding used to select on-road cases. As this is the basis for the selection of cases and there has been some suggestion in the literature of traffic coding being inaccurate (McKenzie & McClure, 2010) it was important to explore this variables validity.

The ICD-10-AM coding guidelines (National Centre for Classification in Health, 2004) specify the following in relation to coding an injury as traffic:

> "*A traffic accident is any vehicle accident occurring on the public highway [i.e. originating on, terminating on, or involving a vehicle partially on the highway]. A vehicle accident is assumed to have occurred on the public highway unless another place is specified, except in the case of accidents involving only off-road motor vehicles, which are classified as non-traffic accidents unless the contrary is stated*" (National Centre for Classification in Health, 2004)

Based on this coding principle, there should be consistency between the *place variable*, *mode of transport (off-road or on-road vehicle)*, and *traffic status*. In order to assess this consistency the traffic coding was compared to the *place* variable and *mode of transport* from the ICD-10-AM external cause code. Off-road vehicles were those coded as V83 – V86 (e.g., Occupant of special vehicle mainly used on industrial premises injured in transport accident, Occupant of special all-terrain or other motor vehicle designed primarily for off-road use).

Queensland Injury Surveillance Unit

Text descriptions (*injury description*) were manually reviewed for a random sample (n = 1000) of cases to assess the selection of transport-related cases for QISU. The sensitivity and specificity of the transport coding was calculated by comparing the manual review (reference standard) with the *external definition* coding.

The validity of the place variable was also assessed by manually reviewing the random selection of transport-related cases. Similar to eARF, the assessment of the on-road status in text was conducted in line with ICD-10-AM coding rules described above. The result of this manual text review was compared to the *place* variable coded to calculate sensitivity and specificity.

As discussed in Section 5.3.2.4, road user was classified by the *external definition* variable. The validity of this classification was assessed with a manual text review of a random sample (n = 1000) of cases. In order to assess the validity of the road user classification, a manual text review of 1000 cases was conducted on the text description. The sensitivity and specificity of the road user coding was calculated. The characterisation of true positives, false negatives, false positives, and true negatives are presented in Table 5.12.

Table 5.12: Characterisation of true positives, false negatives, false positives, and true negatives for road user classification for QISU

Road user	True positives	False negatives	False positives	True negatives
Driver	Classified as driver, driver in text	Not classified as driver, driver in text	Classified as driver, not driver in text[1]	Not classified as driver, not driver in text[1]
Motorcyclists	Classified as motorcyclist, motorcyclist in text	Not classified as motorcyclist, motorcyclist in text	Classified as motorcyclist, not motorcyclist in text[1]	Not classified as motorcyclist, not motorcyclist in text[1]
Cyclists	Classified as cyclist, cyclist in text	Not classified as cyclist, cyclist in text	Classified as cyclist, not cyclist in text[1]	Not classified as cyclist, not cyclist in text[1]
Pedestrian	Classified as pedestrian, pedestrian in text	Not classified as pedestrian, pedestrian in text	Classified as pedestrian, not pedestrian in text[1]	Not classified as pedestrian, not pedestrian in text[1]
Passenger	Classified as passenger, passenger in text	Not classified as passenger, passenger in text	Classified as passenger, not passenger in text[1]	Not classified as passenger, not passenger in text[1]

[1] This refers to cases where another road user is actually identified as the injured person in the text or it refers to something other than the road user. It is not considered a false positive if the case does not specify the road user.

eARF

For the coded transport incidents that were selected for analysis, a random sample of 1000 cases were manually text reviewed to ascertain the proportion of these cases that were not transport-related cases and therefore possibly coded incorrectly. In addition, other cases not selected as transport-related (i.e., crush and fall) may fit the definition of being transport-related (involved a moving vehicle). In order to assess whether additional transport-related cases are coded into these categories, a manual text review of 1000 randomly selected crush and fall incidents was conducted. The sensitivity and specificity of the transport coding was calculated using the results of the text review as the reference standard compared to the coded *case nature*.

A manual text review was conducted on the random sample of coded transport-related cases to ascertain the proportion of these cases that were not road crashes (e.g., off-road, speedway, driveway, property, race track). If the text did not specify where the incident occurred, it was assumed to have occurred on-road, with the exception of cases such as (trail bike, motorcross, quad bike) which were assumed to have occurred off-road. This practice is consistent with ICD-10-AM coding rules described above. The result of this manual text review (the reference standard) was compared to the *location* variable coded in the eARF file to calculate the sensitivity and specificity of the location coding.

As discussed in Section 5.3.2, the variables *case nature* and *vehicle type* were used to classify the road user of each case. The sensitivity and specificity of the road user coding was calculated, with true positives, true negatives, false positives, and false negatives characterised as described for QISU (Table 5.12).

National Coronial Information System

The validity of the identification of road crash cases and road user coding was assessed by comparing these variables to the results of a manual review of the police reports and coroner's findings of all cases. The sensitivity and specificity of the road user coding was calculated, with true positives, true negatives, false positives, and false negatives characterised as described for QISU (Table 5.12).

5.3.3.4 *Assessing representativeness*

There is no gold-standard for what is considered representative of all road crash injuries. However, as the QRCD is used primarily for road safety research, for the purposes of analysing representativeness, QRCD was used as the benchmark for the prevalence and profile of road crash injuries. The other data collections were compared to QRCD on the prevalence of road crash injuries as well as the profile of severity (where possible), road user, age, gender, and ARIA+.

Each data collection relates to QRCD (and each other) in specific ways which influence the selection of cases for comparison. Specifically, QHAPDC only has hospitalisations and fatalities (hospitalisations that result in death within 30 days) so only 'hospitalised' and fatal cases from QRCD were included for the representativeness analysis. Similarly, only ‘hospitalised’ and fatal cases in QRCD were included for comparison with EDIS and QISU also, as the definition of ‘hospitalised’ in QRCD is ‘taken to hospital’. Finally, NCIS only includes fatalities, so only fatal cases from QRCD were included for comparison when examining representativeness.

Bivariate comparisons were made between QRCD and the other data collections on each of the factors (i.e., age, gender, road user, ARIA+, severity) using Chi-square tests of independence, using the criteria described in Section 5.3.3.2. Multivariate analyses (using logistic regressions) were also conducted to allow for an examination of the relationships between the key factors and the data collection while controlling for the relationships of the key factors with each other.

For the purposes of analyses in Section 5.4.4 (representativeness), only cases in 2009 were used for each data set. This was due to comparisons needing to be made with QRCD and this was the latest available full year of data for this data set. Also, for data collections such as EDIS and eARF, this would also represent a full year of data (all included hospitals were collecting EDIS data by this year and the eARF system was in full effect). This year was also the year used for the data linkage study (Chapter 7) which will allow for comparisons between these phases of the program of research.

5.3.3.5 *Assessing serious injury definitions*

For each data collection (where possible) the different measures of severity were compared with each other. Specifically, the proportion of those classified as serious using the three different severity measures (broad severity, AIS, and SRR) will be compared for each data collection.

5.4 Results

5.4.1 *Completeness*

As discussed in Section 5.3.3, the completeness of each data collection was assessed by examining the frequencies of ‘missing’, ‘unspecified’, and/or ‘other’ coded cases for the core minimum, core optional, and supplemental variables.

5.4.1.1 *Queensland Road Crash Database*

All variables relating to the WHO core minimum, core optional, and supplemental data sets that were provided had less than 2% missing with the exception of the *nature of injury* (as indicated by *injury description*) with a total of 70,621 (73.4%) cases classified as ‘unknown’, ‘unspecified’, or missing (see Table 5.13).

Table 5.13: Missing, unknown, unspecified cases for WHO core minimum, core optional, and supplemental variables in QRCD (1st January 2005 to 31st December 2009)

	WHO variable	Variable in QRCD	% missing, unknown, unspecified
Core minimum	Age	Age	1.9%
	Gender	Gender	1.8%
	Nature of injury	Injury description	73.4%
Core optional	Time of injury	Time of crash	0.0%
	Usual residence	Origin town	0.0%
	Injury severity	Casualty severity	0.0%
Supplemental	Mode of transport	Casualty unit type	0.0%
	Road user	Road user	0.0%
	Counterpart	Controller unit type	0.1%

5.4.1.2 *Queensland Hospital Admitted Patients Data Collection*

Firstly, the completeness of the variables relating to the classification of a road crash injury was assessed. It was found that for *external cause* ICD-10-AM code, there was only a small number (n = 1,039; 0.1%) that had ‘other’ or ‘unspecified’ codes and were therefore unable to be classified as being a land transport incident. The second step involved identifying cases using the ICD-10-AM *external cause* code fourth character relating to traffic vs. non-traffic incidents. There were 10,619 (14.1%) cases of land transport-related cases with a code indicating an ‘unspecified’ value for traffic/non-traffic.

As discussed in Table 5.14, in terms of completeness of variables relating to the WHO core minimum data set that were provided, *age* and *gender* had no missing or unspecified values. Approximately one third of land transport-related cases had a code indicating an ‘unspecified’ or ‘other specified’ *place* of occurrence. For land transport cases, *activity* was coded as ‘other’ or ‘unspecified’ for approximately three-quarters of the cases. However, it should be noted that for transport injuries the coding rules dictate that if a the *activity* at the time of the injury is not specified as ‘sport’, ‘leisure’ or ‘working for an income’, ‘unspecified activity’ must be assigned (NCCH, 2009). The *nature of injury* was identified using the *diagnosis string* variable and had less than 5% unspecified (e.g., body region was specified but nature was not).

In terms of core optional data items, all variables provided had less than 5% missing or unspecified. For supplemental data there were unspecified cases (more than 10%) for *counterpart* and less than 5% for *mode of transport*. Also, in terms of being able to classify cases into *road user*, 2,126 (4.1%) road crash cases were classified as car, heavy vehicle, or bus occupants but were unable to be classified into driver or passenger categories, as this information was not specified.

Table 5.14: Missing, unknown, unspecified cases for WHO core minimum, core optional, and supplemental variables in QHAPDC (1^{st} January 2005 to 31^{st} December 2009)

	WHO variable	Variable in QRCD	% missing, unknown, unspecified
Core minimum	Age	Age	0.0%
	Gender	Sex	0.0%
	Place	Place	33.4%
	Activity	Activity	75.2%
	Nature of injury	Diagnosis string	2.4%
Core optional	Time of injury	-	-
	Usual residence	ARIA+	1.7%
	Injury severity	Diagnosis string	2.4%
Supplemental	Mode of transport	External cause string	4.1%
	Road user	External cause string	4.1%
	Counterpart	External cause string	18.1%

5.4.1.3 *Emergency Department Information System*

In order to select cases for road crash injuries, the variable *presenting problem* was used. This variable was a text description field in which text searching was used. In order to assess the completeness of this variable, a random sample of 1000 injury cases were selected for manual text review as it would have been prohibitive to review all cases. Based on this manual review, 4.6% of text descriptions did not include sufficient information that would allow the injury to be classified as transport or not. For example, the description would only include information such as ‘injury elbow’, ‘pain’, or ‘head injury’.

As shown in Table 5.15, in terms of the Core Minimum Data Set variables that were included, there were less than 2% missing or unspecified cases. Of the core optional data set variables provided, *with the exception of the narrative* (*presenting problem*) variable, described above, there were less than 5% missing or unspecified cases. Data relating to supplemental information *road user* could be derived from a text search of the *presenting problem* variable. There were more than a third of cases with insufficient information to determine the road user involved (e.g., “RTC injury”, “MVC head injury”).

Table 5.15: Missing, unknown, unspecified cases for WHO core minimum, core optional, and supplemental variables in EDIS (1st January 2005 to 31st December 2009)

	WHO variable	Variable in QISU	% missing, unknown, unspecified
Core minimum	Age	Age	0.0%
	Gender	Sex	< 0.1%
	Place	-	-
	Activity	-	-
	Mechanism	-	-
	Nature of injury	ICD-10AM diagnosis	1.6%
Core optional	Time of injury	Arrival time	< 0.1%
	Usual residence	Postcode	3.8%
	Injury severity	Triage score	< 0.1%
	Narrative	Presenting problem	4.6%[1]
Supplemental	Mode of transport	-	-
	Road user	Presenting problem	41.7%[1]
	Counterpart	-	-

[1] As determined by a random sample of 1000 cases

5.4.1.4 *Queensland Injury Surveillance Unit*

There were 36,094 (13.1%) cases where *external definition* (used to identify land transport cases) was coded 'unspecified' or 'other'. In terms of the variable used to identify whether a case was a road crash (*place*), there were 37,008 (13.4%) 'unspecified' or 'other' cases.

As shown in Table 5.16, for the core minimum data set variables *age, gender, nature of injury, and mechanism* had less than 10% of cases missing or unspecified. *Activity* however, was unspecified or 'other' for almost one third of cases. Of the core optional data set variables provided, there were less than 1% missing or unspecified cases. Data relating to supplemental information *road user* can be derived from the *external cause* variable which has been discussed previously. For *mode of transport* and *counterpart* (as measured by *major injury factor*) there were less than 2% missing, unknown, or unspecified.

Table 5.16: Missing, unknown, unspecified cases for WHO core minimum, core optional, and supplemental variables in QISU (1st January 2005 to 31st December 2009)

	WHO variable	Variable in QISU	% missing, unknown, unspecified
Core minimum	Age	Age	0.9%
	Gender	Sex	0.8%
	Place	Place	13.4%
	Activity	Activity	32.0%
	Mechanism	Mechanism	2.7%
	Nature of injury	ICD-10AM diagnosis	0.5%
Core optional	Time of injury	Arrival time	0.0%
	Usual residence	Postcode	0.3%
	Injury severity	Triage score	0.3%
	Narrative	Injury test description	< 0.1%
Supplemental	Mode of transport	Major injury factor	1.5%
	Road user	External definition	13.1%
	Counterpart	Major injury factor	1.5%

5.4.1.5 *eARF (Queensland Ambulance Data)*

For selection of cases, *case nature* (*mechanism*) was used to determine whether a case was a land transport-related injury (Section 5.3.2.5). This was also the variable used for the extraction criteria, so only specified case natures were included in the data set. Therefore, no comment can be made on the amount of 'unspecified' or missing data for this variable field.

As shown in Table 5.17, for variables included in the core minimum data set, with the exception of *Nature of injury* (as measured by the *final assessment* variable), all other variables had less than 5% missing or unspecified cases. In terms of the core optional and supplemental data, *injury severity* (as measured by *transport criticality*) had almost one third missing or unspecified cases. It should be noted that unlike QRCD, QHAPDC, QISU, and EDIS, eARF did not have a variable which could be used to classify injury severity (*broad severity*, *AIS*, or *SRR*). The *narrative* (comments) was missing for more than 10% of cases. As discussed in a previous section (Section 5.3.2.5), the variables *vehicle type* in combination with *case nature* could indicate *mode of transport* and *vehicle type* in combination with *case nature and comments* could indicate *road user*. The *vehicle type* variable had approximately 10% of cases classified as 'unknown'.

Table 5.17: Missing, unknown, unspecified cases for WHO core minimum, core optional, and supplemental variables in eARF (1st January 2005 to 31st December 2009)

	WHO variable	Variable in eARF	% missing, unknown, unspecified
Core minimum	Age	Age	4.6%
	Gender	Sex	0.8%
	Place	Place	0.5%
	Activity	-	-
	Nature of injury	Final assessment	23.9%
Core optional	Time of injury	-	-
	Usual residence	-	-
	Injury severity	Transport criticality	30.1%
	Narrative	Comments	15.1%
Supplemental	Mode of transport	Vehicle type	9.8%
	Road user	Vehicle type and comments	15.1%
	Counterpart	-	

5.4.1.6 *National Coronial Information System*

There were two cases where *context* (used to identify road crash cases) was coded 'unspecified'. As shown in Table 5.18, for the core minimum, optional, and supplemental data set variables, there were less than 5% missing or unspecified cases.

Table 5.18: Missing, unknown, unspecified cases for WHO core minimum, core optional, and supplemental variables in NCIS (1st January 2005 to 31st December 2009)

	WHO variable	Variable in NCIS	% missing, unknown, unspecified
Core minimum	Age	Age	0.0%
	Gender	Sex	0.0%
	Place	Location	0.0%
	Activity	Activity	0.4%
	Mechanism	Mechanism	< 0.1%
	Nature of injury	ICD-10AM diagnosis	0.2%
Core optional	Time of injury	Time of incident	0.0%
	Usual residence	Postcode	1.6%
	Injury severity	-	-
	Narrative	Presenting problem	2.6%
Supplemental	Mode of transport	Mode of transport	0.4%
	Road user	User code	0.3%
	Counterpart	Counterpart	0.7%

5.4.2 *Consistency*

The consistency of missing or unspecified variables on key characteristics (e.g., age, gender, ARIA+) was examined for each of the data collections. These analyses were only conducted on variables that had more than 10% missing or unspecified. A summary of these findings (where the effect size was greater than 0.1) are included in Table J.1 in Appendix J.

5.4.2.1 *Queensland Road Crash Database*

The QRCD collection was examined to determine the pattern of missing, unknown, or unspecified data for injury description by year, age, gender, road user, ARIA+, and broad severity (see Table 5.19).

The injury description was less likely to be missing or unspecified for:

- 2006 and 2007 [$\chi^2(4) = 2680.44$, $p < .001$, $\phi_c = .17$] (see Figure 5.1)
- Males and unknown gender [$\chi^2(2) = 68.07$, p < .001, ϕc = .03]
- Cyclists and pedestrians [$\chi^2(4) = 167.50$, $p < .001$, $\phi_c = .04$]
- Fatalities [$\chi^2(2) = 5669.87$, $p < .001$, $\phi_c = .24$]

It should be noted that the associated effect sizes for all of these differences, with the exception of broad severity, were small. There were no significant differences in the proportions of unspecified injury cases by age [$\chi^2(17) = 28.09$, $p < .001$] or ARIA+ [$\chi^2(5) = 7.55$, $p = .183$, $\phi_c = .01$] (see Figure 5.2).

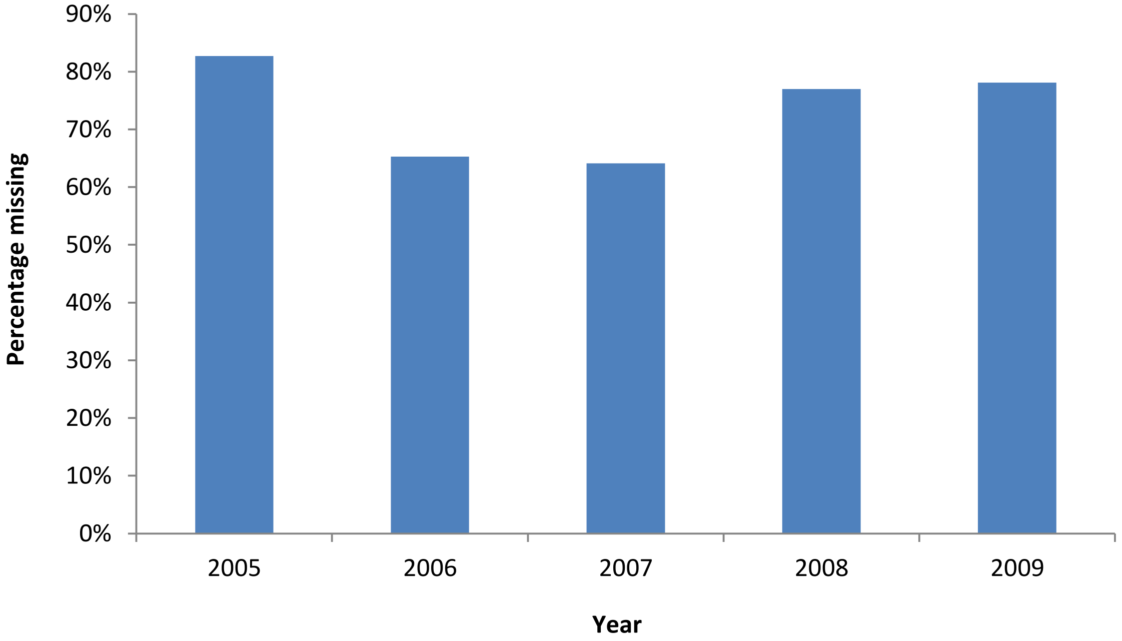


Figure 5.1: Percentage of unspecified injury description cases by year for QRCD 2005-2009

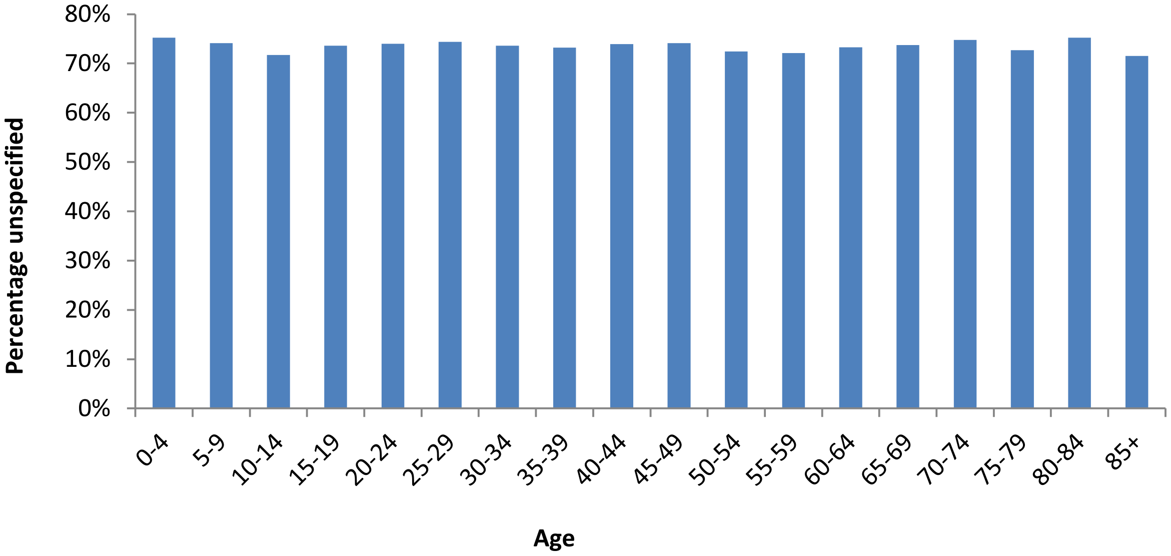


Figure 5.2: Percentage of unspecified injury description cases by age for QRCD 2005-2009

Table 5.19: Unspecified injury description by gender, road user, and ARIS+ for QRCD 2005-2009

		Injury description	
		Specified N (%)	Unspecified N (%)
Gender	Male	**14,050 (27.5)**	37,120 (72.5)
	Female	**11,446 (25.6)**	33,266 (74.4)
	Unknown	**145 (38.2)**	**235 (61.8)**
Road user	Driver	14,890 (26.3)	41,746 (73.7)
	Motorcyclist	**2,255 (24.2)**	7,047 (75.8)
	Cyclist	**1,364 (33.4)**	**2,724 (66.6)**
	Pedestrian	**1,309 (31.0)**	**2,910 (69.0)**
	Passenger	5,823 (26.4)	16,194 (73.6)
ARIA+	Major Cities	14,956 (26.6)	41,273 (73.4)
	Inner Regional	5,333 (26.0)	15,161 (74.0)
	Outer Regional	4,278 (27.4)	11,332 (72.6)
	Remote	668 (27.1)	1,799 (72.9)
	Very Remote	387 (27.9)	1,000 (72.1)
		19 (25.3)	56 (74.7)
Broad severity	Fatality	**1,651 (98.0)**	**33 (2.0)**
	Hospitalisation	**5,834 (18.4)**	**25,927 (81.6)**
	Other injury	**18,158 (28.9)**	**44,661 (71.1)**

Note: Standardised residuals outside +/-3.10 are bolded

5.4.2.2 *Queensland Hospital Admitted Patients Data Collection*

The QHAPDC data were examined to determine the pattern of missing, unknown, or unspecified data for traffic status, place, and activity, by year, age (see Figure 5.4), gender, road user, ARIA+, and broad severity (see Table 5.20).

There were no significant differences in the proportions of unspecified cases by year for traffic status [$\chi^2(5) = 9.83$, $p = .132$]; [$\chi^2(5) = 95.17$, $p = .003$]; or activity [$\chi^2(5) = 111.73$, $p = .002$] (see Figure 5.3).

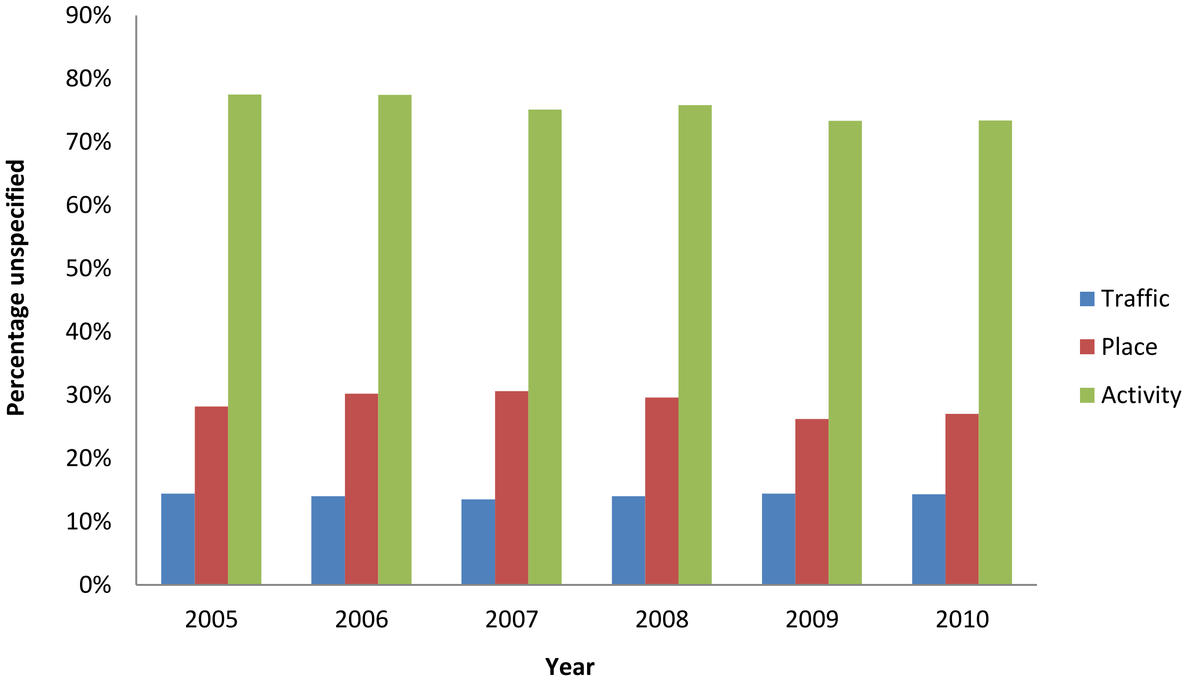


Figure 5.3: Percentage of unspecified traffic status, place, and activity cases by year for QHAPDC 2005-2010

The traffic status for the injury was more likely to be unspecified for:

- Males [$\chi^2(1) = 1416.87$, $p < .001$, $\phi_c = .14$]
- Those aged 5-9 and 45-64 [$\chi^2(17) = 314.82$, $p < .001$, $\phi_c = .07$] (see Figure 5.4)
- Inner Regional, Outer Regional, Remote, and Very Remote areas [$\chi^2(4) = 1758.30$, $p < .001$, $\phi_c = .15$]

Traffic status was less likely to be unspecified for fatalities [$\chi^2(1) = 34.43$, $p < .001$, $\phi_c = .02$].

The effect sizes associated with these differences were small, particularly for age.

Place was more likely to be unspecified for:

- Females [$\chi^2(1) = 103.33$, $p < .001$, $\phi_c = .04$]

- Those aged 0-14 [$\chi^2(17) = 2745.76$, $p < .001$, $\phi_c = .19$] (see Figure 5.4)
- Motorcyclists, cyclists, and unspecified road users [$\chi^2(6) = 11569.12$, $p < .001$, $\phi_c = .39$]
- Inner Regional areas [$\chi^2(4) = 237.22$, $p < .001$, $\phi_c = .06$]

Place was less likely to be unspecified for fatalities [$\chi^2(1) = 146.88$, $p < .001$, $\phi_c = .04$].

With the exception of road user type, the associated effect sizes for these differences were small.

Activity was more likely to be unspecified for:

- Males [$\chi^2(1) = 707.14$, $p < .001$, $\phi_c = .10$]
- Those aged 0-4 and 65+ [$\chi^2(17) = 1578.52$, $p < .001$, $\phi_c = .15$] (see Figure 5.4)
- Drivers, passengers, and pedestrians [$\chi^2(6) = 7849.75$, $p < .001$, $\phi_c = .32$]
- Very Remote areas [$\chi^2(4) = 50.34$, $p < .001$, $\phi_c = .03$]

Activity was less likely to be unspecified for fatalities [$\chi^2(1) = 55.18$, $p < .001$, $\phi_c = .03$].

As with *place*, the associated effect sizes for the differences, with the exception of road user type, were small.

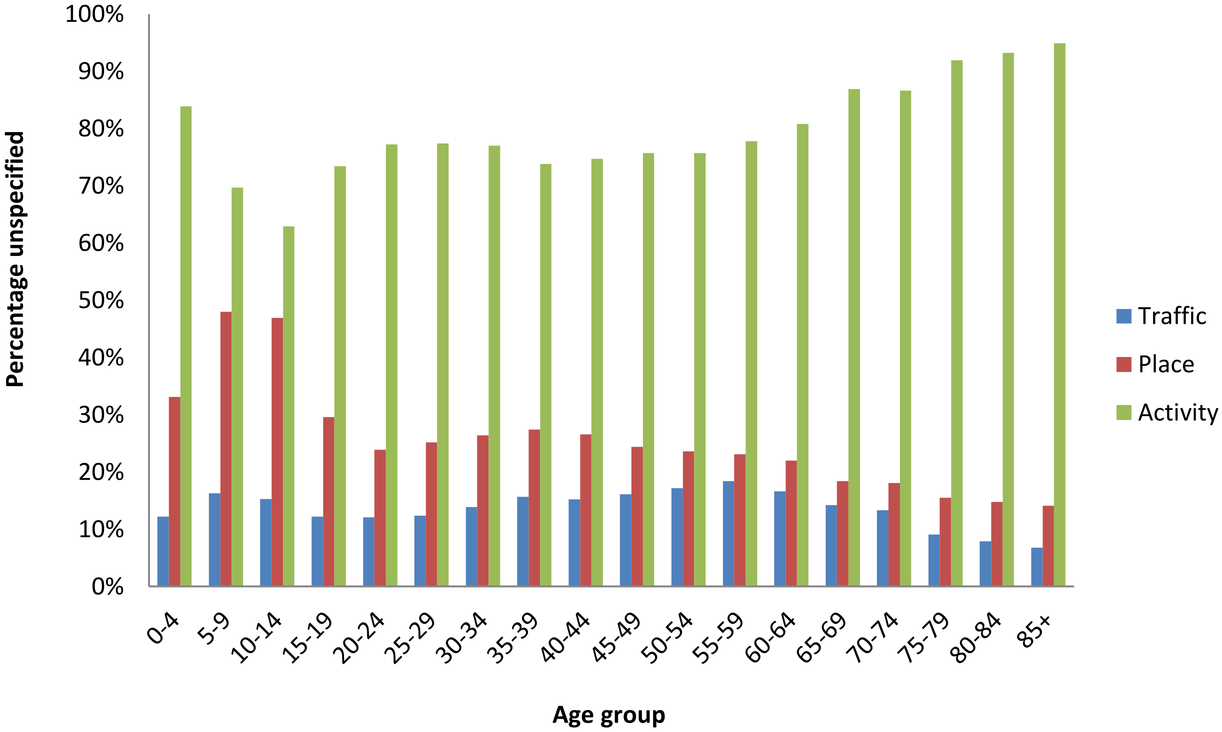


Figure 5.4: Percentage of unspecified traffic status, place, and activity cases by age for QHAPDC 2005-2010

Table 5.20: Unspecified traffic status, place, and activity by gender, road user, ARIA+, and broad severity for QHAPDC 2005-2010

		Traffic n (%)	Place n (%)	Activity n (%)
Gender	Male	**4,671 (21.6)**	**5,628 (26.0)**	**17,747 (82.0)**
	Female	**5,926 (11.0)**	**15,941 (29.7)**	**39,057 (72.8)**
Road user	Driver	-	**353 (2.7)**	**12,024 (91.2)**
	Motorcyclist	-	**7,777 (33.1)**	**16,859 (71.7)**
	Cyclist	-	**5,494 (44.5)**	**7.622 (61.9)**
	Pedestrian	-	**675 (15.0)**	**3,949 (87.9)**
	Passenger	-	**355 (5.1)**	**6,705 (96.6)**
	Car, heavy vehicle, bus occupant (not specified as driver or passenger)	-	1,260 (29.4)	3,761 (88.0)
	'Unspecified'	-	**5,655 (53.4)**	**5,884 (55.5)**
ARIA+	Major Cities	**3,056 (9.0)**	**9,171 (27.0)**	16,851 (74.9)
	Inner Regional	**3,826 (17.0)**	**7,278 (32.4)**	25,903 (76.1)
	Outer Regional	**2,604 (17.3)**	**3,984 (26.5)**	11,109 (74.0)
	Remote	**690 (28.7)**	731 (30.4)	1,829 (76.1)
	Very Remote	**421 (30.5)**	405 (29.4)	**1,112 (80.6)**
Broad severity	Fatality	**30 (5.4)**	**30 (5.4)**	**492 (89.0)**
	Hospitalisation	10,567 (14.1)	10,567 (28.8)	56,312 (75.3)

Note: Standardised residuals outside +/-3.10 are bolded. All unknown traffic cases were cases where the road user was also unknown so this comparison was not performed.

5.4.2.3 *Emergency Department Information System*

The differences in the proportion of unspecified road user (identified by text search) in EDIS were explored by year, gender, age, ARIA+, and broad severity (see Table 5.21).

The road user type was more likely to be unspecified for:

- 2009 and 2010 [$\chi^2(5) = 668.40$, $p < .001$, $\phi_c = .09$] (see Figure 5.5)
- Those aged between 5-19 [$\chi^2(18) = 1131.30$, $p < .001$, $\phi_c = .22$] (see Figure 5.6)
- Females [$\chi^2(1) = 1958.30$, $p < .001$, $\phi_c = .15$]
- Major Cities and unknown areas [$\chi^2(5) = 172.82$, $p < .001$, $\phi_c = .04$]

With the exception of age, the effect sizes associated with these differences were small. There was no statistically significant difference in the proportion of unspecified road user cases by broad severity [$\chi^2(1) = 4.23$, $p = .040$].

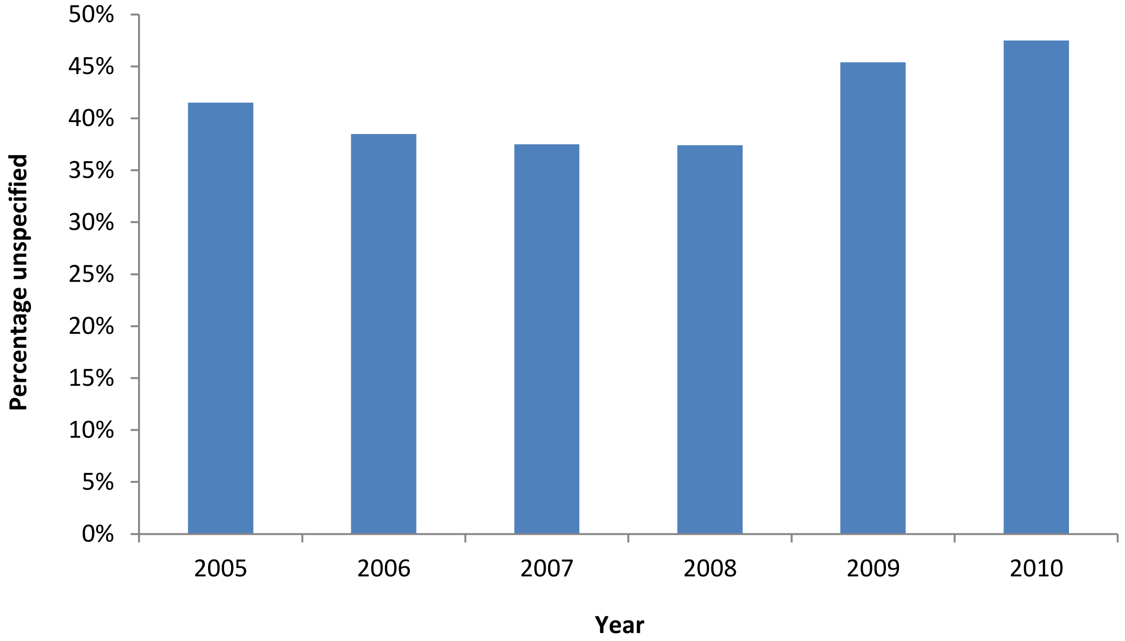


Figure 5.5: Percentage of unspecified road user cases by year for EDIS 2005-2010

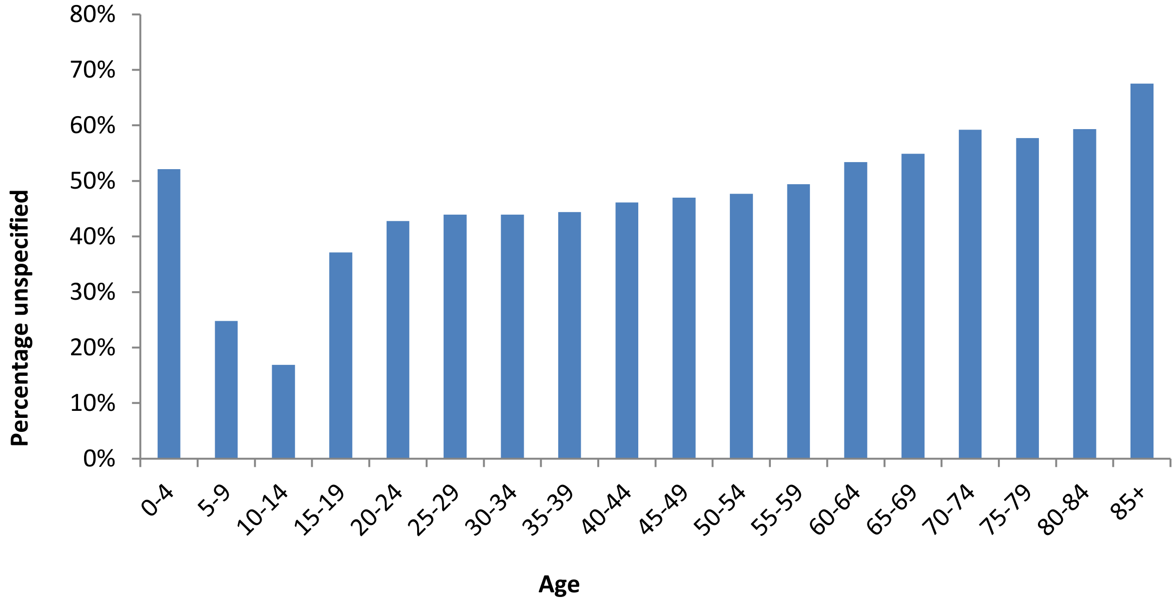


Figure 5.6: Percentage of unspecified road user cases by age for EDIS 2005-2010

Table 5.21: Unspecified road user by gender, ARIA+, and broad severity for EDIS 2005-2010

		Road user	
		Specified N (%)	Unspecified N (%)
Gender	Male	**36,392 (63.8)**	**20,646 (36.2)**
	Female	**15,417 (48.5)**	**16,351 (51.5)**
ARIA+	Major Cities	**25,261 (56.8)**	**19,209 (43.2)**
	Inner Regional	**14,626 (60.6)**	**9,521 (39.4)**
	Outer Regional	**8,077 (60.6)**	**5,259 (39.4)**
	Remote	240 (57.7)	176 (42.3)
	Very Remote	1,343 (61.1)	856 (38.9)
	Unknown	**2,267 (53.2)**	**1,994 (46.8)**
Broad severity	Fatality	22 (44.0)	28 (56.0)
	Hospitalisation	51,792 (58.3)	36,987 (41.7)

Note: Standardised residuals outside +/-3.10 are bolded

5.4.2.4 *Queensland Injury Surveillance Unit*

The pattern of missing, unknown, or unspecified data in the QISU data collection was examined for the *place* and *activity* variables by year, age, gender, road user, ARIA+, and broad severity (see Table 5.22).

Place was more likely to be unspecified for:

- 2010 [$\chi^2(5) = 76.66$, $p < .001$, $\phi_c = .06$] (see Figure 5.7)
- Males [$\chi^2(2) = 224.88$, $p < .001$, $\phi_c = .10$]
- Motorcyclists [$\chi^2(4) = 1332.17$, $p < .001$, $\phi_c = .24$]
- Inner Regional, Outer Regional, and Very Remote areas [$\chi^2(5) = 213.33$, $p < .001$, $\phi_c = .10$]

Place was less likely to be unspecified for those aged 0-9 [$\chi^2(18) = 274.48$, $p < .001$, $\phi_c = .11$] (see Figure 5.8).

With the exception of road user type, the associated effect sizes for all of these differences were small. There was no statistically significant difference in the proportion of unspecified *place* by broad severity [$\chi^2(1) = 0.172$, $p = .679$].

Activity was more likely to be unspecified for:

- 2010 [$\chi^2(5) = 237.46$, $p < .001$, $\phi_c = .10$] (see Figure 5.7)
- Females [$\chi^2(2) = 815.37$, $p < .001$, $\phi_c = .19$]
- Those aged 0-4 and 65+ [$\chi^2(17) = 1578.52$, $p < .001$, $\phi_c = .15$] (see Figure 5.8)
- Drivers, passengers, and pedestrians [$\chi^2(4) = 4981.41$, $p < .001$, $\phi_c = .46$]

- Inner Regional and unknown areas [$\chi^2(5) = 528.95$, $p < .001$, $\phi_c = .15$]

Activity was less likely to be unspecified for those aged 5-14 [$\chi^2(18) = 1131.30$, $p < .001$, $\phi_c = .22$].

As with *place*, the associated effect sizes for the differences, with the exception of road user type, were small. There was no statistically significant difference in the proportion of unspecified activity by broad severity [$\chi^2(1) = 7.24$, $p = .007$].

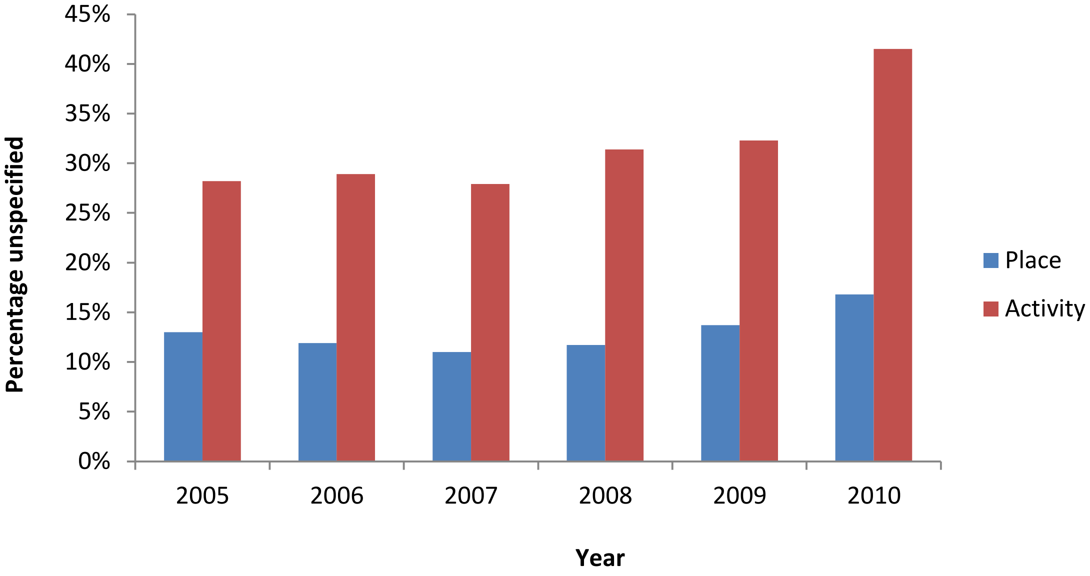


Figure 5.7: Percentage of unspecified place and activity cases by year for QISU 2005-2010

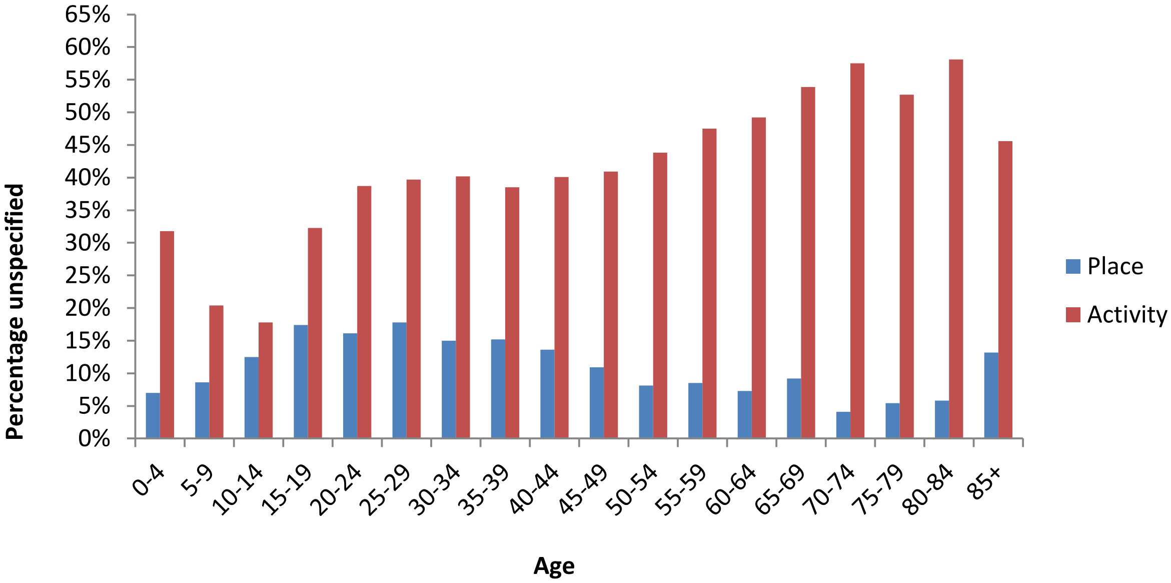


Figure 5.8: Percentage of unspecified place and activity cases by age for QISU 2005-2010

Table 5.22: Unspecified place and activity by gender, road user, ARIA+, and broad severity for QISU 2005-2010

		Place n (%)	Activity n (%)
Gender	Male	**2,509 (15.3)**	**4,324 (26.3)**
	Female	**551 (18.0)**	**3,127 (45.4)**
	Unknown	2 (16.7)	2 (16.7)
Road user	Driver	**230 (4.7)**	**2,737 (56.5)**
	Motorcyclist	**1,719 (25.1)**	**1,465 (21.4)**
	Cyclist	**808 (11.2)**	**646 (9.0)**
	Pedestrian	**88 (9.0)**	**484 (49.3)**
	Passenger	**217 (6.3)**	**2,121 (61.7)**
ARIA+	Major Cities	**992 (9.8)**	3,161 (31.3)
	Inner Regional	**884 (15.2)**	**2,341 (40.3)**
	Outer Regional	**654 (14.9)**	**946 (21.5)**
	Remote	83 (15.7)	96 (18.1)
	Very Remote	**363 (20.3)**	**585 (32.7)**
	Unknown	86 (12.2)	**324 (45.8)**
Broad severity	Fatality	2 (10.0)	12 (60.0)
	Hospitalisation	3,060 (13.1)	7,441 (31.9)

Note: Standardised residuals outside +/-3.10 are bolded.

5.4.2.5 *eARF (Queensland Ambulance Data)*

The pattern of missing, unknown, or unspecified data in eARF was examined for the final assessment and text description variables by year, age, gender, road user, and ARIA+ (see Table 5.23).

Final assessment was more likely to be unspecified for:

- 2009 and 2010 [$\chi^2(3) = 404.80$, $p < .001$, $\phi_c = .09$] (see Figure 5.9)
- Those aged 0-4 and 75+ [$\chi^2(17) = 1308.36$, $p < .001$, $\phi_c = .16$] (see Figure 5.10)
- Unknown gender [$\chi^2(2) = 1330.89$, $p < .001$, $\phi_c = .15$]
- Drivers and unspecified road users [$\chi^2(5) = 1892.84$, $p < .001$, $\phi_c = .18$]
- Outer Regional areas [$\chi^2(5) = 78.23$, $p < .001$, $\phi_c = .04$]

Final assessment was less likely to be unspecified for motorcyclists, cyclists, and pedestrians [$\chi^2(5) = 1892.84$, $p < .001$, $\phi_c = .18$].

The associated effect sizes for these differences were small, particularly for year and ARIA+.

Text description was more likely to be missing for:

- 2007 and 2008 [$\chi^2(3) = 3704.94$, $p < .001$, $\phi_c = .26$] (see Figure 5.9)
- Those aged 10-14 [$\chi^2(17) = 73.84$, $p < .001$, $\phi_c = .04$] (see Figure 5.10)
- Unknown, Remote, and Very Remote areas [$\chi^2(5) = 237.09$, $p < .001$, $\phi_c = .07$]

There was a statistically significant difference in the proportion of missing data for the *text description* variable by gender [$\chi^2(2) = 20.46$, $p < .001$, $\phi_c = .02$]. However, the effect size was very small and there were no significant standardised residuals for any cells. In addition the effect sizes associated with age and ARIA+ were also small.

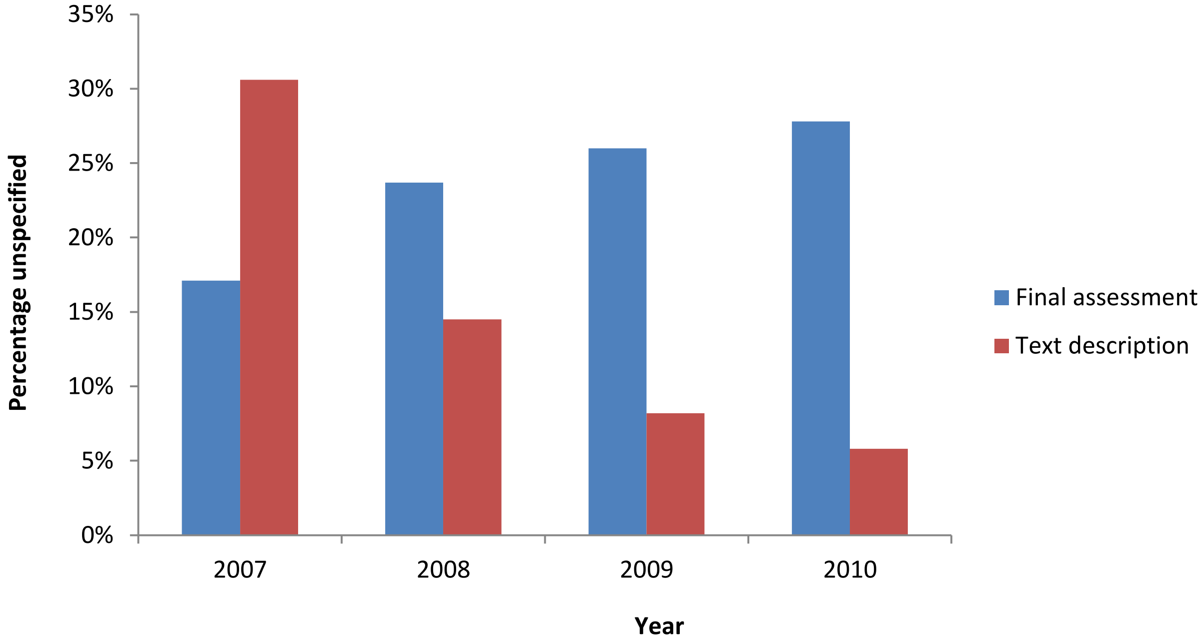


Figure 5.9: Percentage of unspecified final assessment and text description cases by year for eARF 2007-2010

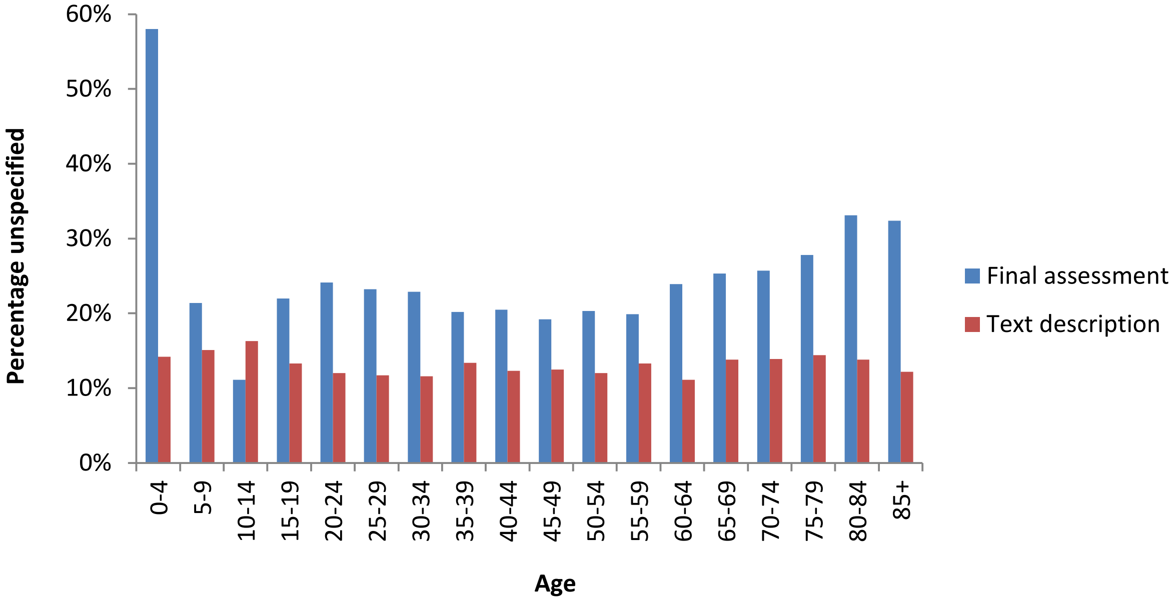


Figure 5.10: Percentage of unspecified final assessment and text description cases by age for eARF 2007- 2010

Table 5.23: Unspecified final assessment and missing text description by gender, road user type, and ARIA+ for eARF 2007-2010

		Final assessment	Text description
		n (%)	n (%)
Gender	Male	7,086 (24.0)	3,992 (13.5)
	Female	6,253 (23.5)	3,300 (12.4)
	Unknown	**459 (92.8)**	**51 (10.3)**
Road user	Driver	**2,710 (26.5)**	-
	Motorcyclist	**808 (12.9)**	-
	Cyclist	**215 (5.1)**	-
	Pedestrian	**65 (8.9)**	-
	Passenger	1,277 (22.4)	-
	Unspecified	**8,723 (29.7)**	-
ARIA+	Major Cities	6,917 (23.6)	3,594 (12.3)
	Inner Regional	3,529 (24.8)	1,798 (12.6)
	Outer Regional	**2,983 (26.7)**	1,527 (13.7)
	Remote	154 (16.4)	**190 (20.2)**
	Very Remote	169 (23.1)	**148 (20.2)**
	Unknown	46 (21.5)	**86 (40.2)**

Note: Standardised residuals outside +/-3.10 are bolded. Due to the text description being used in combination with other variables to assign road user group a comparison of missing data by road user was not performed.

5.4.2.6 *National Coronial Information System*

The consistency of missing, unknown, and unspecified was not explored for any variables as there was less than 10% missing, unknown, and unspecified cases.

5.4.3 *Validity*

As described in Section 5.3.3, the validity of each data collection was assessed by comparing the coding with a reference standard.

5.4.3.1 *Queensland Hospital Admitted Patients Data Collection*

Comparisons between the ICD-10-AM coding of a traffic injury and the place of occurrence revealed that 3,834 (11.7%) on-road vehicle cases coded as occurring on a street and/or highway were not coded as traffic. In addition, all cases coded as an off-road vehicle were not coded as traffic (specifically coded as 'unknown' status), despite the fact that 182 (7.0%) cases had place coded as 'street and highway'. Also, for on-road vehicle cases where place was 'unspecified', 12,206 (46.2%) cases were not coded as traffic (see Table 5.24).

Table 5.24: Correspondence between traffic status and place for on- and off-road vehicles for QHAPDC 2005-2010

		Street/highway	Other place	Unspecified place
On-road vehicle (V10-V82 and V87)	Traffic	34,828 88.3% of street/highway 79.4% of traffic	1,707 14.1% of other 3.9% of traffic	7,337 37.5% of unspecified 16.7% of traffic
	Not traffic[1]	3,834 11.7% of street/highway 14.5% of not traffic	10,371 85.9% of other 39.3% of not traffic	12,206 62.5% of unspecified 46.2% of not traffic
Off-road vehicle (V83-V86)	Traffic	0 0% of street/highway	0 0% of other	0 0% of unspecified
	Not traffic[1]	182 100.0% of street/highway 7.0% of not traffic	1,296 100.0% of other 50.2% of not traffic	1,105 100.0% of unspecified 42.8% of not traffic

[1] Includes non-traffic and unknown traffic status

5.4.3.2 *Queensland Injury Surveillance Unit*

Comparisons between the text descriptions and the transport coding (using external definition) of the random sample revealed that there were 952 true positives, 0 false negatives, 48 false positives, and 1000 true negatives. The results of specificity and sensitivity calculations showed that the transport coding had complete sensitivity (100.0%) and high specificity (95.4%).

In terms of the validity of the place variable for the traffic classification, the manual review revealed there were 572 true positives, 151 false negatives, 45 false positives, 233 true negatives. The sensitivity (79.1%) and specificity (83.8%) were moderately high.

For the road user classification, the comparison between this variable and the manual text review is presented in Table 5.25. The road user classification had high specificity and sensitivity for each category.

Table 5.25: Specificity and sensitivity of road user classification based on text review QISU (n = 849)

Road user classification	% specificity	% sensitivity
Driver	99.4	98.7
Motorcyclist	100.0	99.6
Cyclist	100.0	100.0
Pedestrian	100.0	98.1
Passenger	99.7	97.3

Note: 151 cases had insufficient text descriptions to determine the road user

5.4.3.3 *eARF (Queensland Ambulance Data)*

Validity of selection of transport cases

As shown in Table 5.26, for the random sample of coded transport cases (based on case nature), almost all of the coded transport cases were identified as transport in text. Cases coded as ‘pedestrian’ had a lower percentage correct compared to other cases. It should be noted that text was missing for 140 (14.0%) of the 1,000 cases.

Table 5.26: Number of coded transport cases in eARF identified as transport in text by case nature (n = 1000)

Case nature	**N**	**%**[1]	**% missing**
Motor vehicle collision	639	100.0	12.0
Motorcycle collision	107	100.0	22.5
Bicycle collision	95	99.0	17.9
Pedestrian collision	14	77.8	5.3
Total	**855**	**99.4**	**14.0**

[1] Not including missing

The manual text review of a random sample of 1000 cases not coded as transport identified 14 (1.4%) ‘fall’ cases and 7 (0.7%) ‘crush’ cases that should have been coded as transport-related.

In terms of the specificity and sensitivity of the coding of transport, sensitivity (97.6%) and specificity (99.4%) was high.

Validity of selecting road crash cases

Table 5.27 shows both the proportion of each location category that were found to be on-road from the manual text review as well as the proportion of those identified as on-road with a particular location category. Almost all of the reviewed cases coded as ‘street’, ‘public place’ or ‘vehicle’ were identified as occurring on-road in text. However, approximately 10% of on-road cases identified in text had location coded as an off-road location. Also, a majority of cases coded as ‘private residence in’ and ‘private residence out’ were identified as on-road in text.

Table 5.27: Number of coded transport cases for each location in eARF identified as on-road in text (n = 860)

Classification	**Location**	**N**	**% of on-road**	**% of location on-road**
Traffic	Street	580	75.1	98.3
	Public place	107	13.9	86.3
	Vehicle	5	0.6	100.0
Non-traffic	Private residence in	31	4.0	73.8
	Private residence out	15	1.9	60.0
	Other categories	34	4.5	49.4
	Total	**772**	**100.0**	**89.8**

Note: 140 cases had missing text descriptions

Calculations for specificity and sensitivity revealed high sensitivity (89.6%) and moderate specificity (67.5%).

Validity of road user classification

Table 5.28 shows the specificity and sensitivity of each road user classification (using a combination of case nature, vehicle type, and text terms) as identified by a manual text review. The road user classification had high specificity for each category. However, the sensitivity was only moderate for drivers, pedestrians, and passengers.

Table 5.28: Specificity and sensitivity of road user classification based on text review eARF (n = 1000)

Road user classification	% specificity	% sensitivity
Driver	95.8	66.2
Motorcyclist	96.6	100.0
Cyclist	99.7	100.0
Pedestrian	99.9	57.1
Passenger	97.5	66.5

5.4.3.4 *National Coronial Information System*

The manual document review of all 2009 cases to assess the selection of road crash cases revealed 342 true positives, 17 false negatives, 2 false positives, and 23 true negatives. The sensitivity (95.3%) and specificity were high (92.0%).

For the road user classification, the comparison between this variable and the manual document review is presented in Table 5.29. The road user classification had high specificity and sensitivity for each category.

Table 5.29: Specificity and sensitivity of road user classification based on document review NCIS (n =333)

Road user classification	% specificity	% sensitivity
Driver	98.3	98.0
Motorcyclist	98.4	99.6
Cyclist	100.0	100.0
Pedestrian	100.0	100.0
Passenger	98.1	98.5

5.4.4 *Representativeness*

As described in Section 5.3.3, the representativeness of each of the data collections was assessed by comparing the profile of cases. Table 5.30 shows the corresponding numbers for cases across each data collection. As can be seen, the prevalence of road crash injuries is completely different for each data collection. Even when cases that should correspond in terms of scope are compared (hospitalised in QRCD and EDIS cases or fatal in QRCD and NCIS cases) there are some important discrepancies.

Table 5.30: Correspondence of prevalence for each data collection for 2009

	QRCD	QHAPDC	EDIS	QISU	eARF	NCIS
Road crash	19,018	6,725	19,623	2,380	11,574	333
Fatal	331 (1.7)	71 (1.1)	19 (0.1)	3 (0.1)	-	333
Hospitalised (admitted to hospital non-fatal)	-	6,654	3,957 (20.2)	318 (13.4)	-	-
Hospitalised (taken to hospital non-fatal)	6,672 (35.1)	6,654	19,623	2,380	7,223 (62.4)	-
Attend hospital (via ambulance)	-	-	10,795 (55.0)		7,223 (62.4)	

5.4.4.1 *Queensland Hospital Admitted Patients Data Collection*

Overall, QHAPDC had 6,725 fatal and ‘hospitalised’ cases compared to 7,003 coded fatal (n = 331) and ‘hospitalised’ (n = 6,672) cases in QRCD. In terms of the profile of cases, compared to the QRCD, the QHAPDC had a statistically significantly greater proportion of males, motorcyclists, and cyclists included in the data collection. QHAPDC also had a higher proportion of younger people (14 and younger) [$\chi^2(17) = 125.69$, $p < .001$, $\phi_c =$ .10] and a lower proportion of cases in remote or very remote areas compared to QRCD (see Figure 5.11 and Table 5.31). It should be noted that the effect sizes associated with these differences were small.

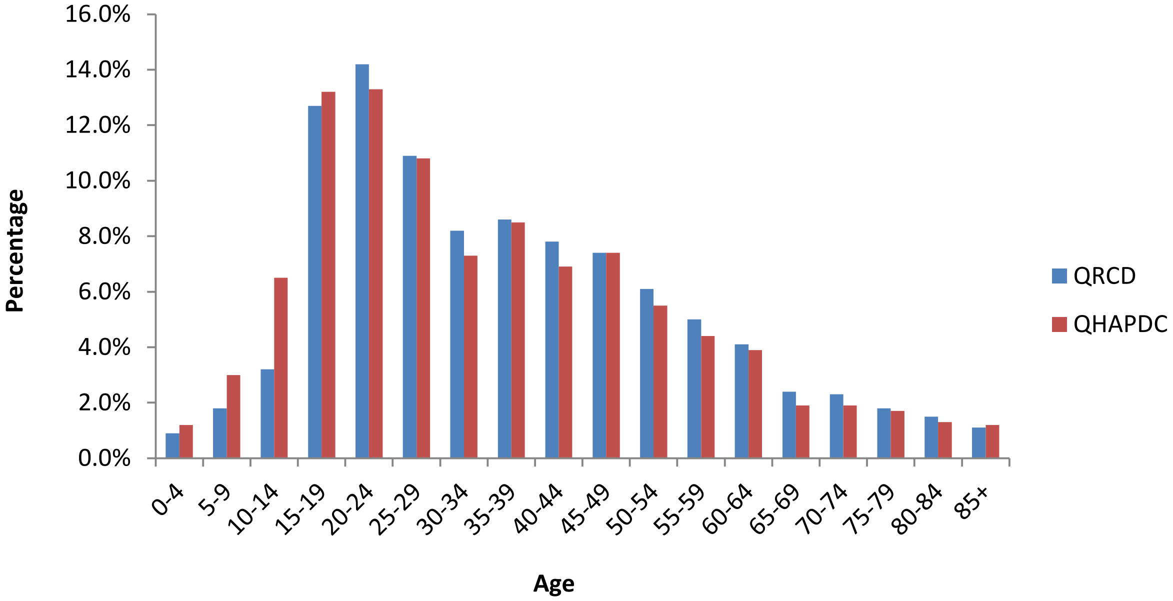


Figure 5.11: Age distribution of QRCD and QHAPDC for 2009

Table 5.31: Demographic characteristics by data source for QRCD and QHAPDC 2009

		Data source		
Variable	Level	QRCD N (%)	QHAPDC N (%)	Significance test
Gender	Male	**4,039 (57.7)**	**4,646 (69.1)**	
	Female	**2,960 (42.3)**	**2,079 (30.9)**	$\chi^2(1) = 191.06$, $p <$ .001, $\phi_c = .12$
ARIA+	Major Cities	3,611 (51.6)	3,753 (55.8)	
	Inner Regional	1,644 (23.5)	1,745 (25.9)	
	Outer Regional	1,320 (18.9)	1,063 (15.8)	
	Remote	**246 (3.5)**	**116 (1.7)**	
	Very Remote	**181 (2.6)**	**48 (0.7)**	$\chi^2(4) = 151.87$, $p <$ .001, $\phi_c = .11$
Road user	Driver	**3,723 (53.2)**	**1,904 (29.5)**	
	Motorcyclist	**1,015 (14.5)**	**2,024 (31.4)**	
	Cyclist	**362 (5.2)**	**1,067 (16.5)**	
	Pedestrian	464 (6.6)	435 (6.7)	
	Passenger	**1,439 (20.5)**	**1,021 (15.8)**	$\chi^2(4) = 162.62$, $p <$ .001, $\phi_c = .11$

Note: Standardised residuals outside +/-3.10 are bolded

In terms of broad severity, QRCD had a greater proportion of fatalities compared to QHAPDC. Based on AIS, QHAPDC had greater proportion of moderate injuries; however, there was no difference on SRR in terms of the proportion of serious vs. non-serious (see Table 5.32). However, it should be noted that much greater proportion of the QRCD were unable to be classified for either AIS or SRR compared to QHAPDC. It should also be noted that, with the exception of the differences for unspecified injuries, the effect sizes associated with the statistically significant differences were small.

Table 5.32: Severity profile by data source for QRCD and QHAPDC 2009

		Data source		
Variable	Level	QRCD N (%)	QHAPDC N (%)	Significance test
Broad severity	Fatality	**331 (4.7)**	**71 (1.1)**	
	Hospitalisation	6,672 (95.3)	6,654 (98.9)	$\chi^2(1) = 162.62$, $p < .001$, $\phi_c = .11$
Unspecified injury	Yes	**5,602 (86.5)**	**31 (0.5)**	
	No	**1,401 (19.3)**	**6,694 (99.5)**	$\chi^2(1) = 8968.61$, $p < .001$, $\phi_c = .81$
AIS	Minor	**633 (45.2)**	2,037 (34.8)	
	Moderate	**424 (30.3)**	**2,789 (47.7)**	
	Serious	**342 (24.4)**	**900 (15.4)**	
	Severe	**0 (0.0)**	89 (1.5)	
	Critical	1 (0.1)	21 (0.4)	
	Maximum	**1 (0.1)**	**16 (0.3)**	$\chi^2(5) = 190.46$, $p < .001$, $\phi_c = .16$
SRR	Serious (< 0.942)	**177 (12.7)**	**921 (13.8)**	
	Non-serious (> 0.941)	**1,218 (87.3)**	**5,733 (86.2)**	$\chi^2(1) = 1.13$, $p = .288$

Note: Standardised residuals outside +/-3.10 are bolded

The relationships between the predictor variables were explored, using Chi-square analyses, to assess any potential confounding. As shown in Table 5.33, there were a number of relationships between the factors (see Appendix H for more detail). It should be noted that age needed to be re-categorised into four groups (0-16; 17-24; 25-59; 60+) due to violations of linearity in the relationship to the outcome when treated as ordinal (5 year intervals). Referent categories for the predictors in the logistic regression were chosen based on either the absence of a condition (e.g., non-serious) or the group with the largest proportion of injuries (e.g., Major Cities, drivers, 25-59 age group).

Table 5.33: Relationships between each of the factors for QRCD and QHAPDC

	Gender	Age	ARIA+	Road user	Broad severity	Serious Injury
Gender		✓	✗	✓	✓	✗
Age	✓		✓	✓	✗	✓
ARIA+	✗	✓		✓	✓	✓
Road user	✓	✓	✓		✓	✓
Broad severity	✓	✗	✓	✓		✓
Serious injury	✗	✓	✓	✓	✓	

Note: The ticks represent statistically significant relationships and crosses represent no statistically significant relationship

In order to adjust for the relationships between the predictors, a logistic regression was performed. With all variables in the model, the model was statistically significant, $\chi^2(14) = 1227.28$, $p < .001$ (Nagelkerke $R^2 = .24$). After controlling for the relationships between the predictors, age and gender were no longer significant. In contrast, road user, broad severity, and serious injury remained statistically significant. Specifically, motorcyclists and cyclists had greater odds (3.7 and 2.9 times respectively) of being included in QHAPDC compared to drivers. Also, serious cases had greater odds (1.7 times) of being included in QHAPDC compared to non-serious cases; and fatal cases had lower odds (33 times) of being included in QHAPDC compared to hospitalisation (see Table 5.34).

Table 5.34: Logistic regression analysis of the profile of road crash injuries in QHAPDC compared to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.64	1.17	0.93 – 1.48	.026
Age	0 – 16	1.93	1.52	0.98 – 2.35	.002
	17 – 24	1.10	1.02	0.73 – 1.41	.880
	25 – 59	1.00	1.00	Referent	
	60 +	0.94	1.16	0.85 – 1.58	.113
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	3.90	3.71	2.58 – 5.14	< .001
	Cyclist	5.76	2.86	1.89 – 4.33	< .001
	Pedestrian	1.83	1.15	0.75 – 1.75	.289
	Passenger	1.39	1.06	0.79 – 1.42	.532
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.02	0.97	0.74 – 1.27	.684
	Outer Regional	0.78	0.77	0.58 – 1.04	.005
	Remote	0.45	0.34	0.18 - 0.64	< .001
	Very Remote	0.26	0.12	0.05 – 0.27	< .001
Broad Severity	Hospitalisation	1.00	1.00	Referent	
	Fatality	0.22	0.03	0.02 – 0.05	< .001
Serious	Non-serious	1.00	1.00	Referent	
	Serious	1.10	1.71	1.19 – 2.44	< .001

[1] Adjusted for all variables in the equation

5.4.4.2 *Emergency Department Information System*

Overall, EDIS had 19,623 road crash cases compared to 7,003 cases in QRCD categorised as fatal or ‘hospitalised’ (taken to hospital). In terms of the profile of cases, compared to the QRCD, EDIS had a statistically significantly greater proportion of males, motorcyclists, and cyclists included in the data collection. EDIS also had a higher proportion of younger people (19 and younger) [$\chi^2(17) = 442.22$, $p < .001$, $\phi_c = .13$] and a lower proportion of cases in outer regional or remote areas compared to QRCD (see Figure 5.12 and Table 5.35). It should be noted that, with the exception of road user type, the effect sizes associated with these differences were small.

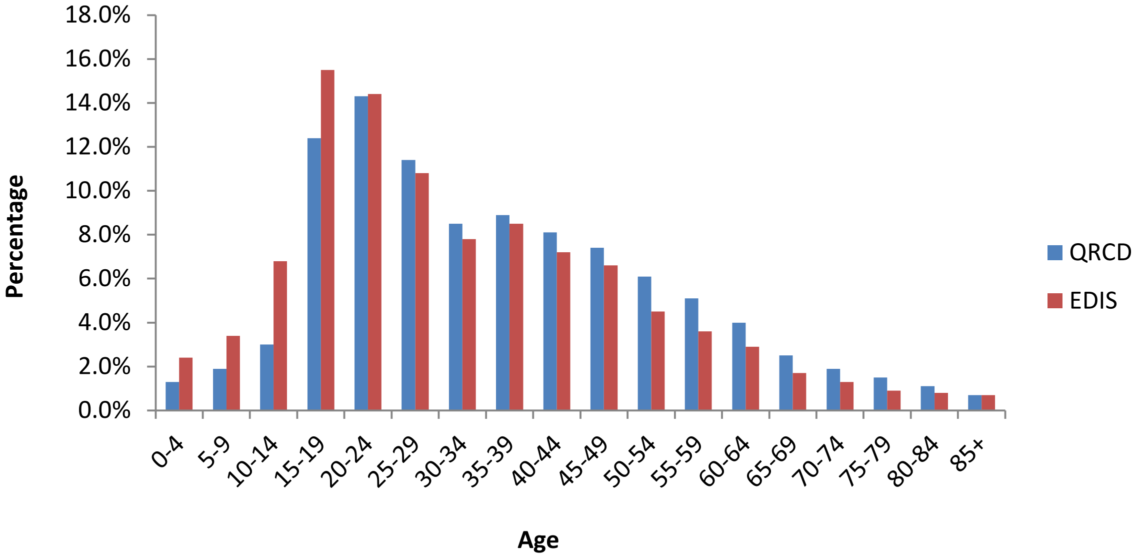


Figure 5.12: Age distribution of QRCD and EDIS 2009

Table 5.35: Demographic characteristics by data source for QRCD and EDIS 2009

Variable	Level	Data source		Significance test
		QRCD N (%)	EDIS N (%)	
Gender	Male	**4,039 (57.7)**	**12,224 (62.3)**	
	Female	**2,960 (42.3)**	**7,395 (37.7)**	$\chi^2(1) = 45.90$, $p < .001$, $\phi_c = .04$
ARIA+	Major Cities	**3,611 (51.6)**	**10,046 (54.0)**	
	Inner Regional	**1,644 (23.5)**	**5,455 (29.3)**	
	Outer Regional	**1,320 (18.9)**	**2,655 (14.3)**	
	Remote	**246 (3.5)**	**89 (0.5)**	
	Very Remote	181 (2.6)	355 (1.9)	$\chi^2(4) = 506.26$, $p < .001$, $\phi_c = .14$
Road user	Driver	**3,723 (53.2)**	**2,437 (22.7)**	
	Motorcyclist	**1,015 (14.5)**	**3,707 (34.6)**	
	Cyclist	**362 (5.2)**	**2,525 (23.5)**	
	Pedestrian	**464 (6.6)**	**177 (1.7)**	
	Passenger	**1,439 (20.5)**	**1,876 (17.5)**	$\chi^2(4) = 2959.84$, $p < .001$, $\phi_c = .41$

Note: A large proportion of cases (45.4%) were not able to be classified into a road user group in EDIS. Standardised residuals outside +/-3.10 are bolded

In terms of broad severity, QRCD had a greater proportion of fatalities compared to EDIS. Based on AIS, EDIS had greater proportion of moderate injuries. Also, QRCD had a greater proportion of serious compared to EDIS (see Table 5.36). However, it should be noted that the effect sizes associated with broad severity and AIS were small and a much greater proportion of the QRCD were unable to be classified for either AIS or SRR compared to EDIS.

Table 5.36: Severity profile by data source for QRCD and EDIS 2009

		Data source		
Variable	Level	QRCD N (%)	EDIS N (%)	Significance test
Broad severity	Fatality	331 (4.7)	19 (0.1)	
	Hospitalisation	6,672 (95.3)	19,604 (99.9)	$\chi^2(1) = 852.78$, $p < .001$, $\phi_c = .18$
Unspecified injury	Yes	**5,602 (80.0)**	**52 (0.3)**	
	No	**1,401 (20.0)**	**19,571 (99.7)**	$\chi^2(1) = 19615.33$, $p < .001$, $\phi_c = .86$
AIS	Minor	**633 (45.2)**	13,539 (75.0)	
	Moderate	**424 (30.3)**	**3,926 (21.8)**	
	Serious	**342 (24.4)**	**523 (2.9)**	
	Severe	**0 (0.0)**	49 (0.3)	
	Critical	1 (0.1)	0 (0.0)	
	Maximum	**1 (0.1)**	**5 (0.1)**	$\chi^2(5) = 1570.86$, $p < .001$, $\phi_c = .28$
SRR	Serious (< 0.942)	**177 (12.7)**	**981 (5.2)**	
	Non-serious (> 0.941)	**1,218 (87.3)**	**18,005 (94.8)**	$\chi^2(1) = 137.18$, $p < .001$, $\phi_c = .08$

Note: Standardised residuals outside +/-3.10 are bolded

The relationships between IVs were explored to assess any potential confounding. There were significant relationships between all the IVs (see Appendix H for more detail).

In order to adjust for the relationships between the IVs, a logistic regression was performed. With all variables[7] in the model, the model was statistically significant, $\chi^2(13) = 1233.57$, $p < .001$ (Nagelkerke $R^2 = .20$). After controlling for the relationships between the predictors, all variables remained statistically significant. Specifically, motorcyclists and cyclists had greater odds (5.6 and 6.3 respectively) and pedestrians had 3.1 times lower odds of being included in EDIS compared to drivers. Also, those cases from Outer Regional, Remote, and Very Remote areas had lower odds (1.7, 12.5, and 3 times respectively) of being included in EDIS compared to those from Major Cities. Those aged 0-16 and 17-24 had greater odds (2.1 and 1.6 times respectively) of being included in

[7] Broad severity was excluded from the analysis as there were too few fatality case in EDIS for interpretation to be meaningful

EDIS compared to those aged 25-59. Males had 1.3 times greater odds of being included in EDIS compared to females. Finally, serious cases had 2.9 times lower odds of being included in EDIS compared to non-serious cases (see Table 5.37).

Table 5.37: Logistic regression analysis of the profile of road crash injuries in EDIS compared to QRCD 2009

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.21	1.27	1.02 – 1.58	< .001
Age	0 – 16	2.37	1.94	1.28 – 2.93	< .001
	17 – 24	1.22	1.48	1.17 – 1.88	< .001
	25 – 59	1.00	1.00	Referent	
	60 +	0.70	0.92	0.66 – 1.29	.410
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	5.58	5.55	4.10 – 7.53	< .001
	Cyclist	10.66	6.27	4.22 – 9.34	< .001
	Pedestrian	0.58	0.32	0.21 – 0.50	< .001
	Passenger	1.99	1.30	1.00 – 1.70	.001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.19	0.99	0.77 – 1.26	.881
	Outer Regional	0.72	0.58	0.44 – 0.76	< .001
	Remote	0.13	0.08	0.04 – 0.17	< .001
	Very Remote	0.71	0.33	0.18 – 0.58	< .001
Serious	Non-serious	1.00	1.00	Referent	
	Serious	0.38	0.34	0.24 – 0.49	< .001

[1] Adjusted for all variables in the equation

5.4.4.3 *Queensland Injury Surveillance Unit*

Overall, QISU had 2,380 road crash cases compared to 7,003 hospital and fatal cases in QRCD. In terms of the profile of cases, compared to the QRCD, QISU had a statistically significantly greater proportion of males, motorcyclists, and cyclists included in the data collection. QISU also had a higher proportion of younger people (19 and younger) [$\chi^2(17) = 796.57$, $p < .001$, $\phi_c = .29$] and a greater proportion of cases in very remote areas compared to QRCD (see Figure 5.13 and Table 5.38). It should be noted that, with the exception of road user type, the effect sizes associated with these differences were small.

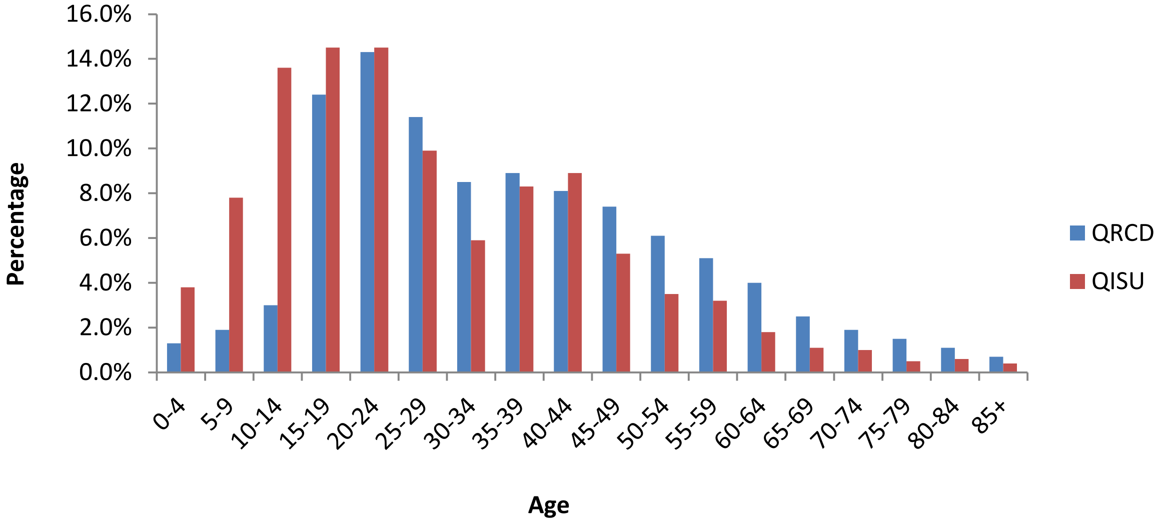


Figure 5.13: Age distribution of QRCD and QISU 2009

Table 5.38: Demographic characteristics by data source for QRCD and QISU 2009

Variable	Level	Data source QRCD N (%)	 QISU N (%)	Significance test
Gender	Male	**4,039 (57.7)**	**1,489 (62.6)**	
	Female	**2,960 (42.3)**	**890 (37.4)**	$\chi^2(1) = 17.48$, $p <$ .001, $\phi_c = .04$
ARIA+	Major Cities	3,611 (51.6)	1,147 (50.2)	
	Inner Regional	1,644 (23.5)	569 (24.9)	
	Outer Regional	1,320 (18.9)	380 (16.6)	
	Remote	246 (3.5)	**30 (1.3)**	
	Very Remote	**181 (2.6)**	**157 (6.9)**	$\chi^2(4) = 121.81$, $p <$ .001, $\phi_c = .12$
Road user	Driver	**3,723 (53.2)**	**840 (35.3)**	
	Motorcyclist	1,015 (14.5)	**435 (18.3)**	
	Cyclist	**362 (5.2)**	**479 (20.1)**	
	Pedestrian	464 (6.6)	116 (4.9)	
	Passenger	1,439 (20.5)	510 (21.4)	$\chi^2(4) = 585.91$, $p <$ .001, $\phi_c = .25$

Note: Standardised residuals outside +/-3.10 are bolded

In terms of broad severity, QRCD had a greater proportion of fatalities compared to QISU. Based on AIS, QRCD had greater proportion of moderate injuries compared to QISU (although this could not be tested for significance due to a violation of the assumption relating to expected cell counts). QRCD also had a higher proportion of cases classified as serious compared to QISU (see Table 5.39). However, it should be noted that the effect size was small and a much greater proportion of the QRCD were unable to be classified for AIS and SRR compared to QISU.

Table 5.39: Severity profile by data source for QRCD and QISU 2009

Variable	Level	Data source QRCD N (%)	 QISU N (%)	Significance test
Broad severity	Fatality	331 (4.7)	**3 (0.1)**	
	Hospitalisation	6,672 (95.3)	2,377 (13.4)	$\chi^2(1) = 109.51$, $p < .001$, $\phi_c = .11$
Unspecified injury	Yes	**5,608 (80.1)**	**368 (15.5)**	
	No	**1,395 (19.9)**	**2,012 (84.5)**	$\chi^2(1) = 3206.19$, $p < .001$, $\phi_c = .44$
AIS	Minor	633 (45.2)	1,513 (75.0)	
	Moderate	424 (30.3)	427 (21.2)	
	Serious	342 (24.4)	64 (3.2)	
	Severe	0 (0.0)	1 (0.1)	
	Critical	1 (0.1)	0 (0.0)	
	Maximum	1 (0.1)	11 (0.5)	-[1]
SRR	Serious (< 0.942)	**177 (12.7)**	**138 (6.0)**	
	Non-serious (> 0.941)	**1,218 (87.3)**	**2,174 (94.0)**	$\chi^2(1) = 50.52$, $p < .001$, $\phi_c = .12$

[1] Chi-square not reported as the assumption of expected cell sizes was violated
Note: Standardised residuals outside +/-3.10 are bolded

The relationships between IVs were explored to assess any potential confounding. There were significant relationships between all the IVs, except age and gender were not related to each other (see Appendix H for more detail).

In order to adjust for the relationships between the IVs, a logistic regression was performed. With all variables[8] in the model, the model was statistically significant, $\chi^2(13) = 509.22$, $p < .001$ (Nagelkerke $R^2 = .18$). After controlling for the relationships between the predictors, all variables (with the exception of gender) remained statistically significant. Specifically, motorcyclists and cyclists had greater odds (1.6 and 2.1 times respectively) and passenger and pedestrians had lower odds (1.6 and 2.5 times respectively) of being included in QISU compared to drivers. It is interesting to note that in bivariate analysis, passengers had greater odds of being included in QISU, but after controlling for the other factors, this relationship was reversed. Also, those cases from Very Remote areas had 1.9 times greater odds and Remote areas had 3.2 times lower odds of being included in QISU compared to those from Major Cities. Those aged 0-16 and 17-24 had greater odds (6.0 and 1.5 times respectively) and those aged 60 and over had 1.6 times lower odds of being included in QISU compared to those aged 25-59. Finally,

[8] Broad severity was excluded from the analysis as there were too few fatality case in QISU for interpretation to be meaningful

serious cases had 2.2 times lower odds of being included in QISU compared to non-serious cases (see Table 5.40).

Table 5.40: Logistic regression analysis of the profile of road crash injuries in QISU compared to QRCD 2009

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.23	1.02	0.83 – 1.25	.107
Age	0 – 16	5.74	5.97	3.78 – 9.42	< .001
	17 – 24	1.44	1.54	1.16 – 2.04	< .001
	25 – 59	1.00	1.00	Referent	
	60 +	0.54	0.62	0.40 – 0.97	< .001
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	1.90	1.56	1.09 – 2.22	< .001
	Cyclist	5.87	2.07	1.33 – 3.24	< .001
	Pedestrian	1.11	0.41	0.24 – 0.69	< .001
	Passenger	1.57	0.65	0.46 – 0.91	< .001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.09	1.07	0.79 – 1.44	.454
	Outer Regional	0.91	0.84	0.60 – 1.17	.085
	Remote	0.38	0.32	0.15 – 0.69	< .001
	Very Remote	2.73	1.90	1.06 – 3.38	< .001
Serious	Non-serious	1.00	1.00	Referent	
	Serious	0.44	0.45	0.30 – 0.70	< .001

[1] Adjusted for all variables in the equation

5.4.4.4 *eARF (Queensland Ambulance Data)*

Overall, eARF had 11,574 road crash cases compared to 19,018 cases in QRCD. In terms of the profile of cases, compared to the QRCD, eARF had a statistically significantly greater proportion of females, motorcyclists, passengers, and cyclists included in the data collection. eARF also had a higher proportion of younger people (4 and younger) [$\chi^2(17) = 213.10$, $p < .001$, $\phi_c = .09$] and a lower proportion of cases in major cities, remote or very remote areas compared to QRCD (see Figure 5.14 and Table 5.41). It should be noted that the effect sizes associated with these differences were small.

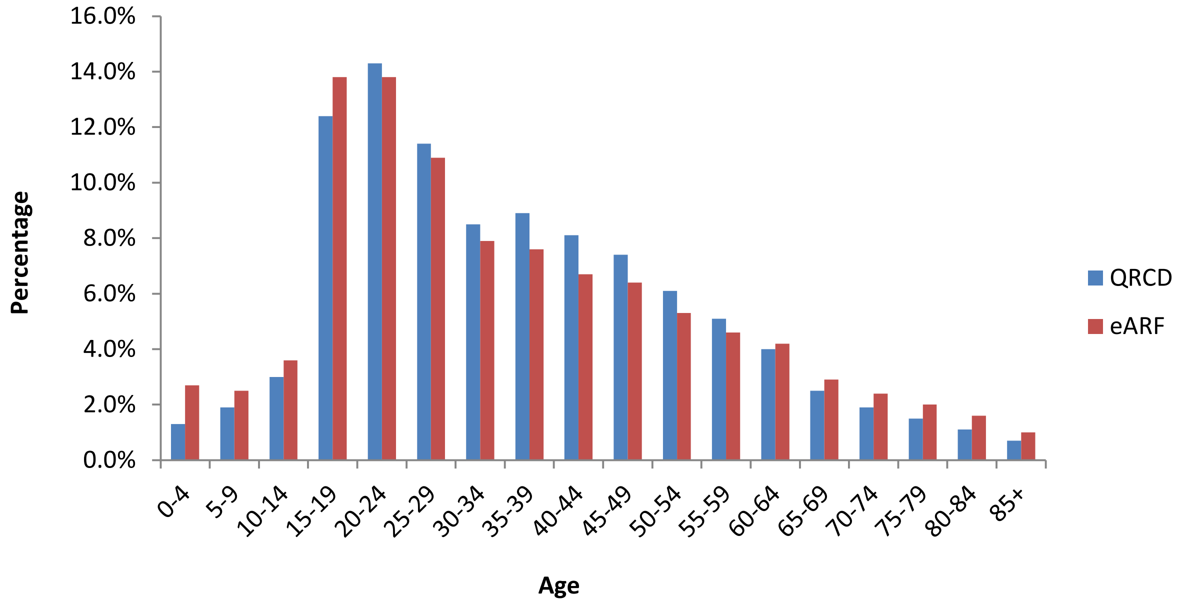


Figure 5.14: Age distribution of QRCD and eARF for 2009

Table 5.41: Demographic characteristics by data source for QRCD and eARF 2009

		Data source		
Variable	Level	QRCD N (%)	eARF N (%)	Significance test
Gender	Male	**9,988 (52.8)**	**5,479 (47.7)**	
	Female	**8,934 (47.2)**	**6,015 (52.3)**	$\chi^2(1) = 74.91$, $p <$ .001, $\phi_c = .05$
ARIA+	Major Cities	**11,022 (58.0)**	**5,735 (49.6)**	
	Inner Regional	**4,041 (21.3)**	**3,213 (27.8)**	
	Outer Regional	**3,135 (16.5)**	**2,354 (20.4)**	
	Remote	**514 (2.7)**	**143 (1.2)**	
	Very Remote	300 (1.6)	**107 (0.9)**	$\chi^2(4) = 376.36$, $p <$ .001, $\phi_c = .11$
Road user	Driver	**11,131 (58.5)**	**2,548 (50.4)**	
	Motorcyclist	1,819 (9.6)	**648 (12.8)**	
	Cyclist	869 (4.6)	**323 (6.4)**	
	Pedestrian	**839 (4.4)**	**62 (1.2)**	
	Passenger	**4,360 (22.9)**	**1,478 (29.2)**	$\chi^2(4) = 288.11$, $p <$ .001, $\phi_c = .11$

Note: A large proportion of cases (56.3%) were not able to be classified into a road user group in eARF. Standardised residuals outside +/-3.10 are bolded

The relationships between IVs were explored to assess any potential confounding. There were significant relationships between all the IVs (see Appendix H for more detail).

In order to adjust for the relationships between the IVs, a logistic regression was performed. With all variables in the model, the model was statistically significant, $\chi^2(12) = 658.23$, $p < .001$ (Nagelkerke $R^2 = .04$). After controlling for the relationships between the predictors, all variables remained statistically significant. Specifically, motorcyclists, cyclists, and passengers had greater odds (1.6, 1.7, and 1.4 times respectively), and pedestrians had 3.4 times lower odds of being included in eARF compared to drivers. Males had 1.3 times the odds of being in eARF compared to females. Those aged 0-16 and 60 and over had greater odds (1.6 and 1.3 times respectively) of being included in eARF compared to those aged 25-59. Finally, those cases from Inner and Outer Regional areas had greater odds (1.5 and 1.4 respectively) and those from Remote areas had 2 times lower odds of being included in eARF compared to those from Major Cities (see Table 5.42).

Table 5.42: Logistic regression analysis of the profile of road crash injuries in eARF compared to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.23	1.32	1.18 – 1.47	< .001
Age	0 – 16	1.60	1.55	1.25 – 1.92	< .001
	17 – 24	1.16	1.14	0.99 – 1.29	.002
	25 – 59	1.00	1.00	Referent	
	60 +	1.36	1.31	1.11 – 1.55	< .001
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	1.56	1.66	1.39 – 1.98	< .001
	Cyclist	1.62	1.70	1.34 – 2.16	< .001
	Pedestrian	0.32	0.30	0.19 – 0.47	< .001
	Passenger	1.48	1.37	1.20 – 1.57	< .001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.53	1.52	1.34 – 1.73	< .001
	Outer Regional	1.44	1.39	1.21 – 1.61	< .001
	Remote	0.54	0.51	0.32 – 0.82	< .001
	Very Remote	0.69	0.62	0.36 – 1.09	.005

[1] Adjusted for all variables in the equation

5.4.4.5 *National Coronial Information System*

Overall, NCIS had 333 road crash cases compared to 331 fatal cases in QRCD. There were no statistically significant differences between the NCIS and QRCD in terms of age [$\chi^2(18) = 3.42$, $p = .998$], gender, road user, or ARIA+ (see Figure 5.15 and Table 5.43).

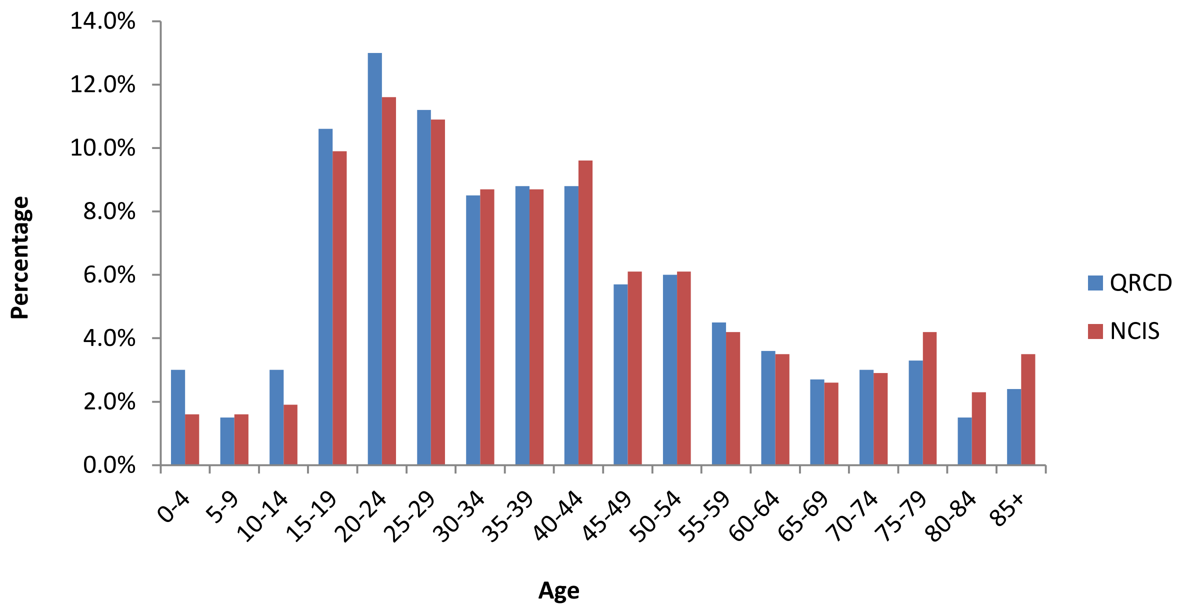


Figure 5.15: Age distribution of QRCD and NCIS 2009

Table 5.43: Demographic characteristics by data source for QRCD and NCIS 2009

Variable	Level	Data source QRCD N (%)	 NCIS N (%)	Significance test
Gender	Male	240 (72.5)	251 (75.4)	
	Female	90 (27.2)	82 (24.6)	$\chi^2(1) = 0.61$, $p =$ .437
ARIA+	Major Cities	96 (29.0)	108 (33.0)	
	Inner Regional	105 (31.7)	106 (32.4)	
	Outer Regional	89 (26.9)	86 (26.3)	
	Remote	29 (8.8)	17 (5.2)	
	Very Remote	12 (3.6)	10 (3.1)	$\chi^2(4) = 4.05$, $p =$ .399
Road user	Driver	152 (45.9)	153 (46.9)	
	Motorcyclist	60 (18.1)	63 (19.3)	
	Cyclist	8 (2.4)	8 (2.5)	
	Pedestrian	40 (12.1)	36 (11.0)	
	Passenger	71 (21.5)	66 (20.2)	$\chi^2(4) = 0.43$, $p =$ .980

5.4.5 *Definitions of serious injury*

5.4.5.1 *Queensland Road Crash Database*

Table 5.44 shows the proportion of serious injuries based on Broad Severity, AIS, and SRR classification criteria. There was a much larger proportion of serious injuries classified when using the broad severity criteria compared to both AIS and SRR. While the SRR and AIS proportions are quite similar, interestingly, only 40 cases were coded as serious under both the AIS and SRR criteria.

Table 5.44: The number and proportion of serious and non-serious injuries based on the three different severity measure criteria, QRCD 2009

	Broad severity (Fatal and 'hospitalised')	AIS (score of 3 or above)	SRR (0.941 or less)
Serious	7,003 (36.8%)	355 (8.6%)	387 (9.3%)
Non-serious	12,015 (63.2%)	3,788 (91.4%)	3,762 (90.7%)

To further explore the broad severity classification, the median of SRRs was calculated for each broad severity category. Table 5.45 shows that the median SRR was lowest (more severe) for fatalities. Surprisingly, the median SRR for other injury was lower than that of hospitalisations, suggesting that other injuries (medical treatment and minor injuries) are more severe than those cases taken to hospital. This table also shows that the range of severities (as measured by SRR) was quite wide within each broad severity category.

Table 5.45: Median and range SRR for each broad severity category, QRCD 2009

	Median SRR	Range (min – max)
Fatality	0.940	0.746 – 1.000
Hospitalisation	0.985	0.500 – 1.000
Other injury	0.954	0.554 – 1.000

5.4.5.2 *Queensland Hospital Admitted Patients Data Collection*

Table 5.46 shows the proportion of serious injuries based on Broad Severity, AIS, and SRR classification criteria. Due to the nature of the data collection (all cases 'hospitalised' or fatality), based on broad severity, all cases are classified as serious. The proportion of serious cases based on AIS was higher than the proportion of serious cases based on SRR. There were 488 cases coded as serious under both the AIS and SRR criteria.

Table 5.46: The number and proportion of serious and non-serious injuries based on the three different severity measure criteria, QHAPDC 2009

	Broad severity (Fatal and 'hospitalised')	AIS (score of 3 or above)	SRR (0.941 or less)
Serious	6,725 (100.0%)	1,026 (17.5%)	921 (13.8%)
Non-serious	0 (0.0%)	4,826 (82.5%)	5,773 (86.2%)

To further explore the broad severity classification, the median of SRRs were calculated for each broad severity category available. Table 5.47 shows that the median SRR was lower (more severe) for fatalities compared to 'hospitalised' cases. The range of severities (as measured by SRR) was quite wide for both fatalities and 'hospitalised'.

Table 5.47: Median and range SRR for each broad severity category, QHAPDC 2009

	Median SRR	Range (min – max)
Fatality	0.867	0.306 – 0.996
'Hospitalised'	0.991	0.306 – 1.000

5.4.5.3 *Emergency Department Information System*

Table 5.48 shows the proportion of serious injuries based on Broad Severity, AIS, and SRR classification criteria. Due to the nature of the data collection (all cases taken to hospital or fatality), based on broad severity, all cases are classified as serious. The proportion of serious cases based on AIS was higher than the proportion of serious cases based on SRR. There were 257 cases coded as serious under both AIS and SRR criteria.

Table 5.48: The number and proportion of serious and non-serious injuries based on the three different severity measure criteria, EDIS 2009

	Broad severity (Fatal and 'hospitalised')	AIS (score of 3 or above)	SRR (0.941 or less)
Serious	19,623 (100.0%)	577 (3.2%)	981 (5.2%)
Non-serious	0 (0.0%)	17,465 (96.8%)	18,005 (94.8%)

To further explore the broad severity classification, the median of SRRs were calculated for each broad severity category available. Table 5.49 shows that the median SRR was lower (more severe) for fatalities compared to 'hospitalised' cases.

Table 5.49: Median SRR for each broad severity category, EDIS 2009

	Median SRR	Range (min – max)
Fatality	0.889	0.735 – 0.988
'Hospitalised'	0.993	0.667 – 1.000

5.4.5.4 *Queensland Injury Surveillance Unit*

Table 5.50 shows the proportion of serious injuries based on Broad Severity, AIS, and SRR classification criteria. Due to the nature of the data collection (all cases taken to hospital or fatality), based on broad severity, all cases are classified as serious. The proportion of serious case based on SRR was higher than the proportion of serious based on AIS. There were only 17 cases coded as serious under both AIS and SRR criteria.

Table 5.50: The number and proportion of serious and non-serious injuries based on the three different severity measure criteria, QISU 2009

	Broad severity (Fatal and 'hospitalised')	AIS (score of 3 or above)	SRR (0.941 or less)
Serious	2,380 (100.0%)	76 (3.8%)	138 (6.0%)
Non-serious	0 (0.0%)	1,940 (96.2%)	2,174 (94.0%)

To further explore the broad severity classification, the median of SRRs were calculated for each broad severity category available. Table 5.51 shows that the median SRR was lower (more severe) for fatalities compared to 'hospitalised' cases.

Table 5.51: Median SRR for each broad severity category, QISU 2009

	Median SRR	Range (min – max)
Fatality	0.917	0.884 – 0.983
'Hospitalised'	0.993	0.775 – 1.000

5.5 Discussion

5.5.1 *Summary of results*

A summary of results for completeness, consistency, validity, representativeness, and severity measurement is provided below for each data collection. There are a number of statistically significant results, however due to the large sample size, some of these results were not considered meaningful. Rather, attention will be given to those cases in which the effect size associated with a result was above 0.1 (more than a small effect).

5.5.1.1 *Queensland Road Crash Database*

There was minimal missing, unspecified, or unknown data in terms of the Core Minimum, Optional, or Supplemental data set variables. One exception was for the injury description variable used to classify severity in terms of AIS and SRR. Almost three-quarters of the cases had missing or unspecified information. This result could be an indication of a reluctance of police to speculate on injury due to it being outside of their expertise.

In terms of consistency, QRCD had bias in the amount of missing and unspecified injury descriptions in terms of broad severity. Specifically, it was found that the injury description was less likely to have complete information when the case was ‘hospitalised’. It is possible that police may be less likely to complete the injury description field in cases where other parties (e.g., ambulance officers or hospital staff) are involved (as would be the case with a ‘hospitalised’ case), as the police officer would defer to medical staff expertise and think they would better capture that information in other data sources. It is also possible that in cases where the injured person is taken to hospital, that the police officer may not have the opportunity to assess the injury due the person being treated at the time or having already left the scene by the time the officer arrives.

The validity of the coding and classification in QRCD was not able to be directly assessed in this study. However, the representativeness was able to be explored. There were differences in the prevalence of road crash injuries between QRCD and the other data collections. There was some indication based on these discrepancies and the profiling differences between this data collection and other data collections (discussed further in the subsequent sections), of possible under-reporting in the QRCD. However, it is also possible that some of the differences are due to scope differences and/or misclassification of road crash injuries in the other data collections. It is not possible, without data linkage, to quantify the extent of misclassification versus under-reporting. This issue will be further explored in the next study, using linked data.

The final issue with the QRCD relates to severity, particularly with the classification of serious injuries. The AIS, SRR, and broad severity classification of serious injury do not correspond. Specifically, using broad severity, the proportion of serious injuries was much greater than when using AIS or SRR. It should also be noted that for the ‘hospitalised’ category, there was a broad range of injury types and SRRs. Also, the category of ‘other injuries’ actually had a lower median SRR (more severe) than the ‘hospitalised’ category suggesting issues with regards to the police assignment of injury types.

5.5.1.2 *Queensland Hospital Admitted Patients Data Collection*

The inclusion of ICD external cause coding in this data set allowed for identification of cases with reasonable ease. However, there were some issues in terms of completeness. There were over 10 percent of cases with an ‘unknown/unspecified’ *traffic status*, making it impossible for these cases to be included as road crash injury cases. It is not known how many of these cases could potentially be road crash cases, and while 10 percent may not be a large proportion, it represents over 10,000 cases. Therefore, estimates of prevalence for this data collection may not only be inaccurate, the potential additional cases could be substantial. Another variable that potentially could be used to identify road crash cases was the variable *place*. However, this variable had approximately one-third ‘unknown/unspecified’ cases. Therefore, using this variable to determine road crash injuries would potentially be less reliable than using *traffic-status*. There were also substantial ‘unknown/unspecified’ cases for activity (approximately three-quarters).

When assessing the consistency of QHAPDC in terms of ‘unspecified/unknown’ cases, it was found that there were some significant inconsistencies. For *traffic*, males were more likely to be recorded as ‘unspecified’ or ‘unknown’ while this was less likely for those cases occurring in Major Cities. For *place*, young people (aged 0-14), motorcyclists, and cyclists were more likely to be ‘unspecified/unknown’. *Activity* was more likely to be ‘unspecified/unknown’ for drivers, passengers, and pedestrians. While it is not clear the underlying reason for these inconsistencies, it is important to note their impact on the conclusions drawn when using these data. The inconsistencies could introduce a bias in terms of the selection of cases.

There was no way to directly assess the validity of the coded variables in QHAPDC as there was no reference standard with which to compare them (such as a text field). However, there were some inconsistencies between variables that are supposed to measure similar things. For example, when comparing the traffic classification with the place variable, there were some discrepancies in terms of street/highway cases not being coded as traffic and vice versa. This discrepancy may have implications for the validity of these variables, which could have impact on identifying the prevalence of road crash cases. However, it is unclear as to which variable is incorrect or even whether both are incorrect. Without an appropriate reference standard it is not possible to determine at this stage. However, it may be possible to explore this issue further in the next study using the links with other data collections to provide a reference.

In terms of overall numbers, the difference between QRCD and QHAPDC was minimal, with QRCD having slightly more cases than QHAPDC. Due to the scope of the data collections, it would be expected that QRCD would have more cases than QHAPDC as it includes all those cases taken to hospital, while QHAPDC has only admitted to hospital cases. When the profiles were compared at a bivariate level, there were significant differences between QRCD and QHAPDC. Specifically, QHAPDC had a greater proportion of males, younger people (aged 0-14), motorcyclists, and cyclists compared to QRCD. These differences provide some evidence of under-reporting within the QRCD, because as noted above it would expected that QRCD should have more cases as the inclusion criteria is broader than QHAPDC. A regression analysis was performed in order to take into account the relationships between the independent variables. Results of this showed that the variables remained significant after controlling for each other with the exception of gender. It is possible that some of the differences found were not due to under-reporting, but instead due to misclassification of road crash injuries in QHAPDC. It is not clear at this stage how valid QHAPDC coding is in terms of identifying road crash cases and road users. This data collection’s primary purpose is not for this type of classification, so it is possible that the accuracy of the coding could be compromised. It is also possible that the classification of ‘hospitalised’ in QRCD is also incorrect, due to the way it is collected. It is possible that some cases not coded as ‘taken hospital’ involved an injury in which the person attended hospital without the police knowing and were ultimately admitted. Further research, using data linkage, may quantify the extent of misclassification versus under-reporting.

In addition to the above differences, QHAPDC had a lower proportion of Remote and Very Remote cases based on *ARIA+* compared to QRCD. This result is perhaps not surprising considering the classification basis for each collection. QHAPDC ARIA+ relates to the location of the hospital, whereas QRCD *ARIA+* relates to the location of the crash. It is likely that even when a crash occurs in a Remote or Very Remote location, the injured person would not be necessarily be treated in a hospital in a Remote or Very Remote location due to the lack of facilities in these locations. This difference would bias this measure somewhat.

For severity, there was no difference between the collections in terms of the proportion classified as serious based on *Survival Risk Ratio (SRR)*. However, QRCD had a greater proportion of fatalities and serious or worse *AIS* classification compared to QHAPDC. The difference between the collections in terms of fatalities is not surprising as there would be a considerable number of fatalities that are not admitted to hospital (i.e., died at scene, died in transit, and died on arrival). Generally, the differences in severity between QRCD and QHAPDC should be treated with caution. QRCD had a considerably greater proportion (87% vs. 0.5%) of missing/unspecified injury descriptions which were used to determine *AIS* and *SRR*. There was also a potential bias in the amount of missing and unspecified injury descriptions in QRCD in terms of broad severity, in that more serious cases were more likely to be described by police, thus biasing this comparison.

5.5.1.3 *Emergency Department Information System*

EDIS has very few variables that are coded. Road crash injuries were only able to be identified and many of the Core Minimum, Core Optional, and Supplemental variables were only able to be classified using the text description variable (presenting problem). While this variable had minimal missing data, there were varying degrees of specificity of the information. A manual review of the variable showed that approximately 40% of the cases lacked the specific information required to classify road user. It was also found that these ‘unspecified’ cases also had a gender and age bias. Specifically, the variable was more likely to be ‘unspecified’ for females and less likely to be unspecified for those aged 5-19 years. As with other data collections it is not clear what the underlying reason for these inconsistencies is. Nonetheless, it is important to note their impact on the conclusions drawn when using these data. The inconsistencies could introduce a bias in terms of the categorisation of road users.

Compared to QRCD, there were many more road crash injury cases included in EDIS. The profiles of EDIS and QRCD were compared and revealed that EDIS had a greater proportion of motorcyclists, cyclists, and younger people (aged 0-19). EDIS also had lower proportions of moderate and serious injuries (based on AIS), serious injuries (based on SRR), and fatalities. There were also location differences (as measured by ARIA+), however, as with QISU, this may simply represent the hospitals that are included in the EDIS collection.

5.5.1.4 *Queensland Injury Surveillance Unit*

There were issues relating to completeness for the *place* and *activity* variables. The implications of this missing information include the difficulty in identifying road crash cases.

When assessing the consistency of QISU in terms of ‘unspecified/unknown’ cases, it was found that there were some significant inconsistencies. For *place*, males, young people (aged 0-9), and motorcyclists were more likely to be ‘unspecified/unknown’. *Activity* was more likely to be ‘unspecified/unknown’ for females, drivers, passengers, pedestrians and Inner Regional areas, and less likely to be ‘unspecified/unknown’ for those aged 5-14 years. While it is not clear the underlying reason for these inconsistencies, it is important to note their potential impact on the conclusions drawn when using these data. The inconsistencies could introduce a bias in terms of the selection of cases.

The selection of transport cases seemed to be valid, with very high sensitivity and specificity. The selection of road crash cases was less valid, with moderately high sensitivity and specificity. As a result, the estimates of road crash injuries may include cases that should not be included and exclude some that it should not exclude. The road user classification was assessed for validity and found to have very high sensitivity and specificity for all road user types.

In terms of representativeness, QISU had considerably fewer cases than QRCD. It would not be expected that the prevalence of road crash injuries in QISU would correspond with that of QRCD, as QISU hospitals are only a subset of hospitals in Queensland at which a road crash injury could present. There were also profile differences between QISU and QRCD. Specifically, QISU had a greater proportion of motorcyclists, cyclists, younger people (0-19 years), and cases from Very Remote locations. It is possible that the age and ARIA+ differences were due to the hospitals that were included in the QISU collection. QISU includes a large hospital that exclusively treats children and does not include several large adult hospitals located in less remote locations. As a result, there is likely an inherent bias in the data collection based on the included hospitals. This same bias may explain the differences in the proportions of motorcyclists and cyclists, as a higher proportion of younger people are in these road user groups; however it does not necessarily explain it completely. Also, in light of the results in other data collections relating to these road user groups, it is possible that this is more evidence of under-reporting of motorcyclists and cyclists in QRCD.

QRCD had a greater proportion of fatalities and cases classified as serious (using SRR) compared to QISU. The difference between the collections in terms of fatalities is not surprising as there would be a considerable number of fatalities that are not taken to hospital (i.e., died at scene, died in transit). Generally, the differences in severity between

QRCD and QISU should be treated with caution due to the incompleteness and inconsistency of data in QRCD relating to injury (see Section 5.3.3.2).

5.5.1.5 *eARF (Queensland Ambulance Service)*

There was no coded variable to directly determine factors such as road user group, mode of transport, and counterpart, making the use of this data collection more problematic than the data collections already discussed. There were issues relating to completeness for the final assessment variable, transport criticality, and the text description. The implications of this missing information include the difficulty in assessing the nature and severity of injuries as well as identifying the different road users in the data, particularly drivers and passengers which are not able to be identified without the text description.

There were some consistency issues with these incomplete variables. For the variable *final assessment*, there were a greater proportion of ‘unspecified/unknown’ cases for ‘unknown’ gender, the very young (0-4) and older cases (75+), drivers and ‘unspecified’ road users. The text description was inconsistently ‘unknown/unspecified’ across the years of the data collection with more ‘unknown/unspecified’ cases in 2007 and 2008. It is possible that the reason for the larger amount of unspecified information in these years was due to the change from the paper based AIMS to the electronic eARF which occurred in 2007. It may be that ambulance officers improved the completion of this field as they became more familiar with the new data collection procedures. This is further supported by the fact that the amount of incomplete data went below 10% by 2009-2010.

The selection of transport cases seemed to be valid, with very high sensitivity and specificity. The selection of road crash cases was less valid, with high sensitivity, but only moderate specificity. It seems the variable used to identify road crash cases was successful at identifying correct cases, but less successful at distinguishing these cases from incorrect cases. As a result, the estimates of road crash injuries may include cases that should not be included. Finally, the road user classification was assessed for validity and found to have very high sensitivity for all road user types combined but only moderate specificity for drivers, passengers, and pedestrians. Similar to the road crash selection, it seems that the categorisation of road user type was good at identifying correct cases for inclusion, but tended to also include incorrect cases. It is possible that the lack of specificity for these road users was due to a reliance on a search of the text descriptions in order to identify them. As noted previously, drivers and passengers are not able to be distinguished from other road users based on any coded variable. While the text searching may be sensitive enough to include the cases it should, it may not be specific enough to avoid the selection of cases it should not. These validity issues could have serious impact on the estimates determined by the eARF data. It is possible that eARF overestimates the number of road crash cases and the involvement of drivers, passengers, and pedestrians.

In comparison to QRCD, eARF had fewer cases overall. It is not clear exactly why eARF has fewer cases than QRCD; however it may be due to the inclusion of minor injuries (which are not medically treated) in QRCD. It is possible that these are the cases where an ambulance was not in attendance. Further examination of this issue will be included in Chapter 7. Despite the overall greater numbers of cases in QRCD, eARF had a greater proportion of motorcyclists, cyclists, and passengers. The higher proportion of passengers may have been influenced by the possible overestimation of this user group discussed in the previous paragraph. However, this does not explain the greater proportion of motorcyclists and cyclists as the sensitivity and specificity was very high for both these groups. This result, similar to those found for QHAPDC; provides further evidence of under-reporting in the QRCD, particularly for these two road user groups. However, like QHAPDC, it is not possible to entirely distinguish between under-reporting and misclassification in eARF, particularly given the lack of specificity for selecting road crash cases.

5.5.1.6 *National Coronial Information System*

The Core Minimum, Core Optional, and Supplemental variables included in NCIS had a high level of completeness (no variables more than 5% missing or unspecified). The validity of the selection of road crash cases and the classification of road user types was good, with high sensitivity and specificity.

NCIS had two more cases than QRCD. It was expected that these data collections would match up exactly as all fatal road crash injuries should be reported to police and to the Coroner. However, there were discrepancies between the collections possibly indicating that the inclusion of road crash deaths in NCIS has a different basis than that of QRCD. For example, if a deceased person had a heart attack and then crashed their vehicle, but was found to have died of the heart attack, not injuries sustained from the road crash incident, QRCD would exclude it, but NCIS may still code one of the mechanisms of injury as transport and therefore be included in this data set. There are other issues of scope that could explain the difference between the data collections in terms of the numbers in this study (e.g., suicide). Also, the time taken for cases to be closed in NCIS could also be affecting the correspondence between the data collections, with the mechanism of injury for some NCIS cases not being finalised, and therefore unable to be extracted.

For the profiles, there were no statistically significant differences between the data collections in terms of gender, age, road user, or ARIA+. It should be noted however that the cases were not completely the same (not just in number but also in distribution), highlighting that there may be some issues with one or both of the data collections in terms of inclusion and/or coding.

5.5.2 *Study limitations*

A limitation of this study was that a number of variables were not able to be assessed for validity. In particular, the QRCD variables were not able to be assessed for validity at all. It is also possible that the proxy gold-standards (reference standard) used were not valid themselves. Specifically, the text fields that were used as a reference standard sometimes had missing cases or insufficient detail, making validity checks difficult. It is possible that, with text descriptions being relatively short, things such as road crash status and road user types were simply not recorded. This may particularly be true with health-related data collections, as it could be argued that these factors may not be clinically relevant, which is the major focus of these collections. Another issue is that in this study, it was not possible to determine with the prevalence and profile differences, how much of this was due to misclassification, or alternatively, under-reporting. Study 3 will attempt to address some of these issues by using data linkage.

While there are still some issues relating to the validity of case selection and road user classification, the benefits of using the health data collections in road safety research are clear. The health data collections contain information about road crash cases not reported to police and contain much more detailed and complete information about injury nature and severity. Both of these information gains have distinct benefit for understanding the nature of road crash injuries and their related costs. Therefore, the use of the health data collections in conjunction with police data (particularly if these data were linked) would potentially provide a more complete picture of the issue.

5.5.3 *Future directions for research*

While this study has identified some potential data quality issues for the QRCD as well as other data collections, and has developed selection criteria for case inclusion and the methods for creating variables, data linkage is required to confirm and expand these findings. Data linkage will be performed as part of Study 3, in the next Chapter, and will allow the data collections to be used as proxy reference standards for each other. For example, the QRCD only includes crashes that occur on designated roads, so for the cases that are linked to QHAPDC, the traffic coding within QHAPDC can be verified. This can be done for other data collections where it is determined that a particular data collection is a good reference standard for another. This will provide a better understanding of the validity of the key variables and selection criteria as well as potentially determining the level of under-reporting versus misclassification in the different data collections. Even for cases where a reference standard is not available, it will be possible to look at the level of convergence between the data collections.

5.6 Chapter Summary

This chapter described the second study conducted as part of the research program. It involved the secondary data analysis of six data collections which collect information relating to road crash injuries in Queensland. This second study was designed to explore research questions three and four. In doing so, it has provided insight into the quality of data collections relating to road crash injury in terms of completeness, consistency, validity, representativeness, and severity classification. The results indicate that there are limitations associated with the police collected Queensland Road Crash Database (QRCD), which is relied on for reporting and research in road safety, in terms of the broadness of the severity definitions and potential under-reporting. Also, the under-reporting, particularly for some road user groups, is problematic for road safety investigation, intervention development, and evaluation and could impact on the allocation of resources. A more precise measure of serious injury would be preferred over current practice as it is more closely related to threat to life and therefore more directly corresponding to the outcomes being measured when cost and impact is determined. Unfortunately, due to the large amount of missing information in police data, and the questionable accuracy of what is there, relying on police data to determine the prevalence and nature of serious injury crashes could be misleading. The inclusion of other data sources, such as hospital data, in the determination of serious injury crash impact has the potential to address the shortcomings of current approaches. However, these data collections often lack other information, which is included in police data, which are needed to determine the nature and circumstances of crashes (e.g., alcohol involvement, speed). As a result, data linkage (combining the data collections when they have individuals in common) is increasingly becoming a popular alternative to using individual data collections. Further research is required however, to assess the possibilities of data linkage, including its feasibility in the context of road safety. This issue will be addressed in the next chapter.

Chapter Six: Data Linkage Process and Assessment Framework

6.1 Introductory Comments

This chapter outlines a process of linkage that was developed to enable the linking of road crash injury data in Queensland. This process was based on the results of the review of data collections (Chapter 3), the interviews (Chapter 4), and the results of the secondary data analysis of non-linked data (Chapter 5). It also outlines the issues relating to how linkage will be assessed in terms of the scope of the data collections. It also highlights the issues relating to and implications of conducting linkage of this nature in Queensland and elsewhere. Finally, this chapter provides the basis for the methodology applied in Study 3 to undertake the linking of specific data collections.

6.2 Study Aims and Research Questions

The aim of the current study was to address the research questions below.

RQ4: What are the facilitators of and barriers to linking road crash injury data collections in Queensland and elsewhere?

> *RQ4c: What is the potential for linkage of the relevant data collections in terms of their common variables and scope?*
>
> *RQ4d: What is the most feasible process for conducting data linkage with road crash injury data in Queensland?*
>
> *RQ4e: What is the framework for assessing linkage success in terms of added information and added cases?*
>
> *RQ4f: What are the barriers to conducting data linkage with road crash injury data in Queensland?*

6.3 Potential for Linkage with Road Crash Injury Data Collections

After consultation with the relevant data custodians and the data linkage unit at QH, the variables in Table 6.1 were identified as those that could potentially be linked.

Table 6.1: Linking variables across QRCD, QHAPDC, EDIS, QISU, and eARF

Link	QRCD	QHAPDC	EDIS	QISU	eARF
Identifiers	-	UR number Facility number eARF number	UR number Facility number eARF number	UR number Facility number	eARF number
Name	Casualty name	Patient name	Patient name	-	Patient name
Date of birth	Casualty date of birth	Patient date of birth	Patient date of birth	Age of patient	Patient date of birth
Sex	Casualty gender	Sex of patient	Sex of patient	Sex of patient	Sex of patient
Address	Casualty address	Address of usual residence	Address of usual residence	Postcode of usual residence	Patient address

It was also necessary to determine the potential for linkage in terms of the selection of cases. Specifically, it was necessary to identify the cases in each data collection that fit the definition of the population of interest (i.e., road crash injury). The following was determined as the selection criteria for each collection to attempt the capture of the population of interest:

- QRCD: all injury cases
- QHAPDC: all cases coded as transport-related (ICD-10-AM External Cause Codes from V00-V99)
- EDIS: all cases coded as an injury (discharge diagnosis S00-S99 and T00-T98)
- QISU: all cases coded as transport-related (external definition of 'motor vehicle – driver'; 'motor vehicle – passenger'; 'motorcycle – driver'; 'motorcycle – passenger'; 'pedal cyclist or pedal cyclist passenger'; and 'pedestrian')
- eARF: all cases with a coded case nature of: motor vehicle collision; bicycle collision; motorcycle collision; and pedestrian collision

A broad approach to the linkage was chosen for two reasons. Firstly, there was no coding that could specifically identify the relevant cases (i.e., EDIS has no coding to identify transport-related cases or more specifically road crash cases). Secondly, the coding for road crash cases (traffic) was in question (i.e., QHAPDC, QISU, and eARF). This approach was based on discussions with custodians as well as the results of Study 2 (see Chapter 4).

Consideration was also given to how the data collections corresponded with each other in terms of their scope. It was not expected that all cases within a particular data collection would be included in another. For example, QHAPDC only includes cases admitted to hospital, so therefore only those cases identified as admitted to hospital in EDIS or QISU could possibly match with a case in QHAPDC. Table 6.2 outlines how the data collections correspond to each other in terms of cases potentially in common.

Table 6.2: Commonalities in the data collections

	QRCD	QHAPDC	EDIS	QISU	eARF
Fatality	Fatality (*Casualty Severity*)	Died in hospital (*Mode of discharge*)	Died in ED (*Mode of discharge*)	Died in ED (*Mode of separation*)	Unknown
Hospitalised	Unknown	By definition	Admitted (*Mode of discharge*)	Admitted (*Mode of separation*)	Unknown
Taken to hospital	Hospitalised (*Casualty Severity*)	-	By definition	By definition	Transported to hospital (*Patient status*)
Other injury	Medical treatment and minor injury (*Casualty Severity*)	-	-	-	By definition

It should be noted however, that because it was not possible to determine the exact correspondence between some collections (e.g., QRCD and eARF have no coding to determine if someone was admitted to hospital) and the validity of the coding of the some of the cases is unknown, all cases in each data collection within the scope of the requested data (as described above) will be attempted to be matched. The correspondence between the collections will be used to explain where linkages may not have occurred, not due to linkage error, but simply because they are out of the scope of each of the particular collections. In other words, these correspondences will inform the assessment framework for linkage (see Section 6.5). In addition to these correspondences, the coding or selection of road crash cases specifically will also need to be considered. It is not expected that off-road transport injury cases will be included in QRCD by definition. Therefore, the correspondence between coded road crash cases in the health data collections and QRCD cases will also need to be taken into account (see Section 6.5).

Another important aspect for the data collections’ potential for linkage relates to whether the data required for linkage (as shown in Table 6.1) are able to be provided. Essentially, there needed to be a mechanism and/or a process that allows the release of data from the custodian to the data linkers. QHAPDC, EDIS, QISU, and eARF all had pre-existing mechanisms to allow for the sharing of identifying data with the appropriate agency for linkage purposes. For QHAPDC, EDIS, and QISU, this is allowed through the ethics approval and the Public Health Act (2005) application. For eARF data, ethics was also required as well as approval from the Commissioner. Access to these data was the same as that for access to the data generally (see Chapter 3, Section 3.4.6) although it did require the applications to be explicit about the release of personal information (to the data linkage unit, not the researcher) and the data linkage process.

For QRCD, as discussed previously (Chapter 3, Section3.3.1), the release of identifying information to other government agencies for the purposes of linkage was possible with a Memorandum of Understanding between the relevant agencies. As a result of negotiations for the completion of this research project, TMR and Queensland Health (QH) signed an MOU allowing for TMR to provide identifying information (name, address, date of birth, date of crash etc.) to QH for the purposes of linking with data QH hold (e.g., Emergency Department Information System). The MOU only allows for the release of the identifying information required for linkage and does not allow the sending of any ‘content’ (specific details of the crash) information to external agencies. The MOU extends beyond the current project to allow researchers in the future to also access linked data if prescribed conditions are met.

Another requirement was to determine if there was appropriate linkage infrastructure. Specifically, it was necessary to determine whether there was a unit or group that had the hardware, software, and skills required to conduct the linkage of these data collections. After discussions with custodians, data linkage experts, and data users it was determined that the data linkage unit with Queensland Health would be appropriate to conduct the linkage. This unit has the hardware, software, and capacity to conduct it. They are also identified as the linkage unit for health data linkage in Queensland by the Population Health Research Network.

The final issue was that the exact process for conducting the linkage needed to be determined. This was done through discussions with custodians and the identified data linkage unit, as well as based on the literature (Chapter 2) and interviews about linkage (Chapter 4, Section 4.4.2). The process of linkage that will be used in this program of research is described in the next section.

6.4 Process of Data Linkage of Road Crash Injury Data in Queensland

The process for linking the selected data collections was based on the ‘separation principle’. Each data custodian provided personal information (see in Section 6.3, Table 6.1) to the QH Data Linkage Unit (DLU). The cases selected were based on the specifications described in Section 6.3. No content (clinical) information was required to be sent to the data linkage unit. The DLU then used these personal data to create links between the data collections. For every link that was found a linkage key was applied in the form of a common person ID. This common person ID was assigned to each common case in the link. A unique person ID was also assigned to those cases that did not link to any other data collection. All person IDs (for links and non-links) were then sent back to the data custodians. The data custodians then attached the person IDs to the content (clinical) data and sent this de-identified data to the researcher. The linkage process is presented in Figure 6.1.

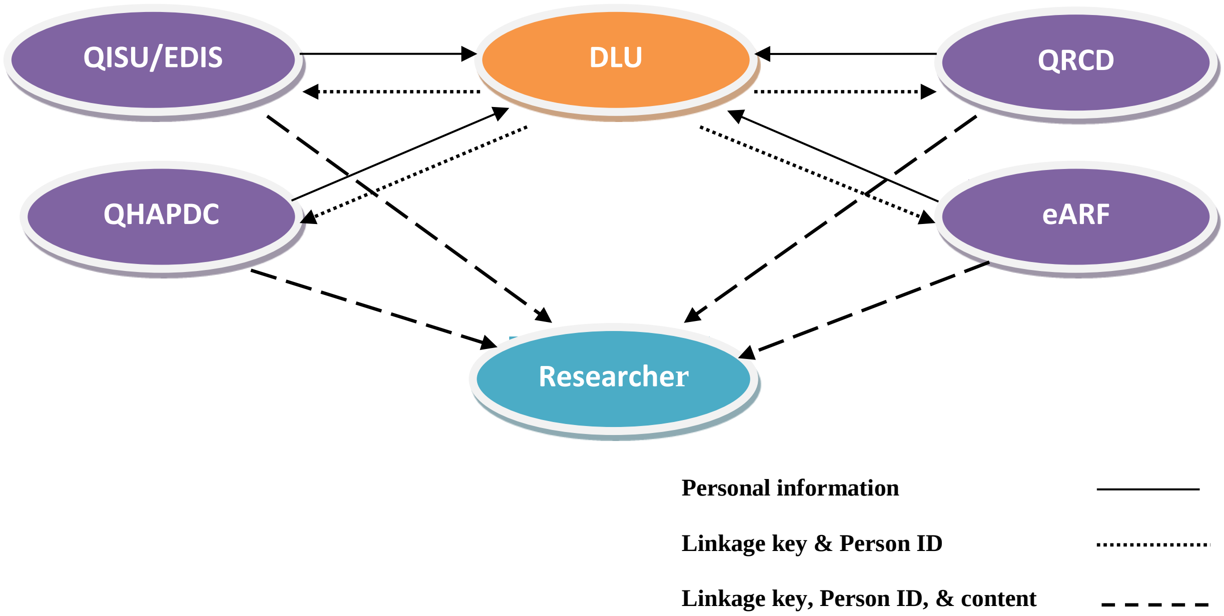


Figure 6.1: Data linkage process

The researcher received no identifying information (i.e., name, address, date of birth) and received all the cases selected including those in which there was no link.

The linkage process as conducted by Queensland Health involved a combination of deterministic linkage (when unique IDs are in common, e.g., UR and Facility Number), probabilistic linkage, and manual review (for grey matches).

In summary, the linkage process required: ethics approval; approval from all custodians; a Public Health Act agreement; and the Memorandum of Understanding being signed between the QRCD custodian and Queensland Health.

6.5 Assessment Framework for Linkage

6.5.1 *Linkage success*

Linkage rates will be calculated to determine the number of QRCD cases that can potentially have extra information added by linking QRCD with other data collections. In these analyses, the QRCD will be the data collection of reference, with the linkage rate determined by examining how many cases in QRCD link to a case in another data collection or collections. Specifically, as shown in Figure 6.2, the linkage rate will be calculated as follows:

Linkage rate % = C/A x 100

With (C) being the number of cases that link with the other data collection/s and (A) being the number of QRCD cases.

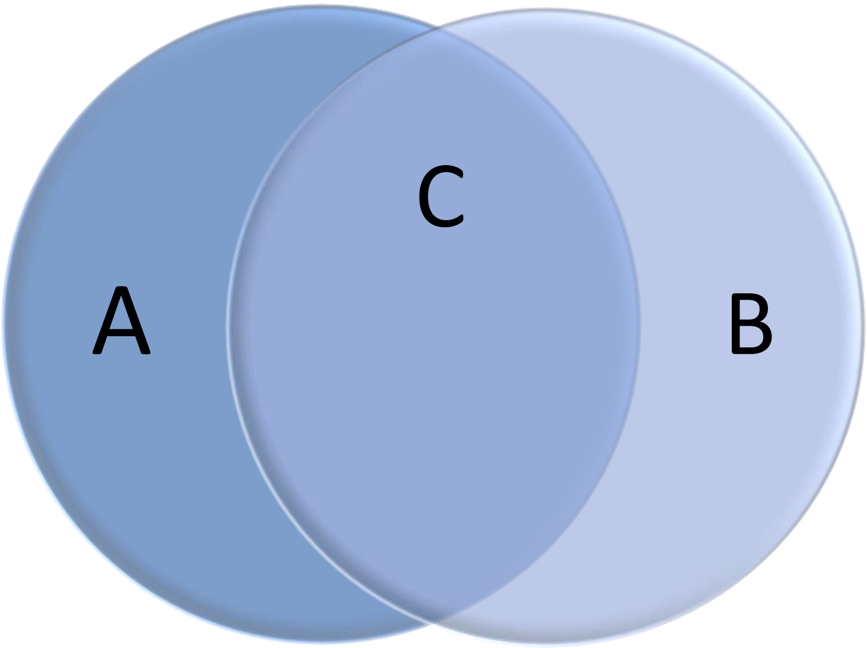


Figure 6.2: Correspondences for linkage rate

Table 6.3 outlines how the linkage rates will be calculated for each one-to-one linkage (i.e., QRCD with each other data collection). In each case, the denominator will be based on the count of QRCD cases. Linkage rate 1 will involve all QRCD injury cases, while linkage rate 2 will involve only those QRCD cases coded the same as the scope of the other data collection. For example, it would not be expected that ‘other injury’ cases in QRCD would be included in EDIS, QISU, or QHAPDC, so if few, or no links occur at that level, this should not necessarily be interpreted as linkage error. For the numerator, it will be a count of linked cases between QRCD and the respective data collection.

Table 6.3: Linkage rates for QRCD with each other data collection

	eARF	EDIS	QISU	QHAPDC
Linkage rate 1 (Broad)	No. of linked cases/No. of QRCD injury cases	No. of linked cases/No. of QRCD injury cases	No. of linked cases/No. of QRCD injury cases	No. of linked cases/No. of QRCD injury cases
Linkage rate 2 (Specific)	No. of linked (QRCD with eARF) medically treated and ‘hospitalised’ injury cases /No. of QRCD medically treated and ‘hospitalised’ injury cases	No. of linked (QRCD with EDIS) ‘hospitalised’ injury cases/No. of QRCD ‘hospitalised’ injury cases	No. of linked (QRCD with QISU) ‘hospitalised’ injury cases/No. of QRCD ‘hospitalised’ injury cases	No. of (QRCD with QHAPDC) ‘hospitalised’ injury cases/No. of QRCD ‘hospitalised’ injury cases

The next step will involve calculating linkage rates between multiple data collections and QRCD. As shown in Table 6.4, in addition to the one-to-one linkage rates (two data collections) described above, every combination (15 combinations) of the data collections will be merged with QRCD (three, four, and all data collections).

Table 6.4: Combinations of QRCD with other data collections

	QRCD	QHAPDC	EDIS	QISU	eARF
Two data collections					
	✓	✓			
	✓		✓		
	✓			✓	
	✓				✓
Three data collections					
	✓	✓	✓		
	✓	✓		✓	
	✓	✓			✓
	✓		✓	✓	
	✓		✓		✓
	✓			✓	✓
Four data collections					
	✓	✓	✓	✓	
	✓	✓	✓		✓
	✓	✓		✓	✓
	✓		✓	✓	✓
All data collections					
	✓	✓	✓	✓	✓

The health data collections will also be represented as a combination in terms of their respective scope. Specifically, EDIS and QISU will be combined to form an emergency department data set; QHAPDC, EDIS, and QISU will be combined to form a *hospital data* set; and QHAPDC, EDIS, QISU, and eARF will be combined to form a *health data* set. These data collections will be combined in a way that takes into account the unique cases from each collection and the common cases. For example, the emergency department data set will include all cases that are either in EDIS or QISU, or both.

The linkage rates will be calculated for each combination, with each of the data collection combinations being used to determine the total number of cases that link across the entire combination. For these calculations, cases in QRCD will be considered linked if they link with every data collection in the combination (e.g., QRCD, QHAPDC, and EDIS). These linkage rates will enable an assessment of how many cases are enhanced (information gained) by being linked through the data collections (e.g., from ambulance through to admission).

It is also of interest however, to determine how many QRCD cases could potentially be enhanced (information gained) by linking with at least one of the data collections. For these calculations, cases in QRCD will be considered linked if they link with at least one of the other data collections in the combination (e.g., the number of QRCD cases linked with QHAPDC, EDIS, QISU, or eARF). The linkage rates from each combination will be compared to determine the unique contribution of each data collection to the linkage rate. This will then be used to determine the most parsimonious combination of data collections for optimum information gain. It is possible that adding a data collection to the linkage provides very few additional linked cases and therefore may not be worth including in future linkage.

There will be some examination of the links between the other data collections separate to the links with QRCD, but only for the purposes of examining data quality issues such as validity (see Section 6.5.3).

6.5.2 *Linkage and completeness*

In addition to the linkage rates discussed above, the number of additional cases that the *health* data collections may identify will also be assessed. This discordance between the *health* data collections and QRCD will give an indication of the underestimation of the population in QRCD, either due to under-reporting or misclassification. Specifically, unlike with the linkage rate, the reference for discordance will be the cases in the *health* data collection and the discordance rate will be determined using the following formula:

Discordance % = (1 - (C/B)) x 100

Where C is the number of cases linked and B is the number of cases in the *health* data collection (see Figure 6.2). Similar to the calculation of linkage rates, the discordance rate will be determined in two ways: one for all transport cases identified in each *health* data collection and the other for those whose coding is consistent with the definition of a road crash (see Table 6.5 for more details). As with the linkage rates, these discordances will be influenced by the validity of the coding (see Section 6.5.3).

Table 6.5: Discordance between each health data collection and QRCD

	eARF	EDIS	QISU	QHAPDC
Discordance 1 (Broad)	No. of eARF cases that do not link to QRCD / No. of eARF cases	No. of EDIS cases that do not link to QRCD / No. of EDIS cases	No. of QISU cases that do not link to QRCD / No. of QISU cases	No. of QHAPDC cases that do not link to QRCD / No. of QHAPDC cases
Discordance 2 (Specific)	No. of road crash coded eARF cases that do not link to QRCD / No. of road crash coded eARF cases	No. of road crash coded EDIS cases that do not link to QRCD / No. of road crash coded EDIS cases	No. of road crash coded QISU cases that do not link to QRCD / No. of road crash coded QISU cases	No. of road crash coded QHAPDC cases that do not link to QRCD / No. of road crash coded QHAPDC cases

As none of the other data collections are the ‘gold standard’ for the population (none of the data collections represent all road crash injuries), using this one to one discordance will not provide a very accurate assessment of the under-representation of QRCD. In order for a more accurate assessment to be made, the *health* data collections will need to be used in combination to determine the population of comparison. Due to the scope of each of the data collections (e.g., QHAPDC - all admitted cases) if these data collections were combined, one would expect to have a better representation of the population than any data collection on its own, and would therefore make a better reference for comparison with QRCD. So as part of the assessment of discordance, a population estimate will be calculated based on the combination of the *health* data collections (after accounting for overlap/linkage of cases within these collections). A data set will be created from this population estimate to represent the largest possible ‘population’ of road crash injuries. Then the number of cases in QRCD not linking with this combined data collection will be considered the initial estimate of under-reporting or under-representation in QRCD. The data set, based on the estimated population of road crashes, will also be used to compare the profile (e.g., gender, ARIA+, road user) of road crash injuries produced from these data with the profile produced from using QRCD cases only. The application of the capture-recapture method was explored as a possibility for estimating the population; however the assumptions of this analysis could not be met with the current data sources. In particular, it could not be assumed that the cases in each *health* data collection are accurately identified as a road crash. As shown in Chapter 5, Section 5.4.3, the selection of road crash cases in the health data collections (particularly EDIS) may not be valid. To equate this with a biological example (from which the capture-recapture method is most commonly applied), if a researcher wanted to estimate the population of fish in a lake, for the capture-recapture method to work, the researcher would need to be certain that they are only counting fish, not some other animal.

Another aspect that will be explored is the completeness of the variables. For QRCD cases that link to another collection, as mentioned above, it is possible that additional information could augment the details included in the QRCD. This information may take the form of added variables, more valid variables, or more complete variables. For example, it is expected that additional information relating to the severity and the nature of injury will be identified. Using linked cases for each combination (as determined above), the amount of additional information provided by each combination of data collections will be determined. This will help identify the most parsimonious combination of data collections for optimal information gain. It should be noted however, that this will obviously only be able to be assessed for the linked data and any bias (Section 6.5.5) or other impacts on linkage success (e.g., validity and quality of linkage, Section 6.5.2 and 6.5.4 respectively) will need to be considered. It should also be noted that for the purposes of assessing the amount of additional information, all linked cases will be considered, not just those that fit the scope of the corresponding data collection (e.g., all linked cases between QRCD and QHAPDC, not just ‘hospitalised’ and fatal QRCD cases that link with QHAPDC).

6.5.3 *Linkage and validity*

As mentioned above, the linkage rates may be affected by the validity of some of the coding within the data collections. As discussed above, the broad severity coding (e.g., hospitalisation) in QRCD may not always be accurate, which could influence the linkage rates. Also, in terms of discordance between the *health* data collections and QRCD, while it is not expected that non-traffic (road crash) cases would be linked, it is possible that the coding of traffic (road crashes) in QHAPDC, QISU, EDIS, and eARF is not always correct. As result, validity assessments of these key variables will need to be conducted and any inaccuracies taken into account when assessing the 'success' of the linkage.

Validity will be assessed in a number of ways, including using the linkage combinations from above, as well as links between the *health* data collections, to identify false negatives and false positives where possible. For example, a false negative for the coding of traffic-related in QHAPDC would be when a case is not coded as traffic in QHAPDC, but does link with a case in QRCD. False positives on the other hand will be those when a case is coded as traffic in QHAPDC but does not link with a case in QRCD. It should be noted however, that false positives will be influenced by the completeness of the data collections that are being compared. As an example, there may be traffic cases in QHAPDC that do not link with QRCD, not because the coding of traffic is incorrect in QHAPDC, but because the case was not reported to police. The use of additional linked data collections may be able to assist with these issues somewhat, by examining the coding of the case in other collections. In this sense, convergent validity assessments will apply. Specifically, validity will be examined by the level of convergence between the data collections. If the same coding is applied to a case in more than two data collections, there could be some degree of confidence that the coding is correct. This is where the linkage between the data collections other than QRCD will need to be considered. It should be noted however that this convergence will only be able to be assessed for those cases that link. As a result, there will still be cases where the validity of the coding will be unknown. Despite this however, it will give some indication of the influence of validity issues on assessing linkage 'success'.

6.5.4 *Quality of the linkage*

Another aspect in assessing the 'success' of data linkage is that of the quality of the linkage process as conducted by the QH data linkage unit. Cases may not link between data collections, not because the data collections do not include them, but rather that the information required for linkage was inaccurate and/or incomplete. For example, it is expected that there would be misspellings, date of birth errors, and missing fields in the data collections that will make linkage difficult. As the author will not be conducting the linkage, the assessment of data linkage quality, in terms of errors in linkage, can only be based on the information provided by the linkage unit within QH.

It may not be possible to exactly determine how much of the discrepancy between the data collections is due to errors in the data and/or the linkage process as opposed to actual discrepencies (e.g., road crash injury admitted to hospital not reported to police).

However, attempts will be made to quantify these issues as much as possible in determining the ‘success’ of the linkage.

6.5.5 *Linkage bias*

Another issue to be examined as part of the linkage assessment framework is the extent of potential bias in linkage. It is important to establish any bias in linkage for a number of reasons. Firstly, it will allow some level of quantification of not just the amount of under-reporting in QRCD, but also whether this under-reporting is more likely to occur in certain circumstances. If this under-reporting bias can be effectively quantified, adjustments in the reporting of road crash injuries can potentially be performed.

The other aspect of bias in linkage relates to the future use of linked data for road crash injuries. If researchers and policy makers are to use linked data as the basis for the reporting and assessments of severity and nature of injury (an example of the additional information provided by these other health related data sources), they will only be reporting on those cases that link. Therefore, if there is a bias in the cases that link, then there will be a bias in what is reported and/or estimated. If for example, more serious injuries link better or injuries that occur in major cities link better than other cases, then a skewed view of the road crash injury problem will result. If the bias is profound, then it may be that using linked data could cause more problems than it solves.

As part of the assessment of data linkage therefore, the linkage rates will be compared across different characteristics that may influence the ‘success’ of the linkage. Also, the discordance between QRCD and the other data collections will be compared across these same characteristics to determine any bias in under-reporting of cases. These will include comparisons based on: age, gender, road user type, ARIA+, and broad severity. These characteristics were chosen based on the literature (Chapter 2), the interviews in Study 1 (Chapter 4), and the results of Study 2 (Chapter 5).

6.6 Summary of Issues Relating to the Conduct of Data Linkage

Each of the data collections includes identifying information that could be used to conduct linkage. However, it is not clear how complete and accurate this information is. While the data linkage will be conducted by another party (not the researcher), and therefore the exact nature of the accuracy and/or completeness of these data may not be known, Study 3 (Chapter 7) will examine some of these issues based on a report provided to the researcher by the data linkers.

Another issue relating to the capability of linkage with these data collections relates to whether the necessary data was able to be provided to the data linkage unit. A review of the legislation and discussions with the custodians determined that this was possible with each of the data collections. It should be noted however, that a further agreement between the data linkers and the QRCD custodian was required. Without this agreement, release of identifying data for the purposes of linkage may not have been possible. However, now that this agreement is in place it has allowed for this project to occur and for other data linkage projects using these data to occur in the future.

The process of data linkage chosen complies with the best practice approach for data linkage used in Australia and other parts of the world. It applies the separation principle which allows for the linkage of data and research using linked data without researchers having access to personal identifying information. While this procedure is based on best practice, it has not been tested for these data in this jurisdiction before, so it is yet to be determined whether the process chosen will be successful and what specific barriers may occur with this process. This will be addressed in Study 3 (Chapter 7) via an application of the data linkage process with these data.

6.7 Chapter Summary

This chapter has explored the potential for data linkage with the identified data collections and has identified the most appropriate process for data linkage given current legislative and organisational circumstances in Queensland. The data collections do correspond with each other to some extent in terms of scope and variables required for linkage, meaning that links between the data collections are possible. An appropriate process for this linkage has been determined based on discussions with data custodians, linkage experts, and users and is based on international best practice linkage. It has also been determined that the release of data required for the linkage to occur is allowable.

This chapter has also described the assessment framework that was used for the linkage. This included the basis for assessing linkage 'success', issues relating to validity, completeness, and linkage bias. The framework has been developed based on the literature review, interviews, and results of Study 2. A more detailed description of the methodologies in the assessment framework as it applies to this linkage is in the next chapter.

While the linking of road crash injury data appears to offer a range of benefits, it remains to be determined how successful linkage will be in the Queensland context and how linked data may provide benefit over non-linked data both qualitatively and quantitatively. An application of data linkage to road crash injuries, including the use of the assessment framework, will be the topic of Chapter 7.

Chapter Seven: Outcomes of Data Linkage

7.1 Introductory Comments

This chapter outlines the third study conducted as part of the research program. It involved secondary data analysis of the linkage between five data collections which include road crash injury information in Queensland:

- Queensland Road Crash Database;
- Queensland Hospital Admitted Patients Data Collection;
- Queensland Ambulance Service (eARF);
- Queensland Injury Surveillance Unit; and
- Emergency Department Information System.

It builds on the findings of Study 2, and includes analysis relating to linkage rates, discordance, validity, and profiles of different combinations of linked data sources. It specifically examines the potential for linked data to enhance the quantification of serious injury and explores issues such as under-reporting of road crash injuries to police.

7.2 Aims and Research Questions

This section of the research aimed to address research question five as described in Chapter Two, Section 2.6. Sub-questions for each of the broad research question are outlined below.

RQ5: What aspects of road crash injury data quality can be improved by using linked data for road safety investigation, intervention development, and evaluation?

> *RQ5a: How many cases in QRCD link to other data collections?*
>
> *RQ5b: How much bias, if any, is there between QRCD cases that link and those that do not link in terms of characteristics such as remoteness, gender, age, and road user type?*
>
> *RQ5c: What is the estimated amount of under-reporting in QRCD?*
>
> *RQ5d: How much bias, if any, is there in the amount of under-reporting in terms of characteristics such as remoteness, gender, age, and road-user type?*
>
> *RQ5e: What extra information, specifically relating to severity, can other data collections provide for linked cases in QRCD?*
>
> *RQ5f: How does the profile of linked QRCD cases differ from the profile of QRCD alone in terms of gender, age, road user, remoteness, and serious injury classification?*

RQ5h: How valid is the coding of case inclusion (road crash) and other attributes (e.g., road user) in each of the health data collections?

7.3 Method

Ethics approval was obtained from the Queensland Health Human Research Ethics Committee (#HREC/12/QHC/45). A Public Health Act agreement was completed by the researcher and signed by each of the Queensland Health (QH) data custodians (EDIS, QHAPDC, and QISU) and the Queensland Health Research Ethics and Governance Unit. Approval was also provided by the Queensland Ambulance Commissioner via mail correspondence. QRCD data was provided following approval (via designated form) from the Manager of the Data Analysis Unit at the Department of Transport and Main Roads (TMR) and a Memorandum of Understanding being signed between TMR and QH.

7.3.1 *Data linkage process*

Information was provided to the researcher by the Queensland Health Data Linkage Unit documenting the linkage process and related output. Person details and demographic data were linked using linkage software applying deterministic & probabilistic methodologies, as well as manual clerical reviews where required. Approximately 100,000 pairs (20%) of pairs were considered grey matches and were manually reviewed. Most of these were considered grey matches due to minor errors in the spelling of names. Due to the extensive manual review, the researcher was told that it was not possible to calculate specificity or sensitivity of the links. The DLU did however comment that they believed the quality of the linkage to be very high and, if anything, may have missed true links, rather than linked cases that should not have been linked.

For the current study, the time taken to gain ethical clearance and data custodian agreements was approximately twenty months. Due to issues with some of the data (incomplete or incorrect personal information), a large number of manual reviews needed to be conducted, so the data linkage process conducted by Queensland Health took approximately five months. As a result, the time taken from applying for ethics to obtaining the data was over two years.

7.3.2 *Data characteristics*

Data were provided from QRCD, QHAPDC, EDIS, QISU, and eARF by each relevant custodian for the specified cases for 2009 as described in Chapter 6, Section 6.3. This selection was also based on the Study 2 selection and represented the cases in each data collection for the year 2009 (1st January, 2009 to 31st December, 2009) that could potentially be considered a road crash injury. The focus of the current study was on serious non-fatal injuries so fatal cases were not included in the analysis. This focus was based on the results of Study 2, which highlighted that there was very little discrepancy between QRCD and NCIS. The variables in each data collection were the same as those used in Study 2 (Chapter 5, Section 5.3.2) with the exception of the person ID added by the DLU to allow for linking across data sets. The process by which these person IDs were attached is described in the previous chapter (Chapter 6, Section 6.4). Details of the cases included in each collection are provided below.

7.3.2.1 *QRCD*

All cases in QRCD for 2009 were included for the study (n = 19,041).

7.3.2.2 *QHAPDC*

QHAPDC had 14,820 land transport cases for 2009 (cases coded with ICD-10-AM External Cause Code from V00-89). As QHAPDC is episode based, there were some duplicate cases in the data. The first case for an individual was considered the index case. For cases where the admission date of a duplicate case was within one day of the discharge date of the index case, the duplicate was removed. When there was more than one duplicate case, if each subsequent case had an admission date within one day of the discharge date of the previous case it was also considered part of the same injury series. If the duplicate was after this date, it was counted as a new injury case (i.e., the person was injured in a separate event). With the removal of duplicates (17.7%), QHAPDC included 12,198 land transport cases. An example of the duplicate removal process is presented below in Table 7.1.

Table 7.1: Example of duplicate removal process

Person ID	**Date of admission**	**Date of discharge**	**Case**
1001	13/01/09	15/01/09	Index
1001	15/01/09	27/01/09	Duplicate
1001	28/01/09	31/01/09	Duplicate
1001	25/02/09	30/02/09	Index

7.3.2.3 *eARF*

All cases attended by an ambulance in Queensland that involved a case nature coded as ‘motor vehicle collision’, ‘motorcycle collision’, ‘bicycle collision’, ‘pedestrian collision’, ‘crush’, and ‘fall’ (n = 72,847).

7.3.2.4 *QISU*

All cases with an *external definition* coded as ‘motor vehicle – driver’; ‘motor vehicle – passenger’; ‘motorcycle – driver’; ‘motorcycle – passenger’; ‘pedal cyclist or pedal cyclist passenger’; and ‘pedestrian’ were included from QISU (n = 5,127). Duplicates were identified and removed using the same method as used for QHAPDC (Section 7.3.2.2). The total number of unique QISU injury cases for the study was 5,071.

7.3.2.5 *EDIS*

All cases coded with a discharge diagnosis between S00-S99 and T00-T98 were included from EDIS (n = 315,491). Duplicates were identified and removed using the same method as used for QHAPDC (see Section 7.3.2.2). The total number of unique EDIS injury cases for the study was 303,870.

7.3.3 *Data merging and linkage coding*

In order to assess the 'success' of the linkage, QRCD was merged with the other data collections. The first set of merges was the one-to-one linkages (e.g., QRCD and QHAPDC). The data sets were merged based on the person ID. If the person ID of a QRCD case matched the person ID of a case in the other data set, then the case was considered to be a link and was coded as such. Non-links, for the purposes of calculating linkage rates and analysing linkage bias, were all cases in QRCD that did not have a person ID in common with any case in the other data collection. Non-links, for the purposes of calculating discordance rates and analysing discordance bias, were all cases in the other data collection that did not have a person ID in common with QRCD.

Merges were then conducted with all other combinations of linkage (see Chapter 6, Section 6.5). Links were then coded in two ways. The first were those cases where the person ID was common between QRCD and all other data collections in the combination. The second were those cases where the QRCD person ID matched the person ID in at least one of the other data collections in the combination. Each of the health data collections were combined to create two population estimates. The hospital (i.e., presented at hospital) data collections (i.e., QHAPDC, EDIS, and QISU) were combined to form a *hospital population* data set and another combined data set with eARF included was formed to provide a *health* data set. These data sets included all cases from each collection that linked to each other as well as the unique (non-linked) cases from each data collection. These combined data sets were then used to assess the convergent validity of coding and as the basis for the population estimates for comparison to QRCD.

7.3.4 *Coding of road crash injury cases and variables*

As with Study 2, cases were also coded based on their alignment with the Queensland Road Crash Data definition of a road crash injury (i.e., resulted from an incident that occurred on a public road and involved a moving vehicle). Only cases that were specifically coded or directly identified in text were coded as a road crash injury case as was done in Study 2 (see Chapter 5, Section 5.3.2). If a case was coded as unknown, unspecified, or other category it was not coded as a road crash injury even though it may be a road crash case.

In terms of variables, as with Study 2 (see Chapter 5), variables were coded to examine bias of linkage and discordance as well as validity and completeness. These codings included: age, gender, severity of injury (broad, AIS, and SRRs), ARIA+, and road user type. The processes for coding these variables were identical to those used in Study 2 and are described in Chapter 5, Section 5.3.2. In addition, a variable 'collision' was also created, where possible, to assess the linkage and discordance on the basis of whether another vehicle was involved.

Collisions were coded for QRCD, QHAPDC, and QISU as follows:

- QRCD collisions were all cases with a *crash nature* of: angle; rear-end; head-on; sideswipe; and hit pedestrian.
- QHAPDC non-collisions were all cases with an *external cause code* of V17, V18, V28, V38, V48, V58, V68, and V78. Collisions will be all other cases.
- QISU collisions were all cases with a *mechanism* of ‘contact with a moving object’ or ‘contact with a person’.

Collisions were not able to be coded for EDIS or eARF.

A summary of the selection criteria for cases and the coding of variables are available in Appendix H as a pull-out A3 sheet for reference.

7.3.5 *Analysis*

7.3.5.1 *Assessing linkage rate*

The number of linked cases and the proportion of cases in QRCD that linked to each other data collection was produced for every combination of linkage as described in Chapter 6, Section 6.5.1

7.3.5.2 *Assessing completeness of cases (discordance rate)*

As described in Chapter 6, Section 6.5.1, in contrast to the linkage rate, the discordance rate was calculated by expressing the number of non-linked cases as a proportion of the health data collection(s).

7.3.5.3 *Assessing completeness of variables*

The level of completeness in terms of the field completeness for QRCD severity of injury was examined for the cases that were linked to each combination of integrated data sets by identifying the proportion of: missing; unknown; other specified; and unspecified values recorded for the severity of injury variable (as represented by the injury description). Injury description is the only variable that was assessed, because in Study 2, it was identified as having more than 10% missing or unspecified cases in QRCD.

In order to assess the completeness of the information about injury severity provided by linkage, the variables in each data collection relating to injury coding were combined from the different data collections to produce a combined variable. In cases where more than one health data collection was combined with QRCD, there was a hierarchy for selection of which data collection would provide the data in the variable. For example, if the linkage between QRCD, QHAPDC, QISU, and EDIS was being examined, if the case has a specified ICD-10-AM principal diagnosis code in QHAPDC, this was the code that was used. The ICD-10-AM coding in QISU was used when QHAPDC was not available and the ICD-10-AM code for EDIS was used in cases where neither QHAPDC nor QISU code is available. This hierarchy was based on the assumption that QHAPDC coding of injury is superior to QISU and EDIS, as it is completed by trained coders with access to the full medical records of the patients. QISU would be considered next best, as it has coded information for most variables, and EDIS last, as many of the variables rely on being created from text searching. It should be noted that for the assessment of the additional completeness of severity of injury, eARF was not included as it does not include any coding of injury nature and therefore was not able to be coded into AIS or SRR severity.

The completeness of this variable for each linked combination was compared with the entire QRCD collection to determine how many more complete cases were available if QRCD is linked to the other data collections as opposed to using QRCD alone.

7.3.5.4 *Assessing bias/consistency*

Linked and non-linked cases were compared on a number of characteristics that may influence the linkage and/or discordance rates. Specifically, linked and non-linked cases were compared on: age, gender, road user, severity (broad severity, AIS and SRR), and ARIA+. The classification of these variables was the same as that used in Chapter 5, Section 5.3.2. In all cases the variables used were those of the reference data collections (e.g., QRCD variables for linkage rates, QHAPDC variables for discordance rates with QRCD). This was done as the reference data collection was the only collection that has data for all the cases (linked and non-linked). Comparisons were made using Chi-square tests of independence. Due to the large sample size, a more stringent alpha of .001 was adopted. Also, Cramer's V (ϕ_c) was calculated in order to provide an estimate of effect size to give a clearer idea of the meaningfulness of any statistical significance found. As suggested by Aron and Aron (1991), a Cramer's V of less than .10 was considered to be a small effect size, between .10 and .30 moderate, and more than 0.30 a large effect size. Post-hoc analyses were also undertaken using an adjusted standardised residual statistic. This statistic can be used to identify those cells with observed frequencies significantly higher or lower than expected. With an alpha level set at 0.001, adjusted standard residuals outside -3.10 and +3.10 were considered significant (Haberman, 1978). As with Study 2, logistic regressions were performed to take into account the relationships between the predictors (e.g., gender and road user). It should be noted that age needed to be re-categorised into four groups (0-16; 17-24; 25-59; 60+) due to violations of linearity in the relationship to the outcome when treated as ordinal (5 year intervals). Referent categories for the predictors in logistic regression were chosen based on either the absence of a condition (e.g., non-serious) or the group with the largest proportion of injuries (e.g., Major Cities, drivers, 25-59 age group).

7.3.5.5 *Profiling serious injury*

Using the different combinations of linked and non-linked data, the following estimates of the number of serious injuries will be produced:

- Police reported hospitalisations (QRCD)

- Hospital attendances (EDIS, QHAPDC, QISU)
- Hospital admissions (QHAPDC)
- Hospital admissions of 24hrs or more (QHAPDC)
- Police reported hospital attendances (QRCD linked with hospital)
- Police reported hospital admissions (QRCD linked with QHAPDC)
- Police reported hospital admissions of 24hrs or more (QRCD linked with QHAPDC)
- Police reported serious injuries as defined by AIS > 3 (QRCD linked with hospital)
- Police reported serious injuries as defined by SRR < .942 (QRCD linked with hospital)
- Hospital serious injuries as defined by AIS > 3 (QHAPDC, EDIS, QISU)
- Hospital serious injuries as defined by SRR < .942 (QHAPDC, EDIS, QISU)

7.3.5.6 *Profiling crash and injured person characteristics*

A profile of road crash injuries based on: gender; age; road user; and ARIA+; was produced and compared for the following data sets:

- All QRCD (linked and non-linked)
- ‘Hospitalised’ QRCD (linked and non-linked)
- *Hospital data* collections (QHAPDC, EDIS, and QISU; linked and non-linked)
- All health data collections (QHAPDC, EDIS, QISU, and eARF; linked and non-linked)
- Linked QRCD cases

7.3.5.7 *Assessing validity*

Firstly, the validity of broad injury severity coding (i.e., ‘hospitalised’, other injury) in QRCD was examined with the linkage rate for each severity level with the *hospital data* (i.e., QHAPDC, EDIS, and QISU). It was assumed that the *hospital data* was the best reference for assessing the validity of the ‘taken to hospital’ definition. The proportion of ‘hospitalised’ cases that did not link to a hospital data collection was considered false positives and ‘other injury’ cases that did link with a *hospital data* collection were considered false negatives. The severity coding based on AIS and SRR was also examined for cases that linked with the *hospital data* by exploring the concordance of the specified serious AIS and SRR cases with the *hospital data* serious AIS and SRR.

For the health data collections, the coding of a road crash was compared based on linkage with QRCD. Specifically, the proportion of cases that were not coded as a road crash in each health collection that did link with QRCD were considered false negatives and the cases that were coded as a road crash in the other data collections that did not link with QRCD were considered false positives. It should be noted however, that the false positives also included those that were not reported to police and therefore would represent a potential overestimation of the false positive rate.

The validity of road user coding for the *health* data sets was also examined by comparing the linked cases with the coding in QRCD. It was assumed that the coding in QRCD was the 'gold standard'. For each road user type, the proportions of correct and incorrect cases were produced. Also, sensitivity and specificity of the coding in each collection, for each road user, was calculated using the method described in Section 5.3.3.3 and based on the following (Table 7.2) characterisation of false positives and false negatives.

Table 7.2: Characterisation of true positives, false negatives, false positives, and true negatives for road user classification for health data sets

Road user	True positives	False negatives	False positives	True negatives
Driver	Driver in health, driver in QRCD	Not driver in health, driver in QRCD	Driver in health, not driver in QRCD	Not driver in health, not driver in QRCD
Motorcyclists	Motorcyclists in health, motorcyclists in QRCD	Not motorcyclists in health, motorcyclists in QRCD	Motorcyclists in health, not motorcyclists in QRCD	Not motorcyclists in health, not motorcyclists in QRCD
Cyclists	Cyclists in health, cyclists in QRCD	Not cyclists in health, cyclists in QRCD	Cyclists in health, not cyclists in QRCD	Not cyclists in health, not cyclists in QRCD
Pedestrian	Pedestrian in health, pedestrian in QRCD	Not pedestrian in health, pedestrian in QRCD	Pedestrian in health, not pedestrian in QRCD	Not pedestrian in health, not pedestrian in QRCD
Passenger	Passenger in health, passenger in QRCD	Not passenger in health, passenger in QRCD	Passenger in health, not passenger in QRCD	Not Passenger in health, not passenger in QRCD

Using the comparison to QRCD to examine the validity of *health data* sets could only be conducted for those cases that linked with QRCD. To examine the validity of both linked and non-linked cases, convergent validity was also explored. The commonalities between the *health* data sets for defining a road crash were examined. This was done using each combination of linkage among the health data collections. The greater the number of data collections with common coding (among the data sets that linked), the higher the convergent validity for a case was considered to be. It should be noted that there will still be some cases that are entirely unique to a data set, so any validity assessments can only be indicative.

7.4 Results

7.4.1 *Linkage rates*

7.4.1.1 *QRCD and QHAPDC*

There were 19,041 road crash casualties in QRCD in 2009. Of these, 4,283 linked to a case in QHAPDC with a linkage rate of 22.5%. Once the coding of ‘hospitalised’ is taken into account, there were 6,674 coded ‘hospitalised’ cases in QRCD, of which 3,264 linked to a case in QHAPDC representing a linkage rate of 48.9%. There were 997 QRCD cases that linked with QHAPDC that were not coded as ‘hospitalised’ in QRCD. These cases are discussed in more detail in Section 7.4.7.

7.4.1.2 *QRCD and EDIS*

Of the road crash cases in QRCD, 9,579 linked to a case in EDIS representing a linkage rate of 50.3%. Once the coding of ‘hospitalised’ was taken into account, there were 6,674 coded ‘hospitalised’ cases in QRCD, of which 4,869 linked to a case in EDIS representing a linkage rate of 73.0%. There were 4,637 QRCD cases that linked with EDIS that were not coded as ‘hospitalised’ in QRCD. These cases are discussed in more detail in Section 7.4.7.

7.4.1.3 *QRCD and QISU*

Of the road crash cases in QRCD, 971 linked to a case in QISU representing a linkage rate of 5.1%. Once the coding of ‘hospitalised’ was taken into account, there were 6,674 coded ‘hospitalised’ cases in QRCD, of which 505 linked to a case in QISU representing a linkage rate of 7.6%. There were 457 QRCD cases that linked with QISU that were not coded as ‘hospitalised’ in QRCD. These cases are discussed in more detail in Section 7.4.7.

7.4.1.4 *QRCD and eARF*

Of the road crashes in QRCD, 11,579 linked to a case in eARF representing a linkage rate of 60.8%. Once the coding of ‘hospitalised’ and medical treatment was taken into account there were 14, 636 case in QRCD, of which 10,351 linked with eARF (70.7% linkage rate). There were also 1,066 QRCD cases that were not coded as medically treated or ‘hospitalised’ that linked with eARF. The linkage numbers and rates for each severity level are shown in Table 7.3.

Table 7.3: Number of QRCD cases that linked to eARF for each QRCD severity level

QRCD Severity	Number of cases in QRCD	Number of linked cases	Linkage rate
Hospitalised	6,674	5,642	84.5%
Medically treated	7,962	4,709	59.1%
Minor injury	4,074	1,066	26.2%

7.4.1.5 *QRCD and other combinations*

Table 7.4 outlines the linkage rates for each combination of linkage where cases needed to link with every data collection in the combination to be considered a link. Not surprisingly, as the number of data collections increase, the linkage rate reduces. This is particularly noticeable with the addition of QISU.

Table 7.4: Number of QRCD cases linked and related linkage rates with all data collections in each combination

Data collection combinations	Number linked	Linkage rate
QHAPDC	4,283	22.5%
EDIS	9,579	50.3%
QISU	971	5.1%
eARF	11,579	60.8%
QHAPDC **and** EDIS	3,672	19.3%
QHAPDC **and** QISU	622	3.3%
QHAPDC **and** eARF	3,922	20.6%
EDIS **and** QISU	1,043	5.5%
EDIS **and** eARF	8,060	42.3%
QISU **and** eARF	1,038	5.5%
QHAPDC, EDIS **and** QISU	294	1.5%
QHAPDC, EDIS **and** eARF	2,884	15.1%
QHAPDC, QISU **and** eARF	319	1.6%
EDIS, QISU, **and** eARF	649	3.4%
QHAPDC, EDIS, QISU, **and** eARF	253	1.3%

Table 7.5 includes the linkage rates for each combination of links with QRCD, where a case need only link to one of the other data collections in the combination. The one-to-one linkage rates (e.g., QRCD and QHAPDC) have already been reported, however they are included here for comparison purposes. Not surprisingly, the maximum number of cases linked to another data collection was achieved by linking QRCD with all other data collections. However, it should be noted that only very few extra links were provided by including QISU (50 cases). For QRCD and any *hospital data* collection, 55.9% linked. These linkage rates increased when only police-reported ‘hospitalised’ cases were considered.

Table 7.5: Number of QRCD cases linked with any data collection in each combination

Data collection combinations	Number linked	Linkage rate (all)	Linkage rate ('hospitalised'[1])
QHAPDC	4,283	22.5%	48.9%
EDIS	9,579	50.3%	73.0%
QISU	971	5.1%	7.6%
eARF	11,579	60.8%	84.5%
QHAPDC **or** EDIS	10,543	55.4%	82.9%
QHAPDC **or** QISU	4,885	25.7%	51.9%
QHAPDC **or** eARF	12,193	64.0%	91.1%
EDIS **or** QISU	9,760	51.3%	74.3%
EDIS **or** eARF	13,351	70.1%	93.0%
QISU **or** eARF	11,765	61.8%	85.5%
QHAPDC, EDIS **or** QISU	10,649	55.9%	83.3%
QHAPDC, EDIS **or** eARF	13,530	71.1%	94.7%
QHAPDC, QISU **or** eARF	12,329	64.7%	91.4%
EDIS, QISU, **or** eARF	13,396	70.4%	93.1%
QHAPDC, EDIS, QISU, **or** eARF	13,566	71.2%	94.8%

[1] 'Hospitalised' refers the police-reported 'taken to hospital'

7.4.2 *Discordance rates*

7.4.2.1 *QHAPDC and QRCD*

There were 12,198 transport cases in QHAPDC in 2009. Of these, 7,396 (63.3%) cases did not link to QRCD. Once the coding of traffic was taken into account, there were 7,278 coded traffic injury cases in QHAPDC of which 3,320 did not link to QRCD (45.6%). These non-linked cases represent possible under-reporting to police and highlight the number of additional cases that linking QRCD with QHAPDC could provide. It should be noted that 329 cases in QHAPDC that were coded as non-traffic did actually link to QRCD (representing 8% of all linked cases). These cases are discussed in more detail in Section 7.4.7.

7.4.2.2 *EDIS and QRCD*

There were 303,870 injury cases in EDIS in 2009. Of these, 294,297 (96.8%) cases did not link to QRCD. Once the coding of road crash injuries in EDIS was taken into account, there were 23,624 coded road crash injury cases in EDIS of which 16,580 did not link to QRCD (70.2%). These non-linked cases represent possible under-reporting to police and the number of additional cases that linking QRCD with EDIS could potentially provide. It should be noted that 2,531 cases in EDIS that were coded as non-road crash did link to QRCD (representing 26% of all linked cases). These cases are discussed in more detail in Section 7.4.7.

7.4.2.3 *QISU and QRCD*

There were 4,620 transport injury cases in QISU in 2009. Of these, 3,661 (79.2%) cases did not link to QRCD. Once the coding of road crash was taken into account, there were 2,478 coded road crash injury cases in QISU of which 1,579 did not link to QRCD (63.7%). Once again, these non-linked cases represent the possible under-reporting to police and the number of additional cases that linking QRCD with QISU could provide. It should be noted that 72 cases in QISU that were coded as non-road crash did link to QRCD (representing 8% of all linked cases). These cases are discussed in more detail in Section 7.4.7.

7.4.2.4 *eARF and QRCD*

There were 15,962 transport injury cases in eARF in 2009. Of these, 8,979 (56.3%) did not link with QRCD. Once the coding of a road crash was taken into account, there were 11,613 cases in eARF of which 5,962 (51.3%) did not link to QRCD. Interestingly, 1,435 cases not coded as a road crash in eARF linked with a case in QRCD (representing 20% of all linked cases). These cases will be discussed in more detail in Section 7.4.7.

7.4.2.5 *Other combinations and QRCD*

Table 7.6 includes the population numbers and discordance rates for each combination of links with QRCD. The one-to-one discordance rates (e.g., QRCD and QHAPDC) have already been reported above however they are included here for comparison purposes. For the entire road crash injury population, as measured by the combination of all *health* data collections, the discordance rate was 67.7%. For the *hospital data* collection population (QHAPDC, EDIS, and QISU), 68.6% did not have a QRCD case.

Table 7.6: Number of population sample set cases linked with QRCD

Population set	Number in population	Discordance rate
QHAPDC	7,278	45.6%
EDIS	23,624	70.2%
QISU	2,478	63.7%
eARF	11,613	51.3%
QHAPDC **and** EDIS	27,292	68.3%
QHAPDC **and** QISU	9,259	50.8%
QHAPDC **and** eARF	17,736	51.5%
EDIS **and** QISU	24,749	70.3%
EDIS **and** eARF	31,698	67.7%
QISU **and** eARF	13,902	54.3%
QHAPDC, EDIS **and** QISU	28,220	68.6%
QHAPDC, EDIS **and** eARF	34,742	67.3%
QHAPDC, QISU **and** eARF	19,330	53.8%
EDIS, QISU, **and** eARF	32,635	68.0%
QHAPDC, EDIS, QISU, **and** eARF	35,536	67.7%

7.4.3 *Linkage bias*

7.4.3.1 *QRCD and QHAPDC*

There was a statistically significant difference in the linkage rate between QRCD and QHAPDC based on road user for police-coded ‘hospitalised’ cases [$\chi^2(4) = 216.89$, $p < .001$, $\phi_c = .18$]. Specifically, drivers had a lower than expected linkage rate and motorcyclists and pedestrians had a higher than expected linkage rate (see Table 7.7).

Table 7.7: Linkage rates for police-coded ‘hospitalised’ QRCD cases and QHAPDC for different road users

Road user	Number of cases in QRCD	Number of linked cases	Linkage rate
Driver	3,573	1,524	**42.7%**
Motorcyclist	955	622	**65.1%**
Cyclist	354	182	51.4%
Pedestrian	424	286	**67.5%**
Passenger	1,365	648	47.5%

Note: Standardised residuals outside +/-3.10 are bolded

There was no statistically significant difference in linkage rate based on age [$\chi^2(18) = 32.80$, $p = .025$, $\phi_c = .07$] (see Figure 7.1).

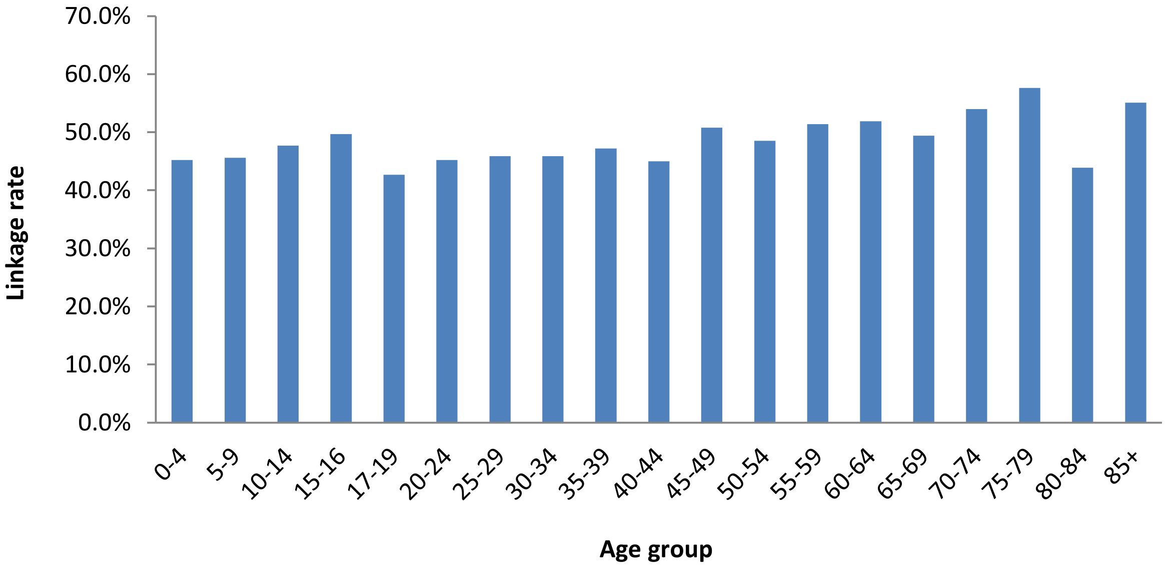


Figure 7.1: Linkage rates for police-coded ‘hospitalised’ QRCD cases and QHAPDC for different age groups

Linkage rates statistically significantly differed on the gender of the injured person for police-coded ‘hospitalised’ cases [$\chi^2(1) = 116.00$, $p < .001$, $\phi_c = .13$]. Specifically, males had a higher than expected linkage rate (54.7%) and females had a lower than expected linkage rate (41.3%).

There was also a statistically significant difference in linkage rates based on ARIA+ for police-coded 'hospitalised' cases [$\chi^2(4) = 103.01$, $p < .001$, $\phi_c = .12$]. Specifically, Major Cities had a lower than expected linkage rate and Remote and Very Remote had a higher than expected linkage rate (see Table 7.8).

Table 7.8: Linkage rates for police-coded 'hospitalised' QRCD cases and QHAPDC for different ARIA+

ARIA+	Number of cases in QRCD	Number of linked cases	Linkage rate
Major Cities	3,603	1,602	**44.5%**
Inner Regional	1,479	749	50.6%
Outer Regional	1,178	630	53.5%
Remote	217	146	**67.3%**
Very Remote	197	137	**69.5%**

Note: Standardised residuals outside +/-3.10 are bolded

While injury severity was not able to be determined in QRCD for a vast majority of cases, the linkage rates for those cases where it was possible were compared. As shown in Table 7.9, for both serious injury based on AIS [$\chi^2(1) = 59.03$, $p < .001$, $\phi_c = .24$] and SRR [$\chi^2(1) = 41.38$, $p < .001$, $\phi_c = .20$], serious police-coded 'hospitalised' cases had a higher than expected linkage rate.

Table 7.9: Linkage rates for police-coded 'hospitalised' QRCD cases and QHAPDC for different serious injury levels

Severity	Seriousness	Number of cases in QRCD	Number of linked cases	Linkage rate
AIS	Non-serious (< 3)	894	417	**46.6%**
	Serious (> 2)	110	94	**85.5%**
SRR	Non-serious (> .941)	837	387	**46.2%**
	Serious (< .942)	161	119	**73.9%**

Note: Standardised residuals outside +/-3.10 are bolded

In order to take into account potential confounding factors, a logistic regression was performed. With all variables in the logistic regression, the model was statistically significant, $\chi^2(13) = 114.66$, $p < .001$ (Nagelkerke $R^2 = .15$). After controlling for the relationships between the predictors, gender, age, and ARIA+ were no longer significant. In contrast, road user and serious injury remained statistically significant. Specifically, motorcyclist and pedestrian police-coded 'hospitalised' cases in QRCD had higher odds of linking to QHAPDC (2.8 and 3.4 times respectively). Also, serious police-coded 'hospitalised' cases in QRCD had higher odds (3.3 times) of linking to QHAPDC compared to non-serious police-coded 'hospitalised' cases (see Table 7.10).

Table 7.10: Logistic regression analysis of the profile of police-coded 'hospitalised' road crash injuries in QRCD that linked to QHAPDC

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.71	1.34	0.84 – 2.13	.038
Age	0 – 16	1.00	1.01	0.61 – 1.67	.973
	17 – 24	0.87	1.11	0.38 – 1.86	.297
	25 – 59	1.00	1.00	Referent	
	60 +	1.20	1.23	0.38 – 1.86	.127
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	2.51	2.83	1.32 – 6.06	< .001
	Cyclist	1.42	1.09	0.45 – 2.62	.750
	Pedestrian	2.79	3.44	1.41 – 8.44	< .001
	Passenger	1.22	0.96	0.53 – 1.75	.838
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.28	1.73	0.98 – 3.05	.002
	Outer Regional	1.44	1.85	0.99 – 3.45	.002
	Remote	2.57	2.12	0.69 – 6.44	.027
	Very Remote	2.85	3.58	0.99 – 12.29	.002
SRR Serious	Non-serious	1.00	1.00	Referent	
	Serious	3.30	3.27	1.70 – 6.29	< .001

[1] Adjusted for all variables in the equation

7.4.3.2 *QRCD and EDIS*

There was a statistically significant difference in the linkage rate between QRCD and EDIS based on road user for police-coded 'hospitalised' cases [$\chi^2(4) = 129.35$, $p < .001$, $\phi_c = .14$]. Specifically, motorcyclists had a higher than expected linkage rate (see Table 7.11).

Table 7.11: Linkage rates for police-coded 'hospitalised' QRCD cases and EDIS for different road users

Road user	Number of cases in QRCD	Number of linked cases	Linkage rate
Driver	3,573	2,534	70.9%
Motorcyclist	955	822	**86.1%**
Cyclist	354	270	76.3%
Pedestrian	424	334	78.8%
Passenger	1,365	907	66.4%

Note: Standardised residuals outside +/-3.10 are bolded

There was a statistically significant difference in linkage rate based on age [$\chi^2(18) = 115.77$, $p < .001$, $\phi_c = .13$] for police-coded 'hospitalised' cases. Specifically, those aged 0-4 had a lower than expected linkage rate (see Figure 7.2).

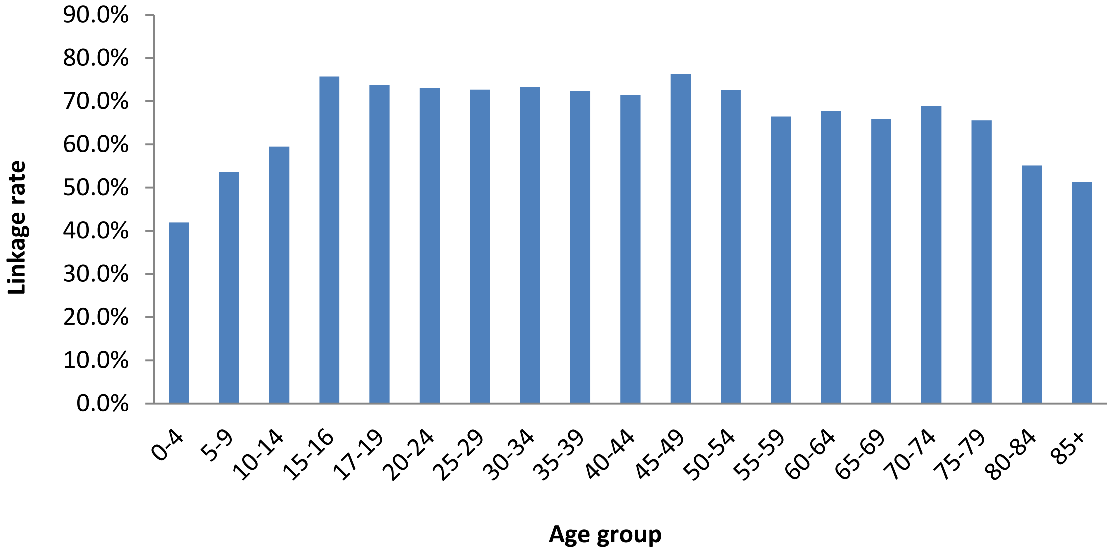


Figure 7.2: Linkage rates for police-coded 'hospitalised' cases and EDIS for different age groups

Linkage rates did not statistically significantly differ in terms of the gender of the injured person [$\chi^2(1) = 5.86$, $p = .015$, $\phi_c = .03$] (75.0% male and 70.4% female).

There was also a statistically significant difference in linkage rates based on ARIA+ for police-coded 'hospitalised' cases [$\chi^2(4) = 146.07$, $p < .001$, $\phi_c = .15$]. Specifically, Very Remote had a lower than expected linkage rate (see Table 7.12).

Table 7.12: Linkage rates for police-coded 'hospitalised' QRCD cases and EDIS for different ARIA+

ARIA+	Number of cases in QRCD	Number of linked cases	Linkage rate
Major Cities	3,603	2,722	75.5%
Inner Regional	1,479	1,153	78.0%
Outer Regional	1,178	773	65.6%
Remote	217	123	56.7%
Very Remote	197	98	**49.7%**

Note: Standardised residuals outside +/-3.10 are bolded

As shown in Table 7.13, for serious injury based on AIS [$\chi^2(1) = 15.22$, $p < .001$, $\phi_c = .12$] serious cases had a higher than expected linkage rate for police-coded 'hospitalised' cases. However, for serious classification based on SRR, there was no statistically significant differences [$\chi^2(1) = 4.79$, $p = .029$, $\phi_c = .07$].

Table 7.13: Linkage rates for police-coded 'hospitalised' QRCD cases and EDIS for different serious injury levels

Severity		Number of cases in QRCD	Number of linked cases	Linkage rate
AIS	Non-serious (< 3)	894	631	**70.6%**
	Serious (> 2)	110	97	**88.2%**
SRR	Non-serious (> .941)	595	242	71.1%
	Serious (< .942)	128	33	79.5%

Note: Standardised residuals outside +/-3.10 are bolded

In order to take into account potential confounding factors, a logistic regression was performed. With all variables in the logistic regression, the model was statistically significant, $\chi^2(13) = 73.52$, $p < .001$ (Nagelkerke $R^2 = .10$). After controlling for the relationships between the predictors, only serious injury remained statistically significant. Specifically, serious police-coded 'hospitalised' cases in QRCD had higher odds (3.2 times) of linking to EDIS compared to non-serious police-coded 'hospitalised' cases (see Table 7.14).

Table 7.14: Logistic regression analysis of the profile of police-coded 'hospitalised' road crash injuries in QRCD that linked to EDIS

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.26	1.04	0.63 – 1.71	.819
Age	0 – 16	0.57	0.66	0.28 – 1.49	.092
	17 – 24	1.02	1.28	0.70 – 2.35	.005
	25 – 59	1.00	1.00	Referent	
	60 +	0.67	0.70	0.33 – 1.49	.120
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	2.53	2.66	0.93 – 7.61	.002
	Cyclist	1.32	0.93	0.36 – 2.41	.794
	Pedestrian	1.52	1.00	0.40 – 2.53	.989
	Passenger	0.81	0.75	0.41 – 1.40	.132
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.15	1.16	0.60 – 2.23	.452
	Outer Regional	0.62	0.70	0.36 – 1.36	.078
	Remote	0.42	0.43	0.14 – 1.30	.012
	Very Remote	0.32	0.32	0.10 – 1.05	.002
SRR Serious	Non-serious	1.00	1.00	Referent	
	Serious	3.11	3.16	1.13 – 8.88	< .001

[1] Adjusted for all variables in the equation

7.4.3.3 *QRCD and QISU*

There was no statistically significant difference in the linkage rate between QRCD and QISU based on road user for police-coded ‘hospitalised’ QRCD cases [χ2(4) = 5.64, p = .228, ϕc = .03].

Table 7.15: Linkage rates for police-coded ‘hospitalised’ QRCD cases and QISU for different road users

Road user	Number of cases in QRCD	Number of linked cases	Linkage rate
Driver	3,573	286	8.0%
Motorcyclist	955	55	5.8%
Cyclist	354	28	7.9%
Pedestrian	424	34	8.0%
Passenger	1,365	102	7.5%

Note: Standardised residuals outside +/-3.10 are bolded

There was, however, a statistically significant difference in linkage rate based on age for police-coded ‘hospitalised’ cases [$\chi^2(18) = 103.17$, $p < .001$, $\phi_c = .13$]. Specifically, those aged 0-14 had a higher than expected linkage rate (see Figure 7.3).

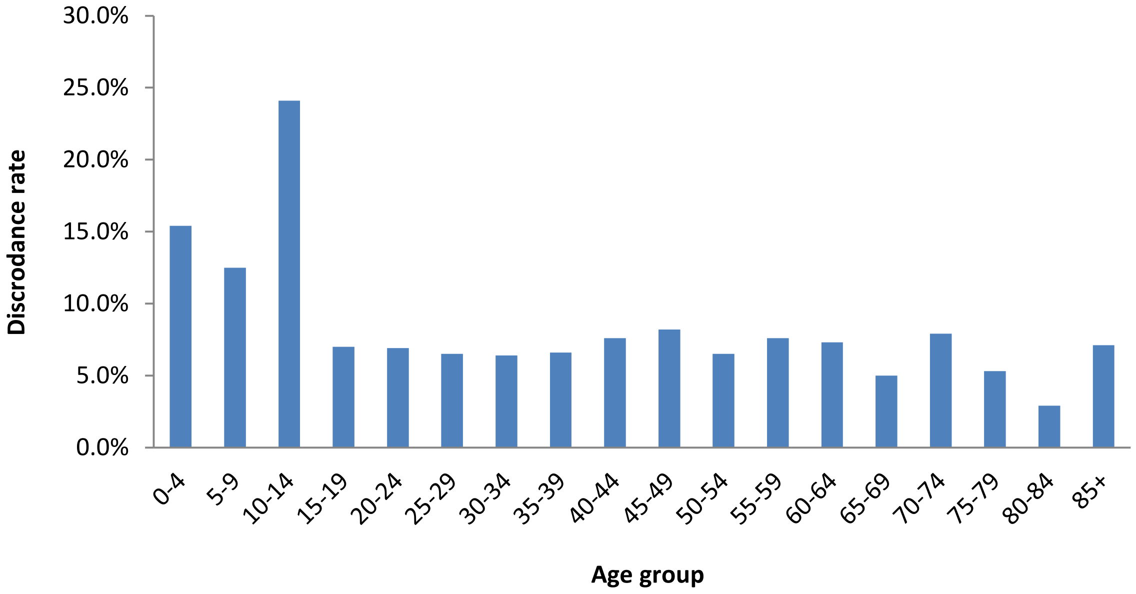


Figure 7.3: Linkage rates for police-coded ‘hospitalised’ QRCD cases and QISU for different age groups

Linkage rates did not statistically significantly differ in terms of the gender of the injured person [$\chi^2(1) = 0.25$, $p = .618$, $\phi_c = .006$] (7.7% male and 7.4% female).

There was a statistically significant difference in linkage rates based on ARIA+ for police-coded ‘hospitalised’ cases [$\chi^2(4) = 86.39$, $p < .001$, $\phi_c = .11$]. Specifically, Outer Regional, Remote, and Very Remote areas had a higher than expected linkage rate (Table 7.16).

Table 7.16: Linkage rates for police-coded 'hospitalised' QRCD cases and QISU for different ARIA+

ARIA+	Number of cases in QRCD	Number of linked cases	Linkage rate
Major Cities	3,603	200	5.6%
Inner Regional	1,479	107	7.2%
Outer Regional	1,178	131	**11.1%**
Remote	217	35	**16.1%**
Very Remote	197	32	**16.2%**

Note: Standardised residuals outside +/-3.10 are bolded

There were no statistically significant differences in linkage rate for serious injury classification based on AIS [$\chi^2(1) = 0.03$, $p = .868$, $\phi_c = .005$] or SRR [$\chi^2(1) = 0.54$, $p = .463$, $\phi_c = .02$].

Table 7.17: Linkage rates for police-coded 'hospitalised' QRCD cases and QISU for different serious injury levels

Severity		Number of cases in QRCD	Number of linked cases	Linkage rate
AIS	Non-serious (< 3)	894	69	**7.7%**
	Serious (> 2)	110	8	7.3%
SRR	Non-serious (> .941)	837	66	7.9%
	Serious (< .942)	161	10	6.2%

Note: Standardised residuals outside +/-3.10 are bolded

A logistic regression was not performed as there were considered too few significant differences to warrant multivariate analysis.

7.4.3.4 *QRCD and eARF*

There was a statistically significant difference in the linkage rate between QRCD and eARF based on road user for police-coded 'hospitalised' and medically treated cases [$\chi^2(4) = 209.23$, $p < .001$, $\phi_c = .12$]. Specifically, motorcyclists had a higher than expected linkage rate and passengers had a lower than expected linkage rate (see Table 7.18).

Table 7.18: Linkage rates for police-coded 'hospitalised' and medically treated QRCD cases and eARF for different road users

Road user	Number of cases in QRCD	Number of linked cases	Linkage rate
Driver	8,359	5,928	70.9%
Motorcyclist	1,486	1,257	**84.6%**
Cyclist	694	478	68.9%
Pedestrian	719	519	72.2%
Passenger	3,375	2,167	**64.2%**

Note: Standardised residuals outside +/-3.10 are bolded

There was also a statistically significant difference in linkage rate based on age for police-coded 'hospitalised' and medically treated cases [$\chi^2(18) = 125.31$, $p < .001$, $\phi_c = .09$]. Specifically, those aged 0-4 had a lower than expected linkage rate (see Figure 7.4). However, it should be noted that the effect size was small (< .1).

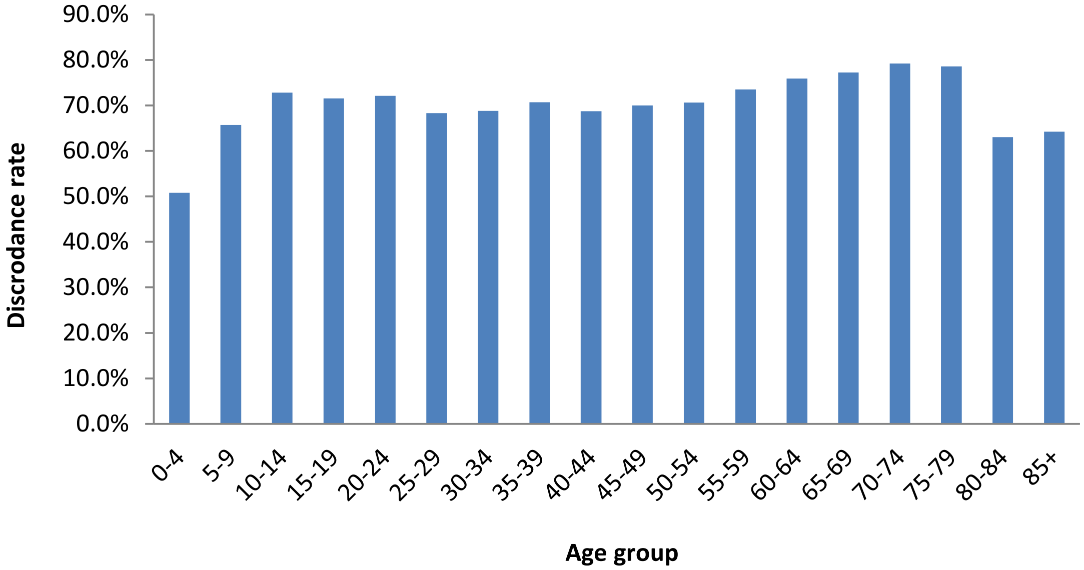


Figure 7.4: Linkage rates for police-coded 'hospitalised' and medically treated QRCD cases and eARF for different age groups

Linkage rates statistically significantly differed in terms of the gender of the injured person for police-coded 'hospitalised' and medically treated cases [$\chi^2(1) = 12.70$, $p < .001$, $\phi_c = .03$] (72.2% male and 69.6% female). However, the effect size associated with this difference was very small.

There was a statistically significant difference in linkage rates based on ARIA+ for police-coded 'hospitalised' and medically treated cases [$\chi^2(4) = 254.68$, $p < .001$, $\phi_c = .13$]. Specifically, Inner Regional and Outer Regional areas had a higher than expected linkage rate and Major Cities and Very Remote areas had a lower than expected linkage rate (see Table 7.19).

Table 7.19: Linkage rates for police-coded 'hospitalised' and medically treated QRCD cases and eARF for different ARIA+

ARIA+	Number of cases in QRCD	Number of linked cases	Linkage rate
Major Cities	8,550	5,705	**66.7%**
Inner Regional	3,007	2,347	**78.1%**
Outer Regional	2,384	1,859	**78.0%**
Remote	414	296	71.5%
Very Remote	278	143	**51.4%**

Note: Standardised residuals outside +/-3.10 are bolded

There were statistically significant differences in the linkage rate for serious injury classification based on AIS [$\chi^2(1) = 14.40$, $p < .001$, $\phi_c = .08$] in that serious injuries had a higher than expected linkage rate for police-coded 'hospitalised' and medically treated cases (see Table 7.20). There was no statistically significant difference, however, based on SRR [$\chi^2(1) = 3.33$, $p = .068$, $\phi_c = .04$].

Table 7.20: Linkage rates for police-coded 'hospitalised' and medically treated QRCD cases and eARF for different serious injury levels

Severity		Number of cases in QRCD	Number of linked cases	Linkage rate
AIS	Non-serious (< 3)	2,472	1,596	64.6%
	Serious (> 2)	119	97	**81.5%**
SRR	Non-serious (> .941)	2,299	1,488	64.7%
	Serious (< .942)	285	200	70.2%

Note: Standardised residuals outside +/-3.10 are bolded

In order to take into account potential confounding factors, a logistic regression was performed. With all variables in the logistic regression, the model was statistically significant, $\chi^2(14) = 507.21$, $p < .001$ (Nagelkerke $R^2 = .05$). After controlling for the relationships between the predictors, age, road user, and ARIA+ remained statistically significant. Those aged 17-24 had greater odds of linking to eARF compared to those aged 25-59. Motorcyclists had higher odds (3.4 times) of linking to eARF compared to drivers (cyclists were no longer significant). Also, those police-coded 'hospitalised' and medically treated QRCD cases in Inner and Outer Regional areas had higher odds (1.8 and 1.9 times respectively) of linking to eARF compared to Major Cities (see Table 7.21).

Table 7.21: Logistic regression analysis of the profile of police-coded 'hospitalised' and medically treated road crash injuries in QRCD that linked to eARF

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.14	0.86	0.64 – 1.15	.087
Age	0 – 16	0.84	1.25	0.72 – 2.18	.180
	17 – 24	1.39	1.64	1.04 – 2.59	< .001
	25 – 59	1.00	1.00	Referent	
	60 +	1.28	1.37	0.91 – 2.06	.012
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	2.25	3.40	1.70 – 6.79	< .001
	Cyclist	0.91	1.24	0.68 – 2.26	.238
	Pedestrian	1.06	1.28	0.68 – 2.42	.204
	Passenger	0.74	0.76	0.53 – 1.08	.011
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.77	1.77	1.19 – 2.65	< .001
	Outer Regional	1.77	1.94	1.26 – 2.97	< .001
	Remote	1.25	1.43	0.63 – 3.23	.150
	Very Remote	0.53	0.61	0.25 – 1.50	.070
AIS serious	Non-serious	1.00	1.00	Referent	
	Serious	1.28	1.88	0.83 – 4.23	.011

[1] Adjusted for all variables in the equation

7.4.3.5 *QRCD and hospital data*

The consistency of the linkage rate across the variables of interest was examined for QRCD police-coded 'hospitalised' cases compared to the combined *hospital data* (i.e., QHAPDC, EDIS, and QISU). There was a statistically significant difference in the linkage rate based on road user for police-coded 'hospitalised' cases [$\chi^2(4) = 75.18$, $p < .001$, $\phi_c = .11$]. Specifically, motorcyclists and pedestrians had a higher than expected linkage rate (see Table 7.22).

Table 7.22: Linkage rates for police-coded 'hospitalised' QRCD cases and hospital data for different road users

Road user	Number of cases in QRCD	Number of linked cases	Linkage rate
Driver	3,573	3,075	86.1%
Motorcyclist	955	906	**94.9%**
Cyclist	354	325	91.8%
Pedestrian	424	398	**93.9%**
Passenger	1,365	1,202	88.1%

Note: Standardised residuals outside +/-3.10 are bolded

There was no statistically significant difference in linkage rate based on age [$\chi^2(18) = 16.75$, $p = .540$, $\phi_c = .05$] (see Figure 7.5).

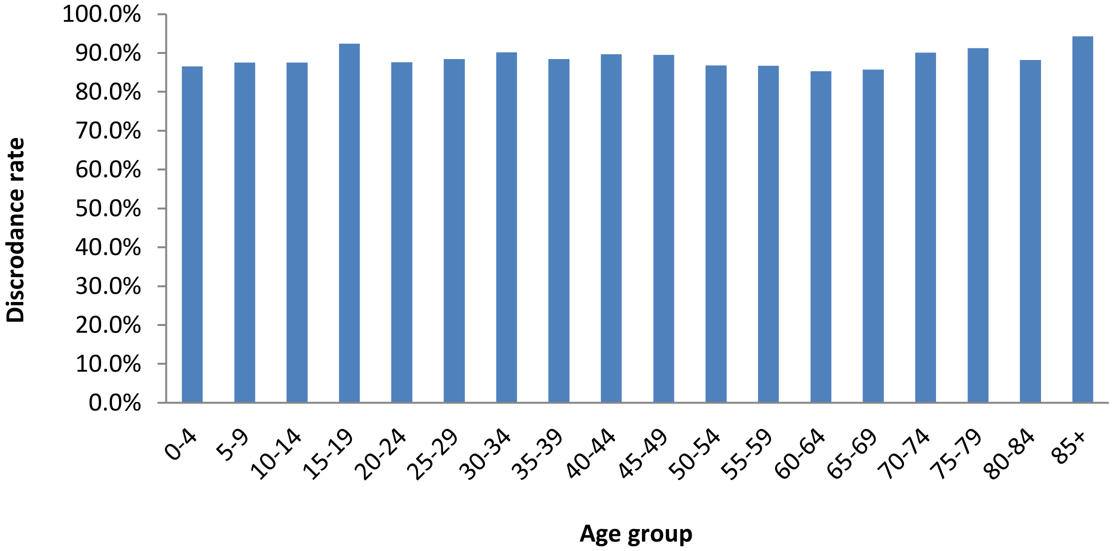


Figure 7.5: Linkage rates for police-coded 'hospitalised' QRCD cases and hospital data for different age groups

Linkage rates statistically significantly differed on the gender of the injured person for police-coded 'hospitalised' cases [$\chi^2(1) = 37.37$, $p < .001$, $\phi_c = .08$] (90.6% male and 85.8% female). However, the effect size associated with this difference was small.

There was no statistically significant difference in linkage rates based on ARIA+ [$\chi^2(4) = 8.63$, $p = .071$, $\phi_c = .04$] (see Table 7.23).

Table 7.23: Linkage rates for police-coded 'hospitalised' QRCD cases and hospital data for different ARIA+

ARIA+	Number of cases in QRCD	Number of linked cases	Linkage rate
Major Cities	3,603	3,163	87.8%
Inner Regional	1,479	1,331	90.0%
Outer Regional	1,178	1,037	88.0%
Remote	217	197	90.8%
Very Remote	197	181	91.9%

Note: Standardised residuals outside +/-3.10 are bolded

Serious cases based on AIS had a higher linkage rate than expected, however this difference was not statistically significant [$\chi^2(1) = 11.11$, $p = .002$, $\phi_c = .09$]. There was also no statistically significant difference for severity based on SRR [$\chi^2(1) = 2.70$, $p = .101$, $\phi_c = .06$].

Table 7.24: Linkage rates for police-coded 'hospitalised' QRCD cases and hospital data for different serious injury levels

Severity		Number of cases in QRCD	Number of linked cases	Linkage rate
AIS	Non-serious (< 3)	721	614	85.2%
	Serious (> 2)	91	89	97.8%
SRR	Non-serious (> .941)	683	585	85.7%
	Serious (< .942)	124	113	91.1%

Note: Standardised residuals outside +/-3.10 are bolded

In order to take into account potential confounding factors, a logistic regression was performed. With all variables in the logistic regression, the model was statistically significant, $\chi^2(14) = 118.61$, $p < .001$ (Nagelkerke $R^2 = .04$). After controlling for the relationships between the predictors, all predictors that were significant at the bivariate level remained significant. Specifically, police-coded 'hospitalised' male cases had higher odds (1.4 times) of linking compared to females and police-coded 'hospitalised' motorcyclists and pedestrians had higher odds (2.7 and 2.6 times respectively) of linking compared to drivers (see Table 7.25).

Table 7.25: Logistic regression analysis of the profile of police-coded 'hospitalised' road crash injuries in QRCD that linked to hospital data

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.60	1.39	1.06 – 1.81	< .001
Age	0 – 16	0.59	0.62	0.39 – 1.02	.002
	17 – 24	0.99	1.09	0.70 – 1.78	.408
	25 – 59	1.00	1.00	Referent	
	60 +	0.79	0.84	0.79 – 1.55	.046
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	2.99	2.69	1.60 – 4.53	< .001
	Cyclist	1.82	1.77	0.90 – 3.47	.005
	Pedestrian	2.48	2.65	1.31 – 5.34	< .001
	Passenger	1.19	1.24	0.87 – 1.76	.044
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.25	1.28	0.92 – 1.80	.014
	Outer Regional	1.02	1.04	0.73 – 1.46	.740
	Remote	1.37	1.50	0.67 – 3.31	.097
	Very Remote	1.57	1.78	0.72 – 4.41	.037

[1] Adjusted for all variables in the equation

7.4.4 *Discordance bias*

7.4.4.1 *QHAPDC and QRCD*

There was statistically significant difference in the discordance rate based on road user for traffic-coded QHAPDC cases [$\chi^2(4) = 1688.94$, $p < .001$, $\phi_c = .50$]. Specifically, motorcyclists and cyclists had a higher than expected discordance rate (see Table 7.26).

Table 7.26: Discordance rates for QRCD and traffic coded QHAPDC cases for different road users

Road user	Number of cases in QHAPDC	Number of non-linked cases	Discordance rate
Driver	2,081	350	**16.8%**
Motorcyclist	2,086	1,357	**65.1%**
Cyclist	1,096	881	**80.4%**
Pedestrian	455	136	**29.9%**
Passenger	1,089	338	**31.0%**

Note: Standardised residuals outside +/-3.10 are bolded

The discordance rate differed on the basis of whether the injury involved another vehicle for traffic-coded QHAPDC cases [$\chi^2(1) = 237.51$, $p < .001$, $\phi_c = .18$]. Specifically, those injuries that did not result from a collision with another vehicle had a higher discordance rate (see Table 7.27).

Table 7.27: Discordance rates for QRCD and traffic coded QHAPDC cases for collision

	Number of cases in QHAPDC	Number of non-linked cases	Discordance rate
Collision	4,979	1,901	**38.2%**
Non-collision	2,028	1,183	**58.3%**

Note: Standardised residuals outside +/-3.10 are bolded

This pattern was consistent for drivers [$\chi^2(1) = 15.79$, $p < .001$, $\phi_c = .09$], motorcyclists [$\chi^2(1) = 26.54$, $p < .001$, $\phi_c = .12$], and cyclists [$\chi^2(1) = 119.01$, $p < .001$, $\phi_c = .34$]. The effect sizes indicate that the relationship between collision status and discordance was much higher for the cyclists and was relatively small for motorcyclists and drivers. There was no difference in discordance rates on the basis of another vehicle being involved for passenger injuries [$\chi^2(1) = 5.41$, $p = .020$, $\phi_c = .07$] (see Table 7.28).

Table 7.28: Discordance rates between QRCD and traffic coded QHAPDC cases for collision with different road user types

Road user	Collision	Number of cases in QHAPDC	Number of non-linked cases	Discordance rate
Driver	Yes	1,576	236	**15.0%**
	No	505	114	**22.6%**
Motorcyclist	Yes	1,273	758	**59.5%**
	No	738	524	**71.0%**
Cyclist	Yes	615	419	**68.1%**
	No	437	418	**95.7%**
Passenger	Yes	792	230	29.0%
	No	297	108	36.4%

Note: Pedestrians were not included in this table as by definition, all cases involve a collision with a vehicle

Standardised residuals outside +/-3.10 are bolded

There was also a statistically significant difference in discordance rate based on age for traffic-coded QHAPDC cases [$\chi^2(18) = 325.33$, $p < .001$, $\phi_c = .21$]. Specifically, those aged 16 years and younger had a higher than expected discordance rate (see Figure 7.6).

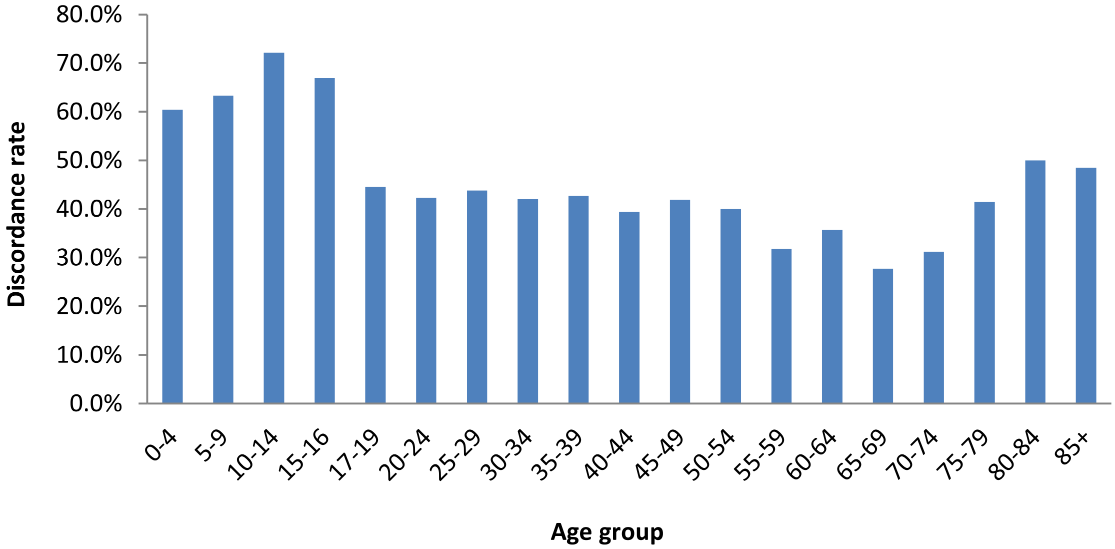


Figure 7.6: Discordance rates for QRCD and traffic coded QHAPDC cases for different age groups

Discordance rates statistically significantly differed on the gender of the injured person for traffic-coded QHAPDC cases [$\chi^2(1) = 159.24$, $p < .001$, $\phi_c = .15$]. Specifically, males had a higher than expected discordance rate (50.7%) and females had a lower than expected discordance rate (34.9%).

There was no statistically significant difference in discordance rates based on ARIA+ of the hospital [$\chi^2(4) = 13.57$, $p = .003$, $\phi_c = .04$].

Table 7.29: Discordance rates between QRCD and traffic coded QHAPDC cases for different ARIA+

ARIA+	Number of cases in QHAPDC	Number of non-linked cases	Discordance rate
Major Cities	3,971	1,755	44.2%
Inner Regional	1,937	935	48.3%
Outer Regional	1,179	529	44.9%
Remote	128	70	54.7%
Very Remote	63	31	49.2%

Note: Standardised residuals outside +/-3.10 are bolded

There was a statistically significant difference in discordance rates based on serious AIS classification for traffic-coded QHAPDC cases [$\chi^2(1) = 32.04$, $p < .001$, $\phi_c = .07$]. Specifically, non-serious cases had a higher than expected discordance rate (see Table 7.30). It should be noted however, that the associated effect size was small. In terms of SRR serious injury classification, there was also a difference in discordance rates for traffic-coded QHAPDC cases [$\chi^2(1) = 81.20$, $p < .001$, $\phi_c = .11$]. Specifically, non-serious cases had a higher than expected discordance rate (see Table 7.30).

Table 7.30: Discordance rates between QRCD and traffic coded QHAPDC cases for different severities

Severity		Number of cases in QHAPDC	Number of non-linked cases	Discordance rate
AIS	Serious (> 2)	776	300	**38.7%**
	Non-serious (< 3)	5,226	2,589	**49.5%**
SRR	Serious (< 0.942)	997	340	**34.1%**
	Non-serious (> 0.941)	6,022	2,980	**49.5%**

Note: Standardised residuals outside +/-3.10 are bolded

With all variables in the logistic regression, the model was statistically significant, $\chi^2(11) = 2073.61$, $p < .001$ (Nagelkerke $R^2 = .36$). After controlling for the relationships between the predictors gender was no longer significant. In contrast, age, road user and serious injury remained statistically significant. Specifically, those aged 0-16 and 17-24 had higher odds of discordance (2.0 and 1.6 times respectively) with QRCD compared to those aged 25-59. All non-driver road user cases in QHAPDC had higher odds of discordance with QRCD, particularly motorcyclists and cyclists (7.7 and 14.3 times respectively). Also, non-serious cases and non-collision cases in QHAPDC had higher odds (1.8 and 1.9 times respectively) of discordance with QRCD (see Table 7.31).

Table 7.31: Logistic regression analysis of the profile of road crash injuries in QHAPDC that did not link to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	0.52	0.97	0.78 – 1.22	.720
Age	0 – 16	2.78	1.99	1.42 – 2.81	< .001
	17 – 24	1.42	1.58	1.18 – 2.11	< .001
	25 – 59	1.00	1.00	Referent	
	60 +	0.69	0.85	0.63 – 1.13	.055
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	9.09	7.69	5.88 – 10.00	< .001
	Cyclist	20.00	14.29	10.00 – 20.00	< .001
	Pedestrian	2.13	2.04	1.35 – 3.13	< .001
	Passenger	2.22	1.79	1.32 – 2.43	< .001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.18	1.04	0.82 – 1.31	.603
	Outer Regional	1.03	1.02	0.77 – 1.33	.861
	Remote	1.52	1.21	0.57 – 2.57	.417
	Very Remote	1.22	1.67	0.46 – 6.07	.193
SRR Serious	Non-serious	1.00	1.00	Referent	
	Serious	0.53	0.56	0.42 – 0.75	< .001
Collision	No	1.00	1.00	Referent	
	Yes	0.44	0.53	0.43 – 0.67	< .001

[1] Adjusted for all variables in the equation

7.4.4.2 *EDIS and QRCD*

There was statistically significant difference in the discordance rate based on road user for road crash-coded EDIS cases [$\chi^2(4) = 3539.06$, $p < .001$, $\phi_c = .49$]. Specifically, motorcyclists and cyclists had a higher than expected discordance rate (see Table 7.32)

Table 7.32: Discordance rates between QRCD and road crash coded EDIS cases for different road users

Road user	Number of cases in EDIS	Number of non-linked cases	Discordance rate
Driver	2,618	1,013	**38.7%**
Motorcyclist	4,773	3,919	**82.1%**
Cyclist	5,396	5,022	**93.1%**
Pedestrian	183	68	**37.2%**
Passenger	1,893	956	**50.5%**

Note: Standardised residuals outside +/-3.10 are bolded

There was also a statistically significant difference in discordance rate based on age for road crash-coded EDIS cases [$\chi^2(18) = 1318.35$, $p < .001$, $\phi_c = .24$]. Specifically, those aged 16 years and younger had a higher than expected discordant rate (see Figure 7.7).

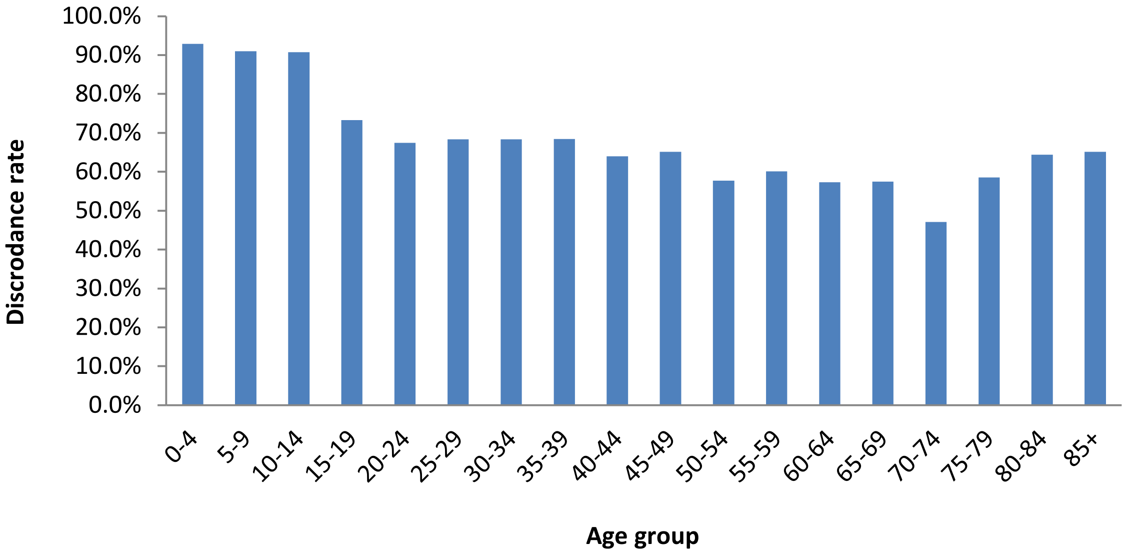


Figure 7.7: Discordance rates between QRCD and road crash coded EDIS cases for different age groups

Discordance rates statistically significantly differed on the gender of the injured person for road crash-coded EDIS cases [$\chi^2(1) = 603.86$, $p < .001$, $\phi_c = .16$]. Specifically, males had a higher than expected discordance rate (75.5%) and females had a lower than expected discordance rate (60.0%).

There was also a statistically significant difference in discordance rates based on ARIA+ of the hospital for road crash-coded EDIS cases [$\chi^2(4) = 245.24$, $p < .001$, $\phi_c = .10$]. Specifically, Inner Regional and Remote had a higher than expected discordance rate (see Table 7.33).

Table 7.33: Discordance rates between QRCD and road crash coded EDIS cases for different ARIA+

ARIA+	Number of cases in EDIS	Number of non-linked cases	Discordance rate
Major Cities	12,380	8,282	**66.9%**
Inner Regional	6,802	5,199	**76.4%**
Outer Regional	3,233	2,261	69.9%
Remote	496	420	**84.7%**
Very Remote	107	85	79.4%

Note: Standardised residuals outside +/-3.10 are bolded

There was a statistically significant difference in discordance rates based on AIS severity for road crash-coded EDIS cases [$\chi^2(4) = 66.97$, $p < .001$, $\phi_c = .06$]. Specifically, non-serious cases had a higher than expected discordance rate. In terms of SRR serious injury classification, there was also a difference in discordance rates for road crash-coded EDIS cases [$\chi^2(1) = 249.59$, $p < .001$, $\phi_c = .11$]. Specifically, non-serious cases had a higher than expected discordance rate (see Table 7.34).

Table 7.34: Discordance rates between QRCD and road crash coded EDIS cases for different severities

Severity		Number of cases in EDIS	Number of non-linked cases	Discordance rate
AIS	Serious (> 2)	637	350	**54.9%**
	Non-serious (< 3)	20,903	14,649	70.1%
SRR	Serious (< 0.942)	1,086	528	**48.6%**
	Non-serious (> 0.941)	20,982	14,927	71.1%

Note: Standardised residuals outside +/-3.10 are bolded

With all variables in the logistic regression, the model was statistically significant, $\chi^2(13) = 3815.66$, $p < .001$ (Nagelkerke $R^2 = .36$). After controlling for the relationships between the predictors gender was no longer significant. In contrast, age, road user, ARIA+, and serious injury remained statistically significant. Specifically, those aged 0-16 and 17-24 had higher odds of discordance with QRCD (3.2 and 1.8 times respectively) compared to those aged 25-59. All non-driver road users for road crash-coded EDIS cases, with the exception of pedestrians, had higher odds of discordance with QRCD, particularly motorcyclists and cyclists (6.7 and 16.7 times respectively). Non-serious road crash-coded EDIS had higher odds (2.3 times) of discordance with QRCD compared to serious cases. Finally, Inner Regional and Remote road crash-coded EDIS cases had higher odds of discordance compared to Major Cities (1.6 and 2.5 times respectively) (see Table 7.35).

Table 7.35: Logistic regression analysis of the profile of road crash injuries in EDIS that did not link to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	0.49	0.86	0.72 – 1.03	.006
Age	0 – 16	2.74	2.53	2.09 – 3.07	< .001
	17 – 24	1.16	1.47	1.21 – 1.79	< .001
	25 – 59	1.00	1.00	Referent	
	60 +	0.76	0.90	0.63 – 1.30	.342
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	7.14	6.67	5.56 – 8.33	< .001
	Cyclist	20.00	16.67	12.50 – 20.00	< .001
	Pedestrian	0.93	0.89	0.50 – 1.59	.506
	Passenger	1.61	1.32	1.04 – 1.64	< .001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.61	1.56	1.30 – 1.85	< .001
	Outer Regional	1.15	0.97	0.78 – 1.22	.679
	Remote	2.70	2.50	1.28 – 5.00	< .001
	Very Remote	1.92	2.17	0.57 – 8.33	.054
SRR Serious	Non-serious	1.00	1.00	Referent	
	Serious	0.38	0.44	0.29 – 0.67	< .001

[1] Adjusted for all variables in the equation

7.4.4.3 *QISU and QRCD*

There was statistically significant difference in the discordance rate based on road user for road crash-coded QISU cases [$\chi^2(4) = 443.16$, $p < .001$, $\phi_c = .43$]. Specifically, motorcyclists and cyclists had a higher than expected discordance rate (see Table 7.36).

Table 7.36: Discordance rates between QRCD and road crash coded QISU cases for different road users

Road user	Number of cases in QISU	Number of non-linked cases	Discordance rate
Driver	865	333	**38.5%**
Motorcyclist	448	358	**79.9%**
Cyclist	483	438	**90.7%**
Pedestrian	121	72	59.5%
Passenger	523	350	66.9%

Note: Standardised residuals outside +/-3.10 are bolded

The discordance rate differed on the basis of whether the injury involved collision with another vehicle for road crash-coded QISU cases [$\chi^2(1) = 118.64$, $p < .001$, $\phi_c = .22$].

Specifically, those injuries that did not result from a collision with another vehicle had a higher discordance rate (see Table 7.37).

Table 7.37: Discordance rates between QRCD and road crash coded QISU cases for collision

	Number of cases in QISU	Number of non-linked cases	Discordance rate
Collision	949	483	**50.9%**
Non-collision	1,448	1,053	**72.7%**

Note: Standardised residuals outside +/-3.10 are bolded

This pattern was consistent for motorcyclists [$\chi^2(1) = 42.65$, $p < .001$, $\phi_c = .31$], and cyclists [$\chi^2(1) = 121.01$, $p < .001$, $\phi_c = .50$]. There was no statistically significant difference in discordance rates on the basis of another vehicle being involved for driver [$\chi^2(1) = 3.53$, $p = .060$, $\phi_c = .07$] or passenger injuries [$\chi^2(1) = 1.24$, $p = .266$, $\phi_c = .05$] (see Table 7.38).

Table 7.38: Discordance rates between QRCD and road crash coded QISU cases for collision with different road user types

Road user	Collision	Number of cases in QISU	Number of non-linked cases	Discordance rate
Driver	Yes	410	145	41.7%
	No	422	176	35.4%
Motorcyclist	Yes	91	51	**56.0%**
	No	345	299	**86.7%**
Cyclist	Yes	68	37	**54.4%**
	No	408	394	**96.6%**
Passenger	Yes	269	188	69.9%
	No	227	148	65.2%

Note: Pedestrians were not included in this table as by definition, all cases involve a collision with a vehicle

Standardised residuals outside +/-3.10 are bolded

There was also a statistically significant difference in discordance rate based on age for road crash-coded QISU cases [$\chi^2(18) = 168.03$, $p < .001$, $\phi_c = .26$]. Specifically, those aged 16 years and younger had a higher than expected discordance rate (see Figure 7.8).

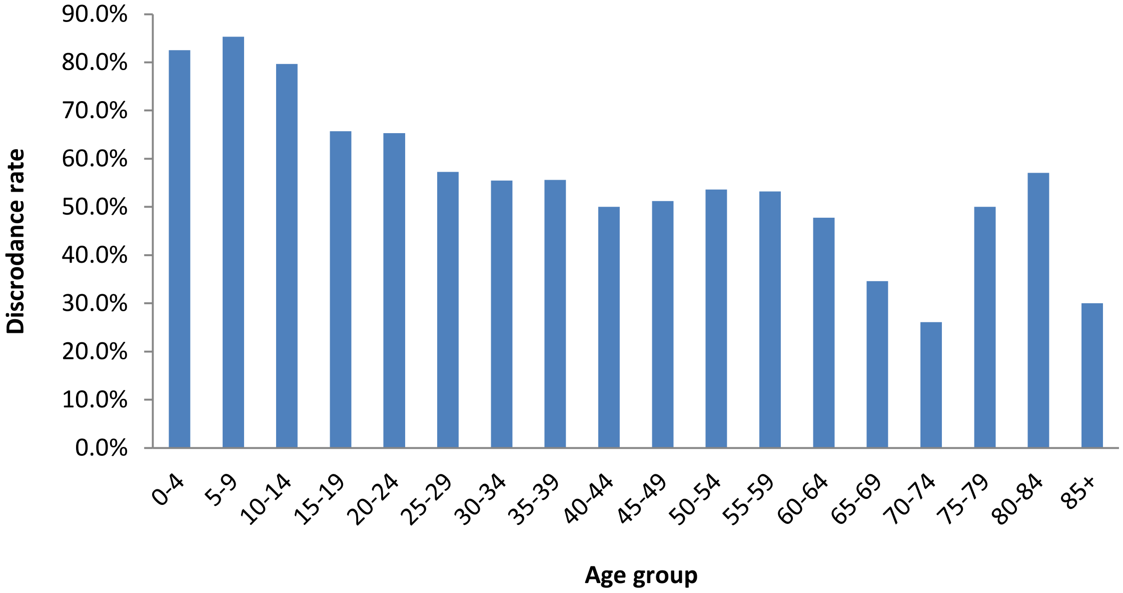


Figure 7.8: Discordance rates between QRCD and road crash coded QISU cases for different age groups

Discordance rates statistically significantly differed on the gender of the injured person for road crash-coded QISU cases [$\chi^2(1) = 26.01$, $p < .001$, $\phi_c = .10$]. Specifically, males had a higher than expected discordance rate (67.6%) and females had a lower than expected discordance rate (57.4%).

There was no statistically significant difference in discordance rates based on ARIA+ of the hospital [$\chi^2(4) = 14.25$, $p = .007$, $\phi_c = .08$].

Table 7.39: Discordance rates between QRCD and road crash coded QISU cases for different ARIA+

ARIA+	Number of cases in QISU	Number of non-linked cases	Discordance rate
Major Cities	1,254	818	65.2%
Inner Regional	555	334	60.2%
Outer Regional	425	257	60.5%
Remote	180	133	73.9%
Very Remote	20	13	65.0%

Note: Standardised residuals outside +/-3.10 are bolded

There was a statistically significant difference in discordance rates based on serious SRR classification for road crash-coded QISU cases [$\chi^2(1) = 34.90$, $p < .001$, $\phi_c = .12$]. Specifically, non-serious cases had a higher than expected discordance rate (see Table 7.40). In terms of AIS serious injury classification, there was no statistically significant difference in discordance rates [$\chi^2(1) = 0.02$, $p = .998$, $\phi_c = .001$].

Table 7.40: Discordance rates between QRCD and road crash coded QISU cases for different severities

Severity		Number of cases in QISU	Number of non-linked cases	Discordance rate
AIS	Serious (> 2)	74	47	63.5%
	Non-serious (< 3)	2,023	1,285	63.5%
SRR	Serious (< 0.942)	132	52	**39.4%**
	Non-serious (> 0.941)	2,246	1,457	**64.9%**

Note: Standardised residuals outside +/-3.10 are bolded

With all variables in the logistic regression, the model was statistically significant, $\chi^2(18) = 625.90$, $p < .001$ (Nagelkerke $R^2 = .34$). After controlling for the relationships between the predictors, gender and collision were no longer significant. In contrast, age, road user and serious injury remained statistically significant. Specifically, road crash-coded QISU cases aged 0-16 had 2.0 times higher odds of being discordant with QRCD compared to those aged 25-59. Road crash-coded QISU motorcyclist and cyclist cases had higher odds of discordance with QRCD (8.3 times and 25 times respectively). Also, non-serious road crash-coded QISU cases had higher odds (2.1 times) of discordance with QRCD (see Table 7.41).

Table 7.41: Logistic regression analysis of the profile of road crash injuries in QISU that did not link to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	0.64	1.03	0.72 – 1.47	.783
Age	0 – 16	2.78	1.82	1.22 – 2.70	< .001
	17 – 24	1.16	1.23	0.90 – 1.69	.025
	25 – 59	1.00	1.00	Referent	
	60 +	1.39	1.30	0.97 – 1.75	.004
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	6.25	8.33	4.35 – 16.67	< .001
	Cyclist	16.67	25.00	10.00 – 100.00	< .001
	Pedestrian	2.33	3.33	0.58 – 20.00	.023
	Passenger	3.23	1.85	0.98 – 3.45	.002
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	0.81	0.89	0.58 – 1.37	.391
	Outer Regional	0.81	0.85	0.53 – 1.37	.272
	Remote	1.52	1.59	0.78 – 3.23	.033
	Very Remote	0.99	1.52	0.25 – 9.09	.453
SRR Serious	Non-serious	1.00	1.00	Referent	
	Serious	0.35	0.47	0.47 – 0.98	< .001
Collision	No	1.00	1.00	Referent	
	Yes	2.57	1.29	0.47 – 1.27	.088

[1] Adjusted for all variables in the equation

7.4.4.4 *eARF and QRCD*

There was statistically significant difference in the discordance rate based on road user for road crash-coded eARF cases [$\chi^2(4) = 247.42$, $p < .001$, $\phi_c = .20$]. Specifically, cyclists had a higher than expected discordance rate (see Table 7.42).

Table 7.42: Discordance rates between QRCD and road crash coded eARF cases for different road users

Road user	Number of cases in eARF	Number of non-linked cases	Discordance rate
Driver	3,375	1,880	**44.3%**
Motorcyclist	659	309	53.1%
Cyclist	337	44	**86.9%**
Pedestrian	360	205	43.1%
Passenger	1,707	788	53.8%

Note: Standardised residuals outside +/-3.10 are bolded

There was also a statistically significant difference in discordance rate based on age for road crash-coded eARF cases [$\chi^2(18) = 252.18$, $p < .001$, $\phi_c = .15$]. Specifically, those aged 14 years and younger and those aged 80 and over had a higher than expected discordance rate (see Figure 7.9).

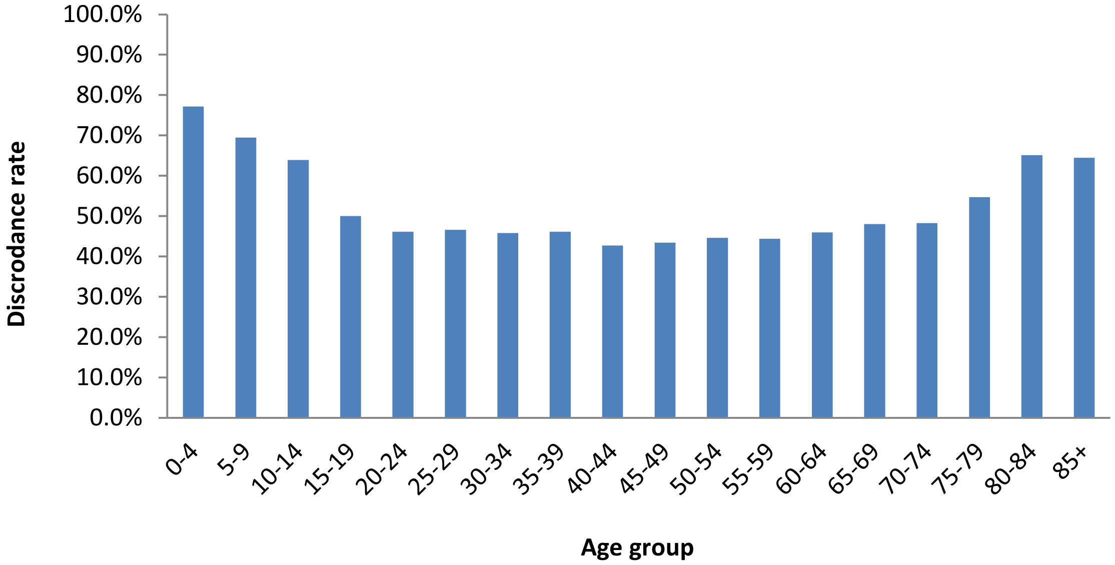


Figure 7.9: Discordance rates between QRCD and road crash coded eARF cases for different age groups

Discordance rates statistically significantly differed on the gender of the injured person for road crash-coded eARF cases [$\chi^2(1) = 25.89$, $p < .001$, $\phi_c = .05$]. Specifically, males had a higher than expected discordance rate (53.5%) and females had a lower than expected discordance rate (48.8%), although the effect size was small.

There was a statistically significant, but small effect on discordance rates based on ARIA+ for road crash-coded eARF cases [$\chi^2(4) = 24.10$, $p < .001$, $\phi_c = .05$], with Remote locations having a higher discordance rate (see Table 7.43).

Table 7.43: Discordance rates between QRCD and road crash coded eARF cases for different ARIA+

ARIA+	Number of cases in eARF	Number of non-linked cases	Discordance rate
Major Cities	5,991	3,038	**49.3%**
Inner Regional	3,268	1,519	53.5%
Outer Regional	2,415	1,124	53.5%
Remote	144	76	47.2%
Very Remote	112	63	43.8%

Note: Standardised residuals outside +/-3.10 are bolded

With all variables in the logistic regression, the model was statistically significant, $\chi^2(12) = 346.61$, $p < .001$ (Nagelkerke $R^2 = .08$). After controlling for the relationships between the predictors, gender and ARIA+ were no longer significant. In contrast age and road user and remained statistically significant. Specifically, road crash-coded eARF cases aged 0-16 had 1.5 times higher odds of discordance with QRCD compared to those aged 25-59. All road crash-coded non-driver road user cases in eARF, with the exception of pedestrians had higher odds of discordance with QRCD, particularly cyclists (7.5 times) (see Table 7.44).

Table 7.44: Logistic regression analysis of the profile of road crash injuries in eARF that did not link to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.21	1.16	0.97 – 1.39	.004
Age	0 – 16	1.77	1.48	0.34 – 0.73	< .001
	17 – 24	0.85	0.80	0.60 – 1.07	.013
	25 – 59	1.00	1.00	Referent	
	60 +	1.30	1.28	0.99 – 1.67	.002
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	1.42	1.36	1.01 – 1.82	< .001
	Cyclist	8.40	7.52	4.31 – 13.16	< .001
	Pedestrian	0.95	1.87	0.95 – 3.66	.002
	Passenger	1.46	1.33	1.08 – 1.65	< .001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.19	1.12	0.91 – 1.31	.069
	Outer Regional	1.18	1.13	0.89 – 1.43	.092
	Remote	0.92	1.54	0.66 – 3.57	.093
	Very Remote	0.80	0.64	0.22 – 1.81	.157

[1] Adjusted for all variables in the equation

7.4.4.5 *Hospital data and QRCD*

The consistency of the discordance rate was examined for QRCD cases and the combined *hospital data* (i.e., QHAPDC, EDIS, and QISU). There was statistically significant difference in the discordance rate based on road user for road crash-coded *hospital data* cases [$\chi^2(4) = 5686.25$, $p < .001$, $\phi_c = .52$]. Specifically, motorcyclists and cyclists had a higher than expected discordance rate (see Table 7.45).

Table 7.45: Discordance rates between QRCD and road crash coded hospital data cases for different road users

Road user	Number of cases in hospital data	Number of non-linked cases	Discordance rate
Driver	4,883	1,571	**32.2%**
Motorcyclist	6,169	5,010	**81.2%**
Cyclist	6,095	5,651	**92.7%**
Pedestrian	721	316	**43.8%**
Passenger	3,034	1,496	**49.3%**

Note: Standardised residuals outside +/-3.10 are bolded

There was also a statistically significant difference in discordance rate based on age for road crash-coded *hospital data* cases [$\chi^2(18) = 1800.32$, $p < .001$, $\phi_c = .25$]. Specifically, those aged 19 years and younger had a higher than expected discordance rate (see Figure 7.10).

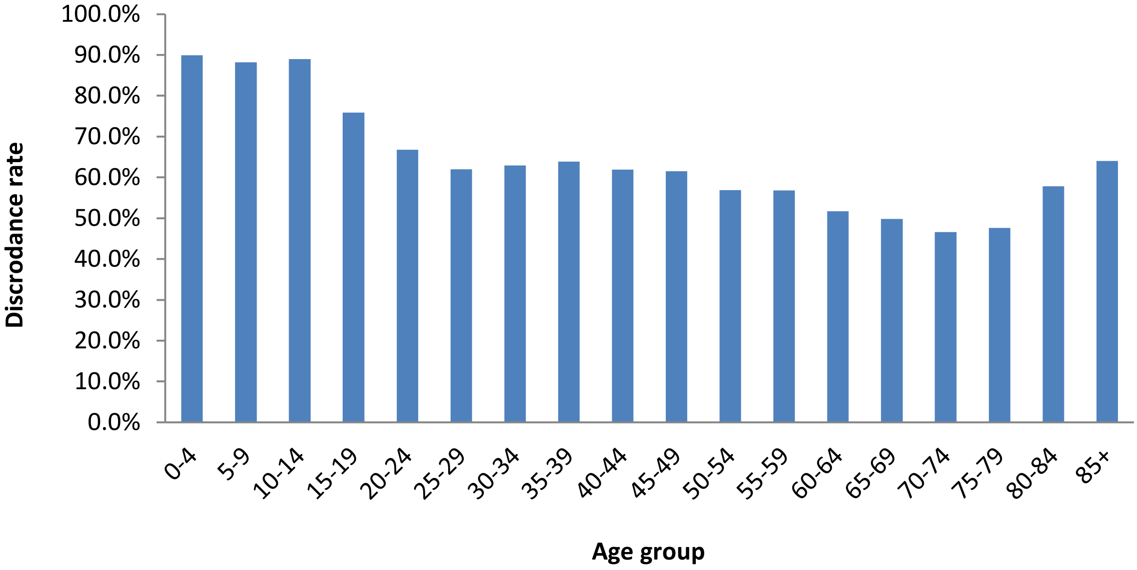


Figure 7.10: Discordance rates between QRCD and road crash coded hospital data cases for different age groups

Discordance rates statistically significantly differed on the gender of the injured person for road crash-coded *hospital data* cases [$\chi^2(1) = 725.02$, $p < .001$, $\phi_c = .16$]. Specifically, males had a higher than expected discordance rate (73.1%) and females had a lower than expected discordance rate (57.6%).

There was a statistically significant, but small difference in discordance rates based on ARIA+ for road crash-coded *hospital data* cases [$\chi^2(4) = 117.63$, $p < .001$, $\phi_c = .06$], with Inner Regional and Remote areas having a higher than expected discordance rate (see Table 7.46).

Table 7.46: Discordance rates between QRCD and road crash coded hospital data cases for different ARIA+

ARIA+	Number of cases in hospital data	Number of non-linked cases	Discordance rate
Major Cities	14,434	9,487	65.7%
Inner Regional	8,530	6,055	**71.0%**
Outer Regional	4,738	3,159	66.7%
Remote	680	544	**80.0%**
Very Remote	203	142	70.0%

Note: Standardised residuals outside +/-3.10 are bolded

There were statistically significant differences in discordance rates based on serious SRR classification for road crash-coded *hospital data* cases [$\chi^2(1) = 259.14$, $p < .001$, $\phi_c = .10$] and AIS serious injury classification for road crash-coded hospital cases [$\chi^2(1) = 133.70$, $p < .001$, $\phi_c = .07$]. Specifically, serious cases had a lower than expected discordance rate (see Table 7.47).

Table 7.47: Discordance rates between QRCD and road crash coded hospital cases for different severities

Severity		Number of cases in hospital	Number of non-linked cases	Discordance rate
AIS	Serious (> 2)	1,110	584	**52.6%**
	Non-serious (< 3)	24,647	17,003	69.1%
SRR	Serious (< 0.942)	1,507	732	**48.6%**
	Non-serious (> 0.941)	26,492	18,159	68.5%

Note: Standardised residuals outside +/-3.10 are bolded

With all variables in the logistic regression, the model was statistically significant, $\chi^2(13) = 6334.93$, $p < .001$ (Nagelkerke $R^2 = .39$). After controlling for the relationships between the predictors, gender was no longer significant. In contrast, age, road user, ARIA+, and serious injury remained statistically significant. Specifically, those aged 0-16 and 17-24 had higher odds (3.5 and 1.9 times respectively) and those aged 60+ had 1.4 times lower odds of being discordant with QRCD compared to those aged 25-59. Motorcyclist and cyclist cases in *hospital data* had higher odds of discordance with QRCD. Remote cases had higher odds of discordance compared to Major Cities. Also, non-serious cases and in *hospital data* had higher odds (2.2 times) of discordance with QRCD (see Table 7.48).

Table 7.48: Logistic regression analysis of the profile of road crash injuries in hospital data that did not link to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	2.00	1.12	0.98 – 1.29	.004
Age	0 – 16	5.03	3.45	2.79 – 4.27	< .001
	17 – 24	1.56	1.90	1.64 – 2.21	< .001
	25 – 59	1.00	1.00	Referent	
	60 +	0.60	0.74	0.60 – 0.91	< .001
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	9.11	7.57	6.43 – 8.90	< .001
	Cyclist	26.83	17.57	14.28 – 21.61	< .001
	Pedestrian	1.65	1.36	1.02 – 1.83	.002
	Passenger	2.05	1.53	1.28 – 1.82	< .001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.28	1.14	0.99 – 1.32	.003
	Outer Regional	1.04	0.96	0.81 – 1.14	.452
	Remote	2.09	2.13	1.41 – 3.22	< .001
	Very Remote	1.21	1.63	0.84 – 3.17	.016
SRR Serious	Non-serious	1.00	1.00	Referent	
	Serious	0.43	0.45	0.35 – 0.56	< .001

[1] Adjusted for all variables in the equation

7.4.4.6 *All health data collections and QRCD*

The consistency of the discordance rate was examined for QRCD cases and the combined *health data* (i.e., QHAPDC, EDIS, QISU, and eARF). There was statistically significant difference in the discordance rate based on road user for road crash-coded *health data* cases [$\chi^2(4) = 5358.52$, $p < .001$, $\phi_c = .46$]. Specifically, motorcyclists and cyclists had a higher than expected discordance rate (see Table 7.49).

Table 7.49: Discordance rates between QRCD and road crash coded health data for different road users

Road user	Number of cases in health	Number of non-linked cases	Discordance rate
Driver	7,642	2,943	**38.5%**
Motorcyclist	6,659	5,254	**78.9%**
Cyclist	6,235	5,749	**92.2%**
Pedestrian	961	448	46.6%
Passenger	4,292	2,249	52.4%

Note: Standardised residuals outside +/-3.10 are bolded

There was also a statistically significant difference in discordance rate based on age for road crash-coded *health data* cases [$\chi^2(18) = 1761.35$, $p < .001$, $\phi_c = .23$]. Specifically, those aged 19 years and younger had a higher than expected discordance rate (see Figure 7.11).

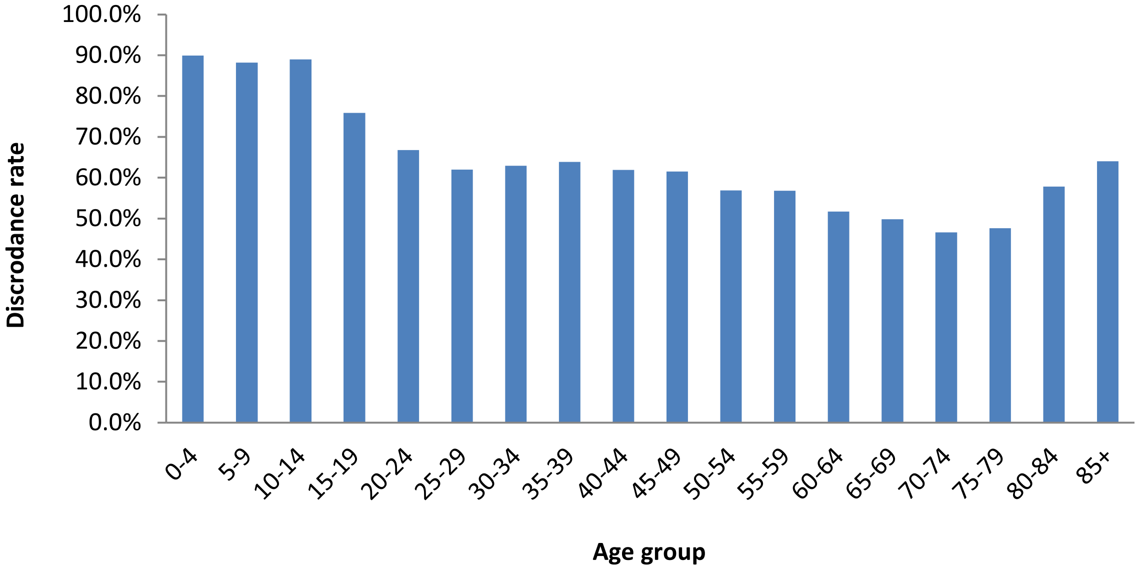


Figure 7.11: Discordance rates between QRCD and road crash coded health data for different age groups

Discordance rates statistically significantly differed on the gender of the injured person for road crash-coded *health data* cases [$\chi^2(1) = 633.66$, $p < .001$, $\phi_c = .13$]. Specifically, males had a higher than expected discordance rate (72.3%) and females had a lower than expected discordance rate (59.3%).

There was a statistically significant, but small difference in discordance rates based on ARIA+ for road crash-coded *health data* cases [$\chi^2(4) = 150.30$, $p < .001$, $\phi_c = .07$], with Inner Regional and Remote areas having a higher than expected discordance rate (see Table 7.50).

Table 7.50: Discordance rates between QRCD and road crash coded health data for different ARIA+

ARIA+	Number of cases in health	Number of non-linked cases	Discordance rate
Major Cities	17,539	11,419	65.1%
Inner Regional	10,134	7,235	**71.4%**
Outer Regional	6,193	4,116	66.5%
Remote	739	566	**76.6%**
Very Remote	363	227	62.5%

Note: Standardised residuals outside +/-3.10 are bolded

In order to take into account potential confounding factors, a logistic regression was performed. With all variables in the logistic regression, the model was statistically significant, $\chi^2(13) = 6334.93$, $p < .001$ (Nagelkerke $R^2 = .39$). After controlling for the relationships between the predictors, all variables remained statistically significant. Specifically, males had higher odds of discordance compared to females. Those aged 0-16 and 17-24 had higher odds (3.3 and 1.7 times respectively) of being discordant with QRCD compared to those aged 25-59. Motorcyclist and cyclist cases had higher odds of discordance with QRCD. Inner Regional and Remote cases had higher odds of discordance compared to Major Cities (see Table 7.51). It should be noted that seriousness is not included in this model as eARF had no serious coding.

Table 7.51: Logistic regression analysis of the profile of road crash injuries in health data that did not link to QRCD

		OR	OR[1]	99.9% CI[1]	p[1]
Gender	Female	1.00	1.00	Referent	
	Male	1.79	1.17	1.05 – 1.31	< .001
Age	0 – 16	4.64	3.30	2.75 – 3.96	< .001
	17 – 24	1.49	1.70	1.50 – 1.91	< .001
	25 – 59	1.00	1.00	Referent	
	60 +	0.82	0.87	0.74 – 1.03	.005
Road user	Driver	1.00	1.00	Referent	
	Motorcyclist	5.97	5.28	4.62 – 6.03	< .001
	Cyclist	18.89	13.39	11.15 – 16.08	< .001
	Pedestrian	1.39	1.15	0.90 – 1.48	.067
	Passenger	1.76	1.40	1.22 – 1.60	< .001
ARIA+	Major Cities	1.00	1.00	Referent	
	Inner Regional	1.34	1.26	1.12 – 1.42	< .001
	Outer Regional	1.06	1.07	0.93 – 1.22	.125
	Remote	1.75	2.02	1.39 – 2.92	< .001
	Very Remote	0.90	1.01	0.64 – 1.58	.969

[1] Adjusted for all variables in the equation

7.4.5 *Completeness of severity of injury*

As shown above (Section 7.4.1) there were 10,649 (55.9%) of cases in QRCD for 2009 that linked with at least one data hospital data collection and therefore possibly have extra and potentially more accurate information about the severity of injury available from other sources. Of these linked cases, 9,198 (86.4%) had an unknown injury severity (AIS and SRR) in QRCD. With the added information from the hospital data collections, 10,442 (54.8%) QRCD cases had complete injury severity coding.

Table 7.52 outlines the contribution of information from each hospital data collection to information about severity of injury.

Table 7.52: Number and percentage of QRCD cases with severity information provided by hospital data collections

	Severity information added (% of QRCD cases)
QHAPDC	4,029 (21.2)
EDIS	8,973 (47.1)
QISU	940 (4.9)
QHAPDC or EDIS	10,321 (54.2)
EDIS or QISU	9,194 (48.3)
QHAPDC or QISU	4,919 (25.8)
QHAPDC or EDIS or QISU	10,442 (54.8)

7.4.6 *Profiling of road crash injuries*

7.4.6.1 *Serious injury*

As shown in Table 7.53, the number of serious injuries differs depending on both the population source and the definition of a serious injury. Using police-reported cases as the population, the highest number of serious injuries would be obtained by including all cases that are reported to police (i.e., are included in the QRCD) and attend hospital (i.e., link with at QHAPDC, EDIS, or QISU). The lowest numbers of serious cases are identified from police reported cases that have an AIS higher than 3. When examining serious injury for cases identified in the *hospital data set* (not necessarily reported to police), attending hospital definition of serious yields the highest number of serious injuries. If the international definition of a serious injury (‘hospitalised’ for 24 hours or more) is applied, almost 30% of police reported and defined ‘hospitalised’ fit this definition. This number doubles if the entire *hospital data set* of QHAPDC, EDIS, and QISU is used (regardless of whether the case is reported to police).

Table 7.53: Number of police reported and hospital serious injuries based on different definitions

Definition	Police reported	Hospital cases
Police definition ‘hospitalised’	6,674	-
Attended hospital	10,649	29,261
Admitted hospital	4,283	8,391
Admitted hospital > 24hrs	1,879	3,474
AIS > 3	672	1,110
SRR < .942	1,041	1,507

7.4.6.2 *Crash and injured person characteristics*

As shown in Table 7.54 and Figure 7.12, there is little difference in the profile of road crash injuries between linked data and QRCD overall. There is however, a difference in profile between the *health* road crash population (i.e., all road crash injuries in QHAPDC, EDIS, QISU, and eARF) and the police-reported road crash injuries (i.e., all QRCD) Specifically, in the *health* population there were a higher proportion of cases aged 0-19 years, males, motorcyclists, cyclists, and cases in Inner Regional areas.

Table 7.54: Profiles of road crash injuries by gender, road user, and ARIA+

Variable	Level	QRCD all (%) n = 19,041	Health (%) n = 35,356	Linked (%) n = 13,566
Gender	Male	9,997 (52.8)	22,004 (62.6)	7,280 (53.7)
	Female	8,947 (47.2)	13,151 (37.4)	6,286 (46.3)
Road user	Driver	11,146 (58.5)	7,642 (29.6)	7,756 (57.2)
	Motorcyclist	1,820 (9.6)	6,659 (25.8)	1,555 (11.5)
	Cyclist	869 (4.6)	6,235 (24.2)	621 (4.6)
	Pedestrian	841 (4.4)	961 (3.7)	693 (5.1)
	Passenger	4,361 (22.9)	4,292 (16.6)	2,937 (21.7)
ARIA+	Major Cities	11,249 (59.1)	17,539 (50.2)	7,545 (55.6)
	Inner Regional	3,885 (20.4)	10,134 (29.0)	3,073 (22.7)
	Outer Regional	3,041 (16.0)	6,193 (17.7)	2,317 (17.1)
	Remote	514 (2.7)	739 (2.1)	394 (2.9)
	Very Remote	349 (1.8)	363 (1.0)	235 (1.7)

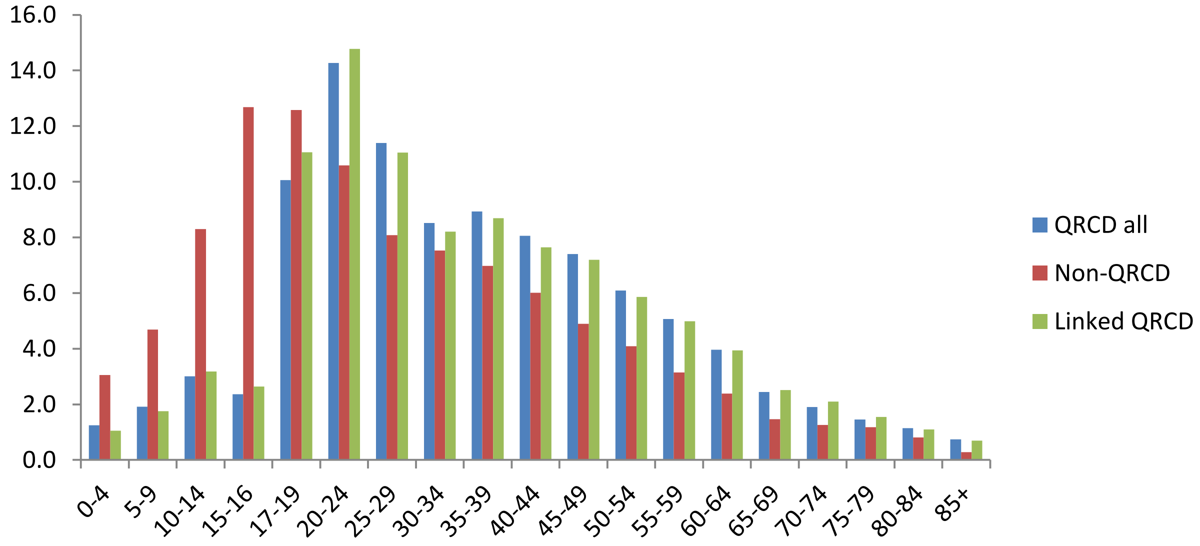


Figure 7.12: Age profile of road crash injuries for each population (QRCD, Health, and Linked QRCD

Looking specifically at police-coded 'hospitalised' cases, as shown in Table 7.55 and Figure 7.13, there is little difference in the profile of road crash injuries between QRCD cases which linked to the *hospital data set* (i.e., QHAPDC, EDIS, or QISU) and police-coded ‘hospitalised’ cases in QRCD. There is however, a difference in profile between the *hospital data set* (linked and non-linked) and police coded ‘hospitalised’ QRCD cases. Specifically, in the *hospital data set* there were a higher proportion of cases aged 0-19 years, males, motorcyclists, cyclists, and cases in Inner Regional areas.

Table 7.55: Profiles of hospital road crash injuries by gender, road user, and ARIA+

Variable	Level	QRCD hospital (%) n = 6,674	Hospital (%) n = 29,261	Linked hospital (%) n = 10,649
Gender	Male	3,800 (57.0)	19,158 (65.5)	5,857 (55.0)
	Female	2,871 (43.0)	10,101 (34.5)	4,792 (45.0)
Road user	Driver	3,573 (53.6)	4,883 (23.4)	5,869 (55.1)
	Motorcyclist	955 (14.3)	6,169 (29.5)	1,341 (12.6)
	Cyclist	354 (5.3)	6,095 (29.2)	513 (4.8)
	Pedestrian	424 (6.4)	721 (3.4)	596 (5.6)
	Passenger	1,365 (20.5)	3,034 (14.5)	2,326 (21.9)
ARIA+	Major Cities	3,603 (54.0)	14,434 (50.5)	5,959 (56.0)
	Inner Regional	1,479 (22.2)	8,530 (29.8)	2,458 (23.1)
	Outer Regional	1,178 (17.7)	4,738 (16.6)	1,753 (16.5)
	Remote	217 (3.3)	680 (2.4)	288 (2.7)
	Very Remote	197 (3.0)	203 (0.7)	190 (1.8)

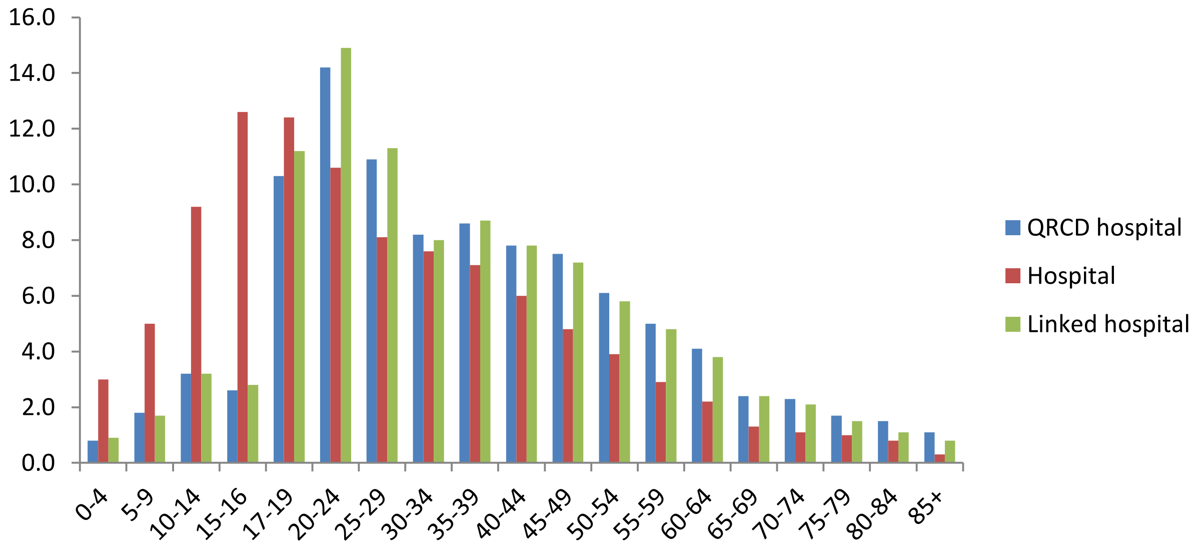


Figure 7.13: Age profile of road crash injuries for each hospital population (QRCD hospital, Hospital, and Linked hospital

7.4.7 *Validity*

7.4.7.1 *QRCD severity coding*

As mentioned in Section 7.4.1, there were 979 QRCD cases in 2009 that linked with QHAPDC that were not coded as ‘hospitalised’. Table 7.56 outlines the coding details of these cases.

Table 7.56: Cases not coded as ‘hospitalised’ in QRCD that linked with QHAPDC

	Number of QRCD cases	Number of linked cases	Linkage rate
Police-coded medically treated	7,962	857	10.8%
Police-coded minor injury	4,074	122	3.0%
TOTAL	12,036	979	8.1%

Similarly, there were 4,636 QRCD cases that linked with EDIS but were not coded as ‘hospitalised’ (see Table 7.57).

Table 7.57: Cases not coded as ‘hospitalised’ in QRCD that linked with EDIS

	Number of QRCD cases	Number of linked cases	Linkage rate
Police-coded medically treated	7,962	3,797	47.7%
Police-coded minor injury	4,074	839	20.6%
TOTAL	12,036	4,636	38.5%

There were 457 QRCD cases that linked with QISU but were not coded as ‘hospitalised’ (see Table 7.58).

Table 7.58: Cases not coded as ‘hospitalised’ in QRCD that linked with QISU

	Number of QRCD cases	Number of linked cases	Linkage rate
Police-coded medically treated	7,962	387	4.9%
Police-coded minor injury	4,074	70	1.7%
TOTAL	12,036	457	3.8%

When QISU and EDIS are examined together to represent emergency department cases, there were 4,731 QRCD cases that linked with emergency department data but were not coded as ‘hospitalised’, translating to a false negative rate of 48.5% (see Table 7.59).

Table 7.59: Cases not coded as 'hospitalised' in QRCD that linked with emergency department data

	Number of QRCD cases	Number of linked cases	Linkage rate
Police-coded medically treated	7,962	3,873	48.6%
Police-coded minor injury	4,074	858	21.1%
TOTAL	12,036	4,731	39.3%

When all of the *hospital data* collections are considered together, 5,005 QRCD cases linked with the *hospital data*, but were not coded as 'hospitalised'. This translates to a false negative rate of 47.0% (see Table 7.60). There were also 1,115 cases that did not link but were coded as 'hospitalised', indicating a possible false positive rate of 13.3%.

Table 7.60: Cases not coded as 'hospitalised' in QRCD that linked with hospital data

	Number of QRCD cases	Number of linked cases	Linkage rate
Police-coded medically treated	7,962	4,108	51.6%
Police-coded minor injury	4,074	897	22.0%
TOTAL	12,036	5,005	41.6%

Assuming the coding in *hospital data* are the 'gold standard', serious injury based on SRR in QRCD was coded correctly for 42.2% of specified cases with specificity and sensitivity being very low. The coding for AIS serious injury was correct for 39.5% of specified cases, also with very low sensitivity and specificity (see Table 7.61).

Table 7.61: Number and percentage of serious injuries (based on AIS) correctly coded in QRCD (compared to hospital data) and corresponding sensitivity and specificity

QRCD Severity	Correct (%)	Incorrect (%)	Sensitivity	Specificity
Serious SRR < .942	84 (42.2)	115 (57.8)	49.4%	9.2%
Serious AIS > 3	47 (39.5)	72 (60.5)	38.5%	5.9%

The validity of the serious injury classification based on the broad severity (i.e., 'hospitalised') in QRCD was examined, by comparing it to the AIS and SRR serious injury classifications for linked cases in *hospital data*. Table 7.62 shows that 'hospitalised' cases were more likely to be serious (based on AIS or SRR) and captured the vast majority of *hospital data* defined serious cases (around 90% for AIS and SRR). It should be noted, however, that there was still a large proportion of cases defined as 'hospitalised' that were not classified as serious by the hospital based definition (i.e., 88.3% AIS; and 83.0% SRR).

Table 7.62: Number and percentage of broad severity classification also classified as serious using hospital based AIS and SRR

		Hospitalised (%)	Other injury (%)	
AIS	Serious (> 2) Non-serious (< 3)	607 (11.7%) 4,580 (88.3%)	65 (1.4%) 4,421 (98.6%)	$\chi^2(1) = 391.21$, $p < .001$, $\phi_c = .20$
SRR	Serious (< 0.942) Non-serious (> 0.941)	924 (17.0%) 4,525 (83.0%)	118 (2.5%) 4,683 (97.5%)	$\chi^2(1) = 587.54$, $p < .001$, $\phi_c = .24$

7.4.7.2 *QHAPDC coding*

Traffic

As shown in Table 7.63, of the linked cases in QHAPDC there were a high proportion of cases with correct traffic coding. However, there were more than two-fifths of non-linked cases coded as traffic. As a result, the sensitivity of QHAPDC traffic coding was high (92.7%), but the specificity was moderate at best (54.4%). It should be noted that the specificity is influenced by the discordance between QRCD and QHAPDC, so may be due to under-reporting in QRCD, rather than incorrect coding in QHAPDC.

Table 7.63: Number and percentage of traffic and non-traffic coded cases in QHAPDC that linked to QRCD

	Traffic in QHAPDC	Non-traffic in QHAPDC
Link with QRCD	3,957 (91.7%)	360 (8.3%)
No link with QRCD	3,320 (42.1%)	4,561 (57.9%)

As mentioned in Chapter 5 (Section 5.3.3.3), within QHAPDC, the variable *place* could also give an indication of whether a transport injury is a road crash injury. Table 7.64 shows that almost two-thirds of those injuries coded as ‘traffic’ that did not link to QRCD were coded in QHAPDC as occurring on a ‘street/highway’. It could be argued that these cases are most likely true road crash injuries, despite not linking with QRCD, as the ‘traffic’ and *place* coding are convergent. Results also showed that almost 30% of injuries coded as ‘traffic’ that did not link with QRCD had an ‘unspecified’ *place*. In this case it is unclear whether these are miscoded or true discordance. In addition, 62% of the cases that linked to QRCD that were coded as non-traffic within QHAPDC were also coded as occurring on a ‘street/highway’. It could be argued that these cases are most likely miscoded in QHAPDC in regards to traffic status given that they were recorded in the police data and were coded as occurring on a street/highway.

Table 7.64: Number and percentage of linked and non-linked traffic and non-traffic injuries in QHAPDC by place coding

	Linked Traffic (%)	Not Linked Traffic (%)	Linked Not Traffic (%)	Not Linked Not Traffic (%)
Street/highway	3,810 (96.3)	2,110 (63.6)	224 (62.2)	433 (9.5)
Other place	39 (1.0)	250 (7.4)	58 (16.1)	2,103 (46.0)
Unspecified place	108 (2.7)	960 (29.0)	78 (21.7)	2,025 (44.5)

Table 7.65 shows the road user type for each of the different linked and traffic coded cases that were coded as ‘street/highway’ in QHAPDC. This table shows that for those that did not link but were coded as both ‘traffic’ and ‘street/highway’, motorcyclists and cyclists represented the majority of injuries. This result indicates that the bias in discordance for these road user types with police data found in Section 7.4.4.1 is, at least in part, a reflection of true discordance bias rather than misclassification. In contrast, 40% of the cases coded as not traffic that were coded as occurring on a ‘street/highway’ had a road user specified as a driver or passenger. This suggests that these cases are probably examples of misclassification of traffic status in QHAPDC. The almost 40% of cases with an unspecified road user which were not coded as traffic may represent a lack of documentation in the medical records.

Table 7.65: Number and percentage of linked and non-linked traffic and non-traffic injuries in QHAPDC by road user for place coded as street/highway

	Linked Traffic (%)	Not Linked Traffic (%)	Linked Not Traffic (%)	Not Linked Not Traffic (%)
Driver	1,705 (44.8)	336 (15.9)	66 (29.5)	29 (6.7)
Motorcyclist	686 (18.0)	746 (35.4)	25 (11.2)	88 (20.3)
Cyclist	200 (5.2)	445 (21.1)	3 (1.3)	141 (32.6)
Pedestrian	314 (8.2)	111 (5.3)	19 (8.5)	29 (6.7)
Passenger	731 (66.3)	321 (15.2)	24 (10.7)	26 (6.0)
Unspecified	174 (4.6)	151 (7.2)	87 (38.8)	120 (27.7)

Table 7.66 shows for the injuries that did not link that were coded as traffic where *place* was ‘unspecified’ a high proportion were motorcyclists and cyclists. This result in combination with the previous table suggests that, while the bias in discordance for these road user groups is still very evident, some of this may be the result of misclassification of traffic status rather than under-reporting. Also, over 50% of ‘unspecified’ place cases that were coded as non-traffic also had an ‘unspecified’ road user type; this again may represent a lack of documentation in the medical records.

Table 7.66: Number and percentage of linked and non-linked traffic and non-traffic injuries by road user for place coded as unspecified

	Linked Traffic (%)	Not Linked Traffic (%)	Linked Not Traffic (%)	Not Linked Not Traffic (%)
Driver	13 (12.0)	6 (0.6)	6 (7.7)	20 (1.0)
Motorcyclist	37 (34.3)	475 (49.5)	18 (23.1)	535 (26.4)
Cyclist	13 (12.0)	383 (39.9)	4 (5.1)	414 (20.4)
Pedestrian	2 (1.9)	6 (0.6)	5 (6.4)	64 (3.2)
Passenger	12 (11.1)	15 (1.6)	5 (6.4)	15 (0.7)
Unspecified	31 (28.7)	75 (7.8)	40 (51.3)	977 (48.2)

Road user

Assuming the coding in QRCD is the gold standard, linked QHAPDC cases had road user coded correctly for 94.9% of specified cases. This rate was slightly lower for cyclists (93.2%) and passengers (91.6%). Specificity and sensitivity was very high for all road users. However, sensitivity was lower for drivers and passengers (see Table 7.67).

Table 7.67: Number and percentage of road users correctly coded in QHAPDC and corresponding sensitivity and specificity

QHAPDC	Correct (%)	Incorrect (%)	Sensitivity	Specificity
Driver	1,722 (95.8)	76 (4.2)	94.9%	96.4%
Motorcyclist	744 (95.8)	33(4.2)	98.8%	99.0%
Cyclist	207 (93.2)	15 (6.8)	97.6%	99.2%
Pedestrian	335 (96.5)	12 (3.5)	96.5%	99.7%
Passenger	719 (91.6)	66 (8.4)	89.7%	97.9%
TOTAL	3,727 (94.9)	202 (5.1)	-	-

Table 7.68 shows that when drivers are incorrectly coded in QHAPDC, they are most often coded as passengers. Motorcyclists are most commonly incorrectly coded as cyclists, cyclists most commonly incorrectly coded as motorcyclists, pedestrians as drivers or passengers, and passengers as drivers. The majority of the unspecified cases in QHAPDC should have been coded (if more specific information was available) as drivers or passengers.

Table 7.68: Number and percentage of road users coded in QHAPDC corresponding to QRCD coding

	QRCD				
QHAPDC	Driver (row %)	Motorcyclist (row %)	Cyclist (row %)	Pedestrian (row %)	Passenger (row %)
Driver	1,722 (95.8)	1 (0.1)	0 (0.0)	1 (0.1)	74 (4.1)
Motorcyclist	18 (2.3)	744 (95.8)	3 (0.4)	7 (0.9)	5 (0.6)
Cyclist	4 (1.8)	7 (3.2)	207 (93.2)	3 (1.4)	1 (0.5)
Pedestrian	6 (1.7)	0 (0.0)	2 (0.6)	335 (96.5)	4 (1.2)
Passenger	64 (8.2)	1 (0.1)	0 (0.0)	1 (0.1)	719 (91.6)
Unspecified	225 (64.1)	13 (3.7)	1 (0.3)	4 (1.1)	108 (30.8)

7.4.7.3 *EDIS coding*

<u>Road crash</u>

Of the linked cases in EDIS, almost three quarters of cases were identified as road crash injuries (using EDIS text searching) and around 6% of non-linked EDIS cases were identified as road crash injuries (see Table 7.69). As a result, the sensitivity of EDIS road crash coding was moderate (54.4%) and the specificity was high (92.7%). As with QHAPDC, the specificity is influenced by the discordance between QRCD and EDIS, so may be due to under-reporting in QRCD, rather than incorrect identification in EDIS.

Table 7.69: Number and percentage of road crash and non-road crash cases in EDIS that linked to QRCD

	Road crash in EDIS	Non-road crash in EDIS
Link with QRCD	7,043 (73.6%)	2,532 (26.4%)
No link with QRCD	16,581 (5.6%)	277,714 (94.4%)

Further analyses were conducted to determine the influences on the identified road crash cases in EDIS that did not link with QRCD. A random sample of 1,000 identified road crash cases that did not link to QRCD text descriptions were manually reviewed. The review showed that 469 (46.9%) injuries involved a motorcycle or bicycle where the place of the injury was not specified; 167 (16.7%) injuries were not actually road crash injuries despite including key words (e.g., fallen off back of truck, finger caught in bike chain); and 364 (36.4%) were likely a road crash injury (e.g., RTC passenger single vehicle 80KPH, Pedestrian hit by a car on street). These results give an indication of the influence of misclassification on the discordance rates.

<u>Road user</u>

EDIS cases linked with QRCD had road user coded correctly for 91.7% of specified cases. This rate was lower for cyclists (82.8%) and passengers (83.9%). Specificity and sensitivity was very high for all road users. However, sensitivity was lower for drivers and pedestrians (see Table 7.70).

Table 7.70: Number and percentage of road users correctly coded in EDIS and corresponding sensitivity and specificity

EDIS	Correct (%)	Incorrect (%)	Sensitivity	Specificity
Driver	1,592 (97.8)	35 (2.2)	88.7%	98.4%
Motorcyclist	812 (92.9)	62 (7.1)	93.9%	98.0%
Cyclist	337 (82.8)	70 (11.2)	96.8%	96.6%
Pedestrian	103 (94.5)	6 (5.5)	83.7%	99.8%
Passenger	814 (83.9)	156 (16.1)	95.8%	95.0%
TOTAL	3,658 (91.7)	329 (8.3)	-	-

Table 7.71 shows that when drivers are incorrectly coded in EDIS, they are most often coded as passengers. Motorcyclists are most commonly incorrectly coded as drivers, cyclists most commonly incorrectly coded as motorcyclists, pedestrians as cyclists or passengers, and passengers as drivers. The majority of the unspecified cases in EDIS should have been coded (if more specific information was available) as drivers or passengers.

Table 7.71: Number and percentage of road users coded in EDIS corresponding to QRCD coding

	QRCD				
EDIS	Driver (row %)	Motorcyclist (row %)	Cyclist (row %)	Pedestrian (row %)	Passenger (row %)
Driver	1,592 (97.8)	4 (0.2)	0 (0.0)	3 (0.2)	28 (1.7)
Motorcyclist	37 (4.2)	812 (92.9)	9 (1.0)	11 (1.3)	5 (0.6)
Cyclist	13 (3.2)	46 (11.3)	337 (82.8)	5 (1.2)	6 (1.5)
Pedestrian	0 (0.0)	1 (0.9)	2 (1.8)	103 (94.5)	3 (2.8)
Passenger	153 (15.8)	2 (0.2)	0 (0.0)	1 (0.1)	814 (83.9)
Unspecified	3,471 (62.1)	389 (7.0)	126 (2.3)	404 (7.2)	1,200 (21.5)

7.4.7.4 *QISU coding*

Road crash

As shown in Table 7.72, a large majority of linked cases in QISU were identified as road crash injuries. However, around one-third of non-linked QISU cases were also identified as road crash injuries. As a result, the sensitivity of QISU road crash coding was high (92.6%) and the specificity was moderate (61.5%). As with QHAPDC and EDIS, the specificity is influenced by the discordance between QRCD and QISU, so may be due to under-reporting in QRCD, rather than incorrect identification in QISU.

Table 7.72: Number and percentage of road crash and non-road crash coded case in QISU that linked to QRCD

	Road crash in QISU	Non-road crash in QISU
Link with QRCD	899 (92.6%)	72 (7.4%)
No link with QRCD	1,579 (38.5%)	2,521 (61.5%)

Road user

QISU cases linked with QRCD had road user coded correctly for 92.7% of specified cases. This rate was lower for motorcyclists (84.5%) and cyclists (87.2%). Specificity and sensitivity was very high for all road users. However, sensitivity was lower for cyclists and passengers and specificity was lower for drivers (see Table 7.73).

Table 7.73: Number and percentage of road users correctly coded in QISU and corresponding sensitivity and specificity

QISU	Correct (%)	Incorrect (%)	Sensitivity	Specificity
Driver	528 (93.6)	36 (6.4)	96.2%	90.9%
Motorcyclist	82 (84.5)	15 (15.5)	95.3%	98.2%
Cyclist	41 (87.2)	6 (12.8)	91.1%	98.4%
Pedestrian	52 (98.1)	1 (1.9)	92.9%	99.9%
Passenger	172 (94.0)	11 (6.0)	82.7%	98.5%
TOTAL	875 (92.7)	69 (7.3)	-	-

Table 7.74 shows that when drivers are incorrectly coded in QISU, they are most often coded as passengers. Motorcyclists are most commonly incorrectly coded as drivers, cyclists most commonly incorrectly coded as drivers, pedestrians as cyclists, and passengers as drivers. There were no unspecified linked QISU cases for road user.

Table 7.74: Number and percentage of road users coded in QISU corresponding to QRCD coding

QISU	QRCD				
	Driver (row %)	Motorcyclist (row %)	Cyclist (row %)	Pedestrian (row %)	Passenger (row %)
Driver	528 (97.8)	3 (0.5)	0 (0.0)	1 (0.2)	32 (5.7)
Motorcyclist	8 (8.2)	82 (84.5)	3 (3.1)	0 (0.0)	4 (4.1)
Cyclist	4 (8.5)	1 (2.1)	41 (87.2)	1 (2.1)	0 (0.0)
Pedestrian	0 (0.0)	0 (0.0)	1 (1.9)	52 (98.1)	0 (0.0)
Passenger	9 (4.9)	0 (0.0)	0 (0.0)	2 (1.1)	172 (94.0)

7.4.7.5 *eARF coding*

Road crash

Around half of linked cases in eARF and 10% of non-linked eARF cases were identified as road crash injuries (see Table 7.75). As a result, the sensitivity of eARF road crash coding was moderate (50.4%) and the specificity was high (90.0%). As with QHAPDC and EDIS, the specificity is influenced by the discordance between QRCD and eARF, so may be due to under-reporting in QRCD, rather than incorrect identification in eARF.

Table 7.75: Number and percentage of road crash and non-road crash coded case in eARF that linked to QRCD

	Road crash in eARF	Non-road crash in eARF
Link with QRCD	5,831 (50.4%)	5,747 (49.6%)
No link with QRCD	6,121 (10.0%)	55,147 (90.0%)

Road user

eARF cases linked with QRCD had road user coded correctly for 91.7% of specified cases. This rate was lower for cyclists (82.8%) and passengers (83.9%). Specificity was very high for all road users, but a little lower for drivers. Sensitivity was high for driver and passengers, very high for motorcyclists and cyclists, but only moderate for pedestrians (see Table 7.76).

Table 7.76: Number and percentage of road users correctly coded in eARF and corresponding sensitivity and specificity

eARF	Correct (%)	Incorrect (%)	Sensitivity	Specificity
Driver	2,970 (91.6)	272 (8.4)	87.3%	91.6%
Motorcyclist	1,215 (82.7)	254 (17.3)	97.2%	95.2%
Cyclist	348 (92.8)	27 (7.2)	95.6%	99.2%
Pedestrian	185 (80.8)	44 (19.2)	64.0%	99.3%
Passenger	1,089 (82.4)	232 (17.6)	83.7%	95.7%
TOTAL	5,807 (87.5)	829 (12.5)	-	-

Table 7.77 shows that when drivers are incorrectly coded in eARF, they are most often coded as passengers. Motorcyclists are most commonly incorrectly coded as drivers, cyclists most commonly incorrectly coded as motorcyclists, pedestrians as drivers or cyclists, and passengers as drivers. The majority of the unspecified cases in eARF should have been coded (if more specific information was available) as drivers or passengers.

Table 7.77: Number and percentage of road users coded in eARF corresponding to QRCD coding

	QRCD				
eARF	Driver (row %)	Motorcyclist (row %)	Cyclist (row %)	Pedestrian (row %)	Passenger (row %)
Driver	2,970 (91.6)	10 (0.3)	31 (1.0)	89 (2.7)	142 (4.4)
Motorcyclist	182 (12.4)	1,215 (82.7)	4 (0.3)	5 (0.3)	63 (4.3)
Cyclist	4 (1.1)	16 (4.3)	348 (92.8)	7 (1.9)	0 (0.0)
Pedestrian	18 (7.9)	9 (3.9)	10 (4.4)	185 (80.8)	7 (3.1)
Passenger	227 (17.2)	0 (0.0)	2 (0.2)	3 (0.2)	1,089 (82.4)
Unspecified	3,282 (66.4)	124 (2.5)	127 (2.6)	279 (5.6)	1,128 (22.8)

Convergent validity of road crash identification

In order to explore the validity of the road crash injury identification of the *health* collections, regardless of whether they linked to QRCD, convergent validity was explored by calculating the number of road crash cases that were identified using each combination of the *health data* collections. Firstly, Table 7.78 presents how many data sets linked with one or two other data collections.

Table 7.78: Number and proportion of cases in each health data collection that linked with other health data collections

	Data collection			
	QHAPDC (%)	EDIS (%)	QISU (%)	eARF (%)
No other data set	2,056 (16.9)	261,872 (86.2)	1,116 (22.1)	35,461 (48.7)
One other data set	10,142 (83.1)	41,998 (13.8)	3,955 (77.9)	37,386 (51.3)
Two other data sets	4,954 (40.6)	5,547 (1.8)	1,034 (20.4)	4,466 (6.1)
Three other data sets	414 (3.4)	414 (0.1)	414 (8.2)	414 (0.6)

As shown in Table 7.79, three-quarters of road crash cases (that linked with at least one other *health data* collection) were identified by only one data source.

Table 7.79: Number and proportion of cases identified as a road crash across different number of data sets

	N	%
One data set	27,217	76.6%
Two data sets	6,929	19.5%
Three data sets	1,304	3.7%
Four data sets	86	0.2%

As shown in Table 7.80, QHAPDC and QISU had higher proportion of cases that were identified as a road crash in at least one other *health* data collection. However, it should be noted, that while this may be indicative of higher validity for these two data collections, this result is influenced by the proportion of those cases that linked with another data collection (see Table 7.78).

Table 7.80: Number and proportion of cases identified in each data collection as a road crash across different number of data sets

	Data collection			
	QHAPDC (%)	EDIS (%)	QISU (%)	eARF (%)
One data set	2,901 (39.9)	16,206 (68.6)	794 (32.2)	7,316 (61.2)
Two data sets	3,240 (44.5)	6,082 (25.7)	1,077 (43.5)	3,459 (28.9)
Three data sets	1,050 (14.4)	1,250 (5.3)	521 (21.0)	1,091 (9.1)
Four data sets	86 (1.2)	86 (0.4)	86 (3.5)	86 (0.7)

7.5 Discussion

7.5.1 *Summary of results*

7.5.1.1 *Linkage rates*

The linkage rates varied depending on the data source(s) QRCD was linked with and on the QRCD cases that were included in the link. Almost three quarters of all cases in QRCD linked with at least one of the *health* data collections (QHAPDC, EDIS, QISU, and eARF) and over half linked with at least one of the *hospital data* collections (QHAPDC, EDIS, and QISU). If only a subset of data is examined (i.e., 'hospitalised' cases in QRCD) the proportion of cases that linked with at least one *health* data collection rises to almost 95% and links with the *hospital data* collections rises to around 80%. As a result, it is possible that for just over half of all QRCD cases and around 80% of 'hospitalised' QRCD road crash cases, *hospital data* coding of the injury (using ICD-10-AM coding) could be applied. This is important considering that the classification of severity based on the broad coding of 'taken to hospital' by police is in question, as is the availability and accuracy of SRR and AIS coding of police data injury descriptions (see Chapter 5, Section 5.4.1 and Section 5.4.4).

In terms of the QRCD cases, where police recorded an injury, that do not link with any of the data collections, there could be several explanations. Firstly, the *health* data collections in this current study do not necessarily cover all the possible road crash cases that are reported to police. EDIS and QISU, for example, do not cover the entire Queensland population of emergency departments (75% of emergency departments are included in EDIS, Toloo et al. (2011) . Also, not all cases are attended by an ambulance (eARF) or admitted to hospital (QHAPDC). Some injuries may be treated at other medical facilities not included in these collections (e.g., General Practitioners (GPs)). It is also possible that some of the QRCD cases do not involve injuries at all, despite being recorded as such, or involve injuries for which no treatment was ever sought. Finally, it is also possible that the discrepancy in linkage may be due in part to the linkage process itself. There may have been incomplete information or errors in some cases (either in QRCD or the other data collections), making linking impossible. In other words, some of the discrepancy was due to a link not being able to be found, rather than the case not being present in the combination of data collections.

7.5.1.2 *Completeness of cases*

The completeness of road crash cases reported to police was examined via discordance rates between the QRCD and the *health* data collections. The level of discordance varied depending on the population being compared and the definitions within those data collections. When all data collections are examined together the estimated population of road crash injuries was approximately 35,000, with around two-thirds not linking to any record in QRCD. This discordance indicates the level of under-reporting of road crash injuries to police and is somewhat similar to the level of discordance found in other studies (Alsop & Langley, 2001; Amoros et al., 2006; Boufous et al., 2008). It should be noted however that this discordance was lower (around 50%) when only linkage with QHAPDC or eARF was considered. There may be a number of reasons for less discordance with these data collections. The lower discordance with QHAPDC could indicate that when a road crash injury is more severe (i.e., requiring hospital admission); the more likely it would be to come to the attention of police. For the lower rate of discordance with ambulance, it may be that when an ambulance attends the scene of a road crash, the police may be more likely to also attend the scene. It is possible that those cases where an injured person attends hospital in private transport (instead of arriving by ambulance), the less likely the injury would be reported to police. It is also possible, that the discordance rates with EDIS and QISU may also be the result of misclassification of cases. This may particularly be the case with EDIS, where the identification of cases relies on text searching which may overestimate the population. This issue will be discussed further in Section 7.5.1.6 on validity. Regardless of the differences in the discordance rates, the results show that there is a significant level of under-reporting to police.

7.5.1.3 *Consistency*

Comparisons between cases in QRCD that linked and did not link with QHAPDC showed that linked cases were more likely to be serious (as measured by AIS and SRR), and to involve motorcyclists and pedestrians. For linking with EDIS, linked cases were also more likely to be serious. For QISU, linked cases were more likely to be young and be from outside Major Cities. Linked cases with eARF were more likely than non-linked cases to be motorcyclists in Inner Regional and Outer Regional areas. The linkage bias found across the collections is likely to represent the nature of the *health* data collections. Firstly, these data collections would be more likely to have serious cases included within them and therefore serious cases in QRCD would be more likely to link with cases from at least one of the other collections. This may also explain the higher than expected linkage rates for motorcyclists and pedestrians. These road users are more vulnerable (i.e., more likely to sustain more serious injuries) and therefore are more likely to be included in health data collections. The linkage bias for QISU (younger cases and location) likely represents the hospitals included in the QISU collection. Specifically, QISU includes the two largest emergency children's hospitals and a number of hospitals in regional areas. Despite the differences between linked and non-linked data, this does not appear to have substantial impacts on the profile of linked data compared to all of QRCD (see Section 7.5.1.6).

The pattern of discordance rates between all *health* collections combined and QRCD was also examined to explore the issue of bias in under-reporting to police. It was found that for QHAPDC discordance was higher for young people, motorcyclists and cyclists and lower for more serious injuries and cases involving another vehicle. This pattern was similar for QISU, EDIS and eARF. Although it should be noted that eARF did not include serious injury information and both eARF and EDIS did not have information about another vehicle being involved. For the *hospital data* there was also a difference in discordance on the basis of ARIA+ location. Specifically, Remote and Inner Regional locations had higher discordance rates compared to Major Cities. This may possibly reflect greater levels of under-reporting to police in these locations. As noted earlier, there were lower discordance rates (i.e., lower levels of under-reporting when another vehicle was involved in the incident that caused the injury). It is possible that injuries resulting from collision with another vehicle are more serious and therefore more likely to be reported. It could also be argued that when another motorist is involved there would be insurance implications that could provide the impetus to report the crash to police. The bias in under-reporting found in this study is similar to that found elsewhere (Alsop & Langley, 2001; Boufous et al., 2008; Langley, Dow, et al., 2003).

These results indicate that not only is there a level of under-reporting to police; there are certain types of injury cases that appear to be less likely to be reported. It should be noted however, that the bias found in discordance for road user may be exaggerated due to validity issues, particularly with EDIS. See Section 7.5.1.7 on validity for more discussion of this issue.

7.5.1.4 *Completeness of data*

The number of cases with unknown injury description in QRCD and therefore undetermined severity (based on AIS and SRR) was more than halved by the linkage with *hospital data*. Almost all of those cases that still had unknown information were due to them not linking to a *hospital data* collection. The results showed that more than half of the QRCD cases would have more complete and potentially accurate injury nature and severity information added to them by linking to *hospital data*.

7.5.1.5 *Severity profile of road crash injuries*

There was a large amount of variation in the estimates of serious road crash injuries depending on the population of reference and the definition or measure used. If the current reporting practice definition within QRCD is used (i.e., police-reported 'hospitalised'), there were around 6,000 serious road crash injuries in 2009. If the number of police-reported road crash injuries that were actually 'taken to hospital' is considered (based on the cases linked with QHAPDC, EDIS, or QISU), the number of serious injuries rises to approximately 10,000. If the international definition of a serious road crash injury is applied (i.e., admitted to hospital for 24 hours or more), there was slightly less than 2,000 serious road crash injuries reported to police. When AIS and SRR are used to classify serious injury the numbers are approximately 600 (AIS > 3) and 1,000 (SRR < .942) serious injuries respectively, reported to police. The number of serious injuries increases dramatically, if requirement for reporting to police is removed. Specifically, if all cases 'taken to hospital' (regardless of whether they are reported to police) are counted, there were almost 30,000 serious injuries. Admitted to hospital for 24 hours or more was around 3,500 and using AIS and SRR based definitions there were approximately 1,000 and 1,500 respectively.

7.5.1.6 *Profile of road crash injuries*

There were no real differences between profiles of the police-reported linked cases and police-reported cases (linked and non-linked) for both total police-reported and 'hospitalised' police-reported cases on the basis of age, gender, road user, or ARIA. This highlights the potential for linked data to be used (due to the added injury nature and severity information) without biasing the profile of characteristics. There were however, differences in the profiles of police-reported road crash injuries and the *health* population (both all cases and those in *hospital data* only). Specifically, the *health* population cases had a greater proportion of injuries to young people, males, motorcyclists and cyclists, and cases in Inner Regional areas. These results represent the discordance bias discussed previously (Section 7.5.1.3) and highlight the potential bias in profile if only police-reported road crash injuries are examined.

7.5.1.7 *Validity*

Results of comparisons between the broad severity categorisation in QRCD (i.e., 'hospitalised', 'medically treated', 'minor injury') and the *hospital data* collections revealed that there were approximately 5,000 (26%) cases that did link with a *hospital data* collection but were not coded as 'hospitalised' in QRCD. Also, there were approximately 1,000 (6%) cases coded as 'hospitalised' that did not link to any of the *hospital data* collections. This result indicates a potential issue with the classification of 'hospitalised' (in this case taken to hospital) by police. It is possible that the some of the cases coded as 'hospitalised' that did not link to any *hospital data* are due to these cases not being included in the current collections as they appeared for treatment elsewhere (e.g., a hospital not included in EDIS or QISU) or were not actually admitted to a hospital and thus did not appear in QHAPDC. The result of under-ascertainment (i.e., did link but were not coded as 'hospitalised', n ≈ 5,000) indicates that a certain proportion of cases that are potentially serious are not being coded as such by police, as they are unaware that the injured person ultimately sought treatment at a hospital.

Another aspect of severity classification using police-reported data was the issue of severity coding based on injury description. Firstly, as shown in Chapter 5, Section 5.4.1, the injury description on which AIS or SRR could be based is often missing or insufficiently sufficient for a classification to be applied. In addition, the results of this study indicate that even when QRCD has a specified injury coding (AIS or SRR), it is incorrect the majority of the time (around 60%). The final validity issue with the classification of injury severity in QRCD relates to whether the cases coded as 'hospitalised' in QRCD were more serious (as measured by AIS and SRR) than those coded as 'other injury'. Results showed that 'hospitalised' cases were more likely to be serious than 'other injury cases' and did capture the vast majority of *hospital data* defined serious cases (around 90% for both AIS and SRR). However, there were still a large proportion of cases defined as 'hospitalised' that were not classified as serious by the hospital based definition.

An examination of the validity of the selection of road crash injury cases in each of the *health* data collection produced mixed results. For QISU, the ability to correctly identify a road crash case (sensitivity) was very good. However, this data collection was only moderately good at correctly identifying that a case was not a road crash (specificity). QHAPDC, EDIS, and eARF had moderate sensitivity and high specificity. This result indicates that *health* data are generally good at rejecting cases when they should. However, they are less capable of including a case when they should. It should be noted that there is an interaction between the specificity results and the discordance between these data collections and QRCD and it is not possible tease out their respective effects just by comparing them with QRCD.

In an attempt to clarify this issue somewhat, further analyses were conducted on the cases identified as road crash injuries in both the QHAPDC and EDIS data collections which did not link with QRCD. For QHAPDC, it was found that approximately 60% of 'traffic' coded injuries that did not link to QRCD (discordant cases) were coded as 'street/highway' for the *place* variable. It could be argued that these cases are most likely to be true road crash injuries, despite not linking with QRCD, as the *traffic status* and *place* coding are convergent. Results also showed that around 30% of injuries coded as 'traffic' that did not link with QRCD had an 'unspecified' *place* in QHAPDC. For these cases it is not able to be determined if these are true road crash injuries or false positives (i.e., a product of misclassification). These results suggest that somewhere between 60% and 90% of QHAPDC road crash injuries that did not link with QRCD are actually road crash injuries and thus represent under-reporting of road crash injuries to police. A manual review of a random sample of EDIS cases revealed that almost half of the identified road crash injuries that did not link to QRCD were identified in text as involving a 'motorcyclist' or a 'cyclist' and the place of the injury was not specified. For these cases, it is still unclear what proportion represents under-reporting and what proportion represent misclassification. Around 17% were identified in text as not being road crashes injuries and almost 40% were identified as likely to be road crash injuries. These results suggest that somewhere between 40% and 83% of EDIS identified road crash injuries that did not link with QRCD are actually road crash injuries and thus represent under-reporting of road crash injuries to police.

To further examine validity of coding for road crash injury identification, convergent validity was explored. The commonalities between the *health* data sets for defining a road crash were examined. The results indicated that approximately three-quarters of cases (across all data collections) were identified as a road crash injury by only one data collection. This does suggest some doubt over the selection of road crash cases in *health* data. When each of the data collections were examined separately, QHAPDC and QISU had a higher proportion of cases (over 60%) that were identified as a road crash in more than one data collection, suggesting reasonable validity. EDIS and eARF, however, had only one-third of cases identified as a road crash in more than one data collection. This result, does suggest that while the respective effects of discordance and misclassification are not entirely clear, that it is possible that misclassification could be influencing the level of discordance with QRCD.

The examination of the validity of the *health* data collections for classifying road user revealed that when road user is specified, the linked *health* data collections have a relatively high proportion of correctly classified cases (using QRCD as the reference standard). However, particularly in EDIS and eARF, when the road user was not specified, the case was more likely to be a driver or a passenger. It seems that when text fields are used to determine the road user of a case, the text is more likely to specify a road user when it is a cyclist, motorcyclist, or a pedestrian. It is possible that when emergency and ambulance personal are completing these descriptions, they may believe it is more clinically relevant to mention these vulnerable road user groups as opposed to the light and heavy vehicle occupants. Whatever the reason, these results indicate that the road user bias in discordance rates may be overestimated. Also, as it is possible that cyclists and motorcyclists may be more likely to be injured off-road (although not identified as such in the text description), this finding may have an impact on the overall discordance rates.

7.5.2 *Limitations*

One of the limitations of this study, as with other studies using probabilistic linkage methods, is that it cannot be determined how many of the non-linked cases were due to linkage errors rather than being true non-links. While the Queensland Health DLU commented that they thought the quality of linkage was very high, specificity and sensitivity were not able to be calculated due to a large number of manual reviews. Also, the DLU suggested that any errors in the linkage were more likely to be the rejection of links that did exist. Another linkage issue relates to the less specific personal information being available in QISU. Unlike the other data collections, the QISU data collection does not include name and date of birth, thus affecting the linkage rates for this data collection. As a result of all of these issues, while probably only affecting a relatively small amount of cases, it is possible that some of the non-links are due to linkage error rather than true non-links. Despite attempts being made to explore the issue of misclassification in the form of validity analyses, it was still not possible to exactly quantify how much of the misclassification of cases and/or variables influenced discordance rates. Further research into this issue is required to tease out the relative influence of these factors.

Also, in order to identify cases and classify variables such as road users, this study used methods (text terms, coding practices) that are commonly used in research of this nature. However, these methods, as highlighted to some extent in the validity analyses, may result in inaccurate identification and classification of cases. It is possible that these methods could be improved through more elaborate search and/or data mining tools and/or techniques that are increasingly being applied in this type of research. Related to this issue was that the technique for dealing with duplicates in this study differed from Study 2. In Study 2, duplicate cases were not able to be directly identified as there was no person ID. As a result, the method used in Study 2 was a crude one, in which all transfers in the *hospital data* were removed from analyses. In the current study, not all duplicates were removed and instead only those where it was clear (based on dates of arrival, admission, and discharge) that it was the same injury case were removed. This resulted in a larger number of cases identified in Study 3 than Study 2. While the method is Study 2 would have most certainly resulted in an underestimation of the total number, it is also possible that the method applied in this current study, although consistent with methods applied in other studies (Davie, Samaranayaka, Langley, & Barson, 2011; Lujic, Finch, Boufous, Hayen, & Dunsmuir, 2008) resulted in an overestimation. This should be taken into account in terms of any conclusions relating to under-reporting.

Another issue worth noting is the mapping of ICD-10-AM coding to AIS and SRR. For SRR, the mapping corresponds directly to ICD-10-AM. However, the AIS mapping corresponds to ICD-10 and is then extrapolated to ICD-10-AM. The correspondence between ICD-10 and ICD-10-AM is at a level less specific than would otherwise be the case. As a result the reliability of the assignment of AIS may be in question. In addition, for both AIS and SRR, there were still a number of cases in the *hospital data* that could not be assigned a value, while this was not a large proportion it may still be considered significant. Further research should be conducted to improve the current severity mapping practices. Also, status of ICD-11 should be monitored as this new coding system may better allow for mapping to these measures. A related limitation is the use of a single SRR rather than using multiple SRRs to form an International Classification of Diseases Based Injury Severity Score (ICISS). It was not possible to compute ICISS in this study as only one diagnosis was available in the EDIS and QISU data collections. While there has been some research suggesting that a single SRR may be just as useful as the multiplicative method (Henley & Harrison, 2009) this assumes the single diagnosis is the 'worst injury' that an injured person has. It could be argued that the principal diagnosis could represent the 'worst injury'; further examination of this issue with the current data may be the subject of future research. The other limitation is relation to severity coding is the use of 'threat to life' measures. Further research could examine the potential of other injury severity indicators (e.g., Disability Adjusted Life Years (DALYs), length of stay), to explore the impact of injuries not just in terms of 'threat to life', but also the impacts of disability and the burden on the health system.

The final limitation is that this study did not include all the possible data collections that could potentially hold information or cases relating to road crash injuries in Queensland. For example, the Queensland Motor Accident Insurance Commission holds data relating to personal injury insurance claims in Queensland. However, it is a requirement that the crashes that lead to an injury claim be reported to police and therefore each injury in MAIC should by definition be included in QRCD. Also, not every injury crash can or will result in an injury claim and therefore, these data would likely only be a subset of the QRCD. However, while MAIC data may have additional information relating to the injury itself, it is not expected that it would include any injury information above and beyond what is included in *hospital data*. Another possible data source was the Queensland Trauma Registry. These data include coded injury information for acute hospital injury cases with an admission of greater than 24 hours. As with MAIC data, the QTR would only be a subset of another data source included in this study (i.e., QHAPDC). Also, while the QTR has detailed follow-up information about acute injuries which may be of interest, the collection of QTR cases ceased at the end of 2012. Therefore, it was considered unnecessary to explore the feasibility of this data collection in future linkage research as it is no longer being collected.

7.6 Chapter Summary

This chapter described the third study conducted as part of the research program. It involved the secondary data analysis of five linked data collections that include road crash injury cases in Queensland. This study has shown how data linkage can be used to investigate issues of data quality particularly in relation to defining serious injury and estimating the extent of under-reporting of road crash injuries to police. In addition, it has been shown that by linking other data sources with QRCD, improvements to reporting and the classification of serious injury can be achieved. This study has also shown however, that some caution is needed in assuming that the *health* data collections include all relevant cases and that these cases are always accurately identified. Further research on this issue is required, including the refinement of the methods used to identify cases and classify road users in these data. It is also possible that data linkage in the future could restrict the data collections linked with QRCD to those that are most relevant to the purpose of use and have the most accurate information. Despite some limitations, this study has shown that linking road crash data in Queensland is possible. It has also shown how the methodology applied here could be utilised (possibly with some refinement) in other jurisdictions. It has also demonstrated the potential improvements to the understanding of the road safety problem, particularly serious injury, by conducting data linkage. Even if linkage was not performed routinely, further research could be conducted to develop adjustments based on linked data, which could then be applied routinely to current reporting, for a more accurate representation of the road trauma problem.

Chapter Eight: Discussion

8.1 Introductory Comments

This program of research has explored the quality of current sources of road crash injury data and the linkage opportunities that exist within Queensland in order to provide a more comprehensive picture of road crashes and the resulting injuries. It also addressed not only whether road safety data linkage is feasible in Queensland, but whether data linkage provides qualitative and quantitative improvement to current practice. This final chapter will draw together the findings from Studies 1a, 1b, 2, and 3 and discuss the practical implications for road safety. The limitations of the research will also be discussed, along with suggestions for future research.

The first section will review the main findings of the program of research in terms of the research questions identified at the end of Chapter 2, which have been used to guide the program of research.

8.2 Review of Findings

8.2.1 *How well do data collections which collect road crash injury information in Queensland conform to the core/minimum requirements for road crash injury data?*

Study 1a results suggest that the relevant data collections vary in the extent to which they conform to the core/minimum requirements for road crash injury data. Some of the data collections conform very well, others less so. Arguably, QRCD included the most data elements recommended by the guidelines. This is perhaps not surprising given its primary purpose is for road safety reporting and research.

Overall, QRCD, QHAPDC, NCIS and QISU have a high level of completeness of the Core MDS, Core ODS, and Supplemental data sets. eARF and EDIS, however, have only half of these variables at best. In terms of the other recommended variables, QRCD is clearly the most complete, with the other data collections lacking coded variables on many of these factors (e.g., information on specific circumstances (e.g., speed, fatigue), or other crash or road user characteristics (e.g., road environment, seating position, licence status)).

8.2.2 *What are the strengths and weaknesses of each of the road crash injury data collections within the context of road safety investigation, intervention development, and evaluation?*

Studies 1b and 2 described the strengths and weaknesses of the identified data collections. A key issue emerging from the interviews is the possibility that the classification of injury severity in QRCD may be lacking. The results of the interviews also gave further indication of the under-reporting of road crash injuries to police. As mentioned previously, this under-reporting and lack of precision in assigning injury severity would impact on the estimated impact on and cost of crashes to our community. In contrast, based on the interviews with the custodian and data users, QRCD was seen as highly consistent overtime. It would seem therefore that while the data may not be entirely accurate or complete, any inaccuracies or incompleteness would be consistent over time. Therefore, while not being an accurate representation of all road crash injuries, it would be reliable enough to establish trends in the data and to be confident that any changes in the number of road crash injuries would represent actual changes which would be important for road safety evaluation and monitoring.

While Queensland Road Crash Database (QRCD) has a lot of relevant information and is mostly complete on the Core Minimum, Core Optional, and Supplemental variables, it is lacking in the key area of severity. Using the broad classification of fatal; 'hospitalised'; and other, while complete, is lacking in precision. Specifically, the category of 'hospitalised' (which is currently used to define serious injuries) is very broad in its range of more objective measures of severity, such as Survival Risk Ratio (SRR). Also, based on SRR, 'other injury' (which is currently used to define non-serious injuries) had a lower median SRR (indicating a greater level of serious injury) than 'hospitalised' cases. It seems then that the use of this broad severity measure (which is currently the case) is potentially a misrepresentation of the true seriousness of cases.

It could be argued that the fact that the 'hospitalised' category refers to cases where injured people are taken to hospital could explain the lack of precision in measuring the severity of a case. Rather, basing this category on whether a person was admitted to hospital for 24 hours or more, as it is specified by the International Road Traffic and Accident Database definition (IRTAD, 2005) could be a better indicator. However, examination of QHAPDC showed that it too had quite a broad range of SRRs among the 'hospitalised' cases. Also, it has been suggested that using these broad measures (based on the nature of medical intervention), means that the severity of the cases is influenced by things such as admission policies and that these policies are not necessary a reflection of the true clinical severity of a case and can often vary over time. Due to these issues, it is preferable to base indications of severity on clinical measures such as AIS and SRR. The problem here, for the police data, lies in that this study has found a large amount of missing and unspecified information in these data to determine a clinically based severity measure. There is also a bias in the completeness of this variable that could affect the determination of the seriousness of cases. The incompleteness and inconsistency of the information required for determining objective severity measures provides further evidence that using police data alone for determining severity is problematic.

These potential limitations with the QRCD have an impact on road safety research and policy. An accurate representation of the road crash injury problem in terms of severity and prevalence is essential for: prioritising funding and resources; targeting road safety interventions into areas of higher risk such as in different road user groups or locations (urban/rural); calculating the cost of road crash injuries in order to estimate the burden of road crash injuries in terms of future disability; and calculating the cost/ benefit ratio for evaluating interventions aimed at reducing road crash injuries.

Based on these demonstrated limitations of the QRCD, it is possible that the health data collections could potentially add to the understanding of the prevalence and severity issues. However, as these data collections are not designed with the specific purpose of road crash injury surveillance, there are shortcomings in these collections which impact on the reliable identification of the relevant cases. This is not just in terms of the validity of the selections, which varied across data collections in this study, but also in the ease with which these selections can be made. For example, EDIS only has text descriptions to determine whether a case is a road crash. More particularly, the analyses conducted in this study suggest that using this selection method could lead to an overestimation of road crash cases. QHAPDC, QISU, and eARF are better in this regard as they have coded variables allowing the selection of cases. However, there were still instances of missing or unspecified cases for some key variables (such as place), that could impact on the validity of the estimates of road crash cases for these data collections also.

8.2.3 *To what extent are the road crash injury data collections consistent with one another in terms of scope, data classification, and epidemiological profile?*

The results of Study 2 highlight the scope, classification and profile differences between QRCD and the other data collections. Each of the health data collections only represents a subset of the road crash injuries that are reportable to police. eARF is fairly comprehensive, however there would be some cases where an ambulance would not attend a road crash injury incident, or that an injured person could alert police or attend a medical facility without requiring an ambulance. QHAPDC is comprehensive in that it covers every hospital in Queensland; however it only includes cases that are admitted to hospital. This results in not only a limited estimate of the number of road crash injuries, it is also biases it towards particular injuries and injured persons that are more likely to involve a hospital admission. EDIS and QISU capture more than just admitted patients, which in some ways increases their scope, however both include only cases that present at hospitals included in the collection and in both collections this is not all hospitals in Queensland. There is also some bias in the included hospitals, particularly for QISU, which includes a large children’s emergency department, but excludes the largest emergency department (both children and adult) in Queensland. Also, as discussed previously, the included hospitals in each of the collections has changed over the years, thus affecting the ability to reliably estimate the number of road crash injuries over time.

In terms of overall numbers, the difference between QRCD and QHAPDC was minimal, with QRCD having slightly more cases than QHAPDC. When the profiles were compared at a bivariate level, there were significant differences between QRCD and QHAPDC (e.g., age and road user type). These differences provide some evidence of under-reporting within QRCD, because as noted above it would expected that QRCD should have more cases as the scope is broader than QHAPDC. However, it is possible that some of the differences found are not due to under-reporting, but instead due to misclassification of road crash injuries in QHAPDC. For severity, there was no difference between the collections in terms of the proportion classified as serious based on *Survival Risk Ratio (SRR)*. However, QRCD had a greater proportion of fatalities and serious or worse *AIS* classification compared to QHAPDC. In comparison to QRCD, eARF had fewer cases overall. It is not clear exactly why eARF has fewer cases than QRCD; however it may be due to the inclusion of minor injuries (which are not medically treated) in QRCD. It is also possible that these are the crashes resulting in injuries included within the QRCD where an ambulance was not in attendance. QISU had considerably fewer cases than QRCD. It would not be expected that the prevalence of road crash injuries in QISU would correspond with that of QRCD, as QISU hospitals are only a subset of hospitals in Queensland in which a road crash injury could present. Compared to QRCD, there was many more road crash injury cases included in EDIS. NCIS had two more cases than QRCD. It was expected that these data collections would match up exactly as all fatal road crash injuries should be reported to police and to the Coroner. However, there was some indication that the inclusion of road crash deaths in NCIS has a different basis than that of QRCD. It should also be noted that the cases were not completely the same (not just in number but also in distribution), highlighting that there may be some other differences with one or both of the data collections in terms of inclusion and/or coding.

8.2.4 *What are the facilitators of* and *barriers to linking road crash injury data collections in Queensland and elsewhere?*

Based on interviews with custodians, expert data users, and data linkage experts (Study 1b), the results indicated that there are many perceived benefits of data linkage including efficiency, increased samples sizes, and the ability to conduct research on issues that would not be possible with only one data collection. Specifically, it was suggested that the major potential benefit of data linkage to road safety research would be the ability to gain a more complete picture of both the circumstances and outcomes relating to road crash injury. There were also some barriers to data linkage highlighted relating to lack of resourcing, skills, and information, as well as potential reluctance among the relevant custodians to share the data required for linkage to occur. Overall, however, most participants were keen to see linkage trialled with road crash injury data in this jurisdiction.

In Study 3, some of the barriers identified above did actually pose problems for undertaking the linkage process. The time taken to gain ethical clearance and data custodian agreements was approximately twenty months. Due to issues with some of the data (incomplete or incorrect personal information), a large number of manual reviews needed to be conducted, so the data linkage process conducted by Queensland Health took approximately five months. As a result, the time taken from applying for ethics to obtaining the data was over two years. While it did take a considerable time to gain approval and for the data linkage to be completed, many of these issues were due to this being the first study of its kind in Queensland. Now that agreements are in place and the method has been established, it would be arguably easier and less time consuming to conduct linkage of this type in the future.

8.2.5 *What aspects of road crash injury data quality can be improved by using linked data for road safety investigation, intervention development, and evaluation?*

The results of Study 3 showed that the use of linked data highlights and could potentially quantify data quality issues with road crash data. Firstly, the results of Study 3 confirmed that there are a number of road crash injuries that are not reported to police as shown in studies elsewhere in Australia (Boufous et al., 2008; Rosman & Knuiman, 1994) and in other countries (Alsop & Langley, 2001; Amoros et al., 2006; Langley, Dow, et al., 2003). It has also confirmed the pattern of under-reporting found elsewhere in terms of bias towards certain types of road users (i.e., cyclists and motorcyclists). While the level of discordance (i.e., road crash injuries that did not link to QRCD) varied depending on the population being compared and the definitions within those data collections it tended to range between 46% and 69%. It is possible, however that the discordance rates may also be the result of misclassification of cases. This may particularly be the case with EDIS, where the identification of cases relies on text searching which may inaccurately estimate the population. Regardless of the differences in the discordance rates, the results suggest that there is still a substantial level of under-reporting of road crash injuries to police. Based on validity analyses and discordance rates it is estimated that this may be somewhere between 30% and 60%. In addition to the level of under-reporting, the data linkage in Study 3 highlighted the issue of bias in under-reporting. Specifically it was found that for QHAPDC, discordance was higher for young people, motorcyclists and cyclists and lower for more serious injuries and cases involving another vehicle. This pattern was similar for QISU, EDIS and eARF. Although it should be noted that eARF did not include serious injury information and both eARF and EDIS did not have information about another vehicle being involved. For the *hospital data* there was also a difference in discordance in the basis of ARIA+ location. Specifically, Remote and Inner Regional locations had higher discordance rates compared to Major Cities. This may possibly reflect greater levels of under-reporting to police in these locations. The bias in under-reporting found in this study is similar to that found elsewhere (Alsop & Langley, 2001; Boufous et al., 2008; Langley, Dow, et al., 2003).These results indicate that not only is there a level of under-reporting to police; there are certain types of injury cases that are less likely to be reported.

Another data quality issue with QRCD highlighted by the use of linked data related to the classification of serious injury. Validity analysis demonstrated that there were some cases coded as ‘taken to hospital’ that did not link with any *hospital data* (6% of police-reported injuries). Also, there were quite a number of cases that were not coded as ‘taken to hospital’ but were in fact recorded in the *hospital data* (26% of police-reported injuries). It was also found that the many of the police defined serious cases did not align with the AIS or SRR definition derived from the *hospital data*. These results demonstrate that relying on police data for serious injury reporting has clear limitations.

Another benefit of using linked data is the potential for obtaining additional information about cases in the QRCD (police data), from other data sources. More particularly, this study examined linkage rates of police-reported cases to hospital data collections (with police-reported road crash injuries as the denominator), rather than just focussing on the discordance (or under-reporting) in the police data (with the hospital data as the denominator). For example, Study 3 showed that the number of cases with unknown injury description in QRCD and therefore undetermined severity (based on AIS and SRR) was more than halved by the linkage with *hospital data*. Almost all of those cases that still had unknown information were due to them not linking to a hospital data collection. The results showed that more than half of the QRCD cases would have more complete and potentially accurate injury nature and severity information added to them by linking to hospital data. This added injury information has the potential benefit of better representing the serious road crash injury problem than current practice. As mentioned earlier, eARF does not include information about injury severity, so in the interest of parsimony, it may not be included in linkage for the purpose of serious injury classification (as opposed to under-reporting estimates).

An additional benefit of using linked data surrounds the validity of the health data sources in identifying road crash injuries. Combined with the results of Study 2, there are some issues with the identification of relevant cases, particularly with those data collections (e.g., EDIS) that rely on text searching. It has been suggested as a result of the analyses conducted in this study that using the current method for identifying road crash injury cases may lead to an inaccurate estimation of road crash cases. In addition, it was shown that the classification of road users, particularly for some data collections (i.e., EDIS and eARF) was also problematic. Specifically, it was found that motorcyclists and cyclists may be easier to identify in text and therefore some of the bias in under-reporting may be somewhat exaggerated.

8.3 Limitations

A limitation of this research is that it cannot be determined how many of the non-linked cases in Study 3 were due to linkage errors rather than true non-links. While the Queensland Health DLU commented that they thought the quality of linkage was very high, specificity and sensitivity were not able to be calculated due to a large number of manual reviews. Also, the DLU suggested that any errors in the linkage were more likely to be the rejection of links that did exist. As a result, while probably only affecting a relatively small amount of cases, it is possible that some of the non-links are due to linkage error rather than true non-links. On a related issue, despite attempts being made to explore the issue of misclassification in the form of validity analyses, it was still not possible to exactly quantify how much of the misclassification of cases and/or variables influenced discordance rates. Further research into this issue is required to tease out the relative influence of these factors.

Also, in order to identify cases and classify variables such as road users, this study used methods (text terms, coding practices) that are commonly used in research of this nature.

However, these methods, as highlighted to some extent by the validity analyses, may not be sufficiently accurate in identifying and classifying cases. It is possible that these methods could be improved through more elaborate search and/or data mining tools and/or techniques that are increasingly being applied in this type of research.

As mentioned previously (see Section 7.5.2) there may be limitations with the use of AIS and SRR. Firstly, the mapping for AIS involved the extrapolation from ICD-10 to ICD-10-AM and this is at a level less specific than would otherwise be the case. As a result the reliability of the assignment of AIS may be in question. There were also issues in terms of there still being a small number of cases in the hospital data collections that were unable to be mapped due to lack of specificity in the coding. It has also been suggested that the use of a single diagnosis SRR is not ideal and that other injury severity indicators such as DALYs and length of stay could be utilised in further research.

The final limitation of this research is that it did not include all the possible data collections that could potentially hold information or cases relating to road crash injuries. As discussed in the previous chapter (see Section 7.5.2), it is possible that the Queensland Motor Accident Insurance Commission (MAIC) data or the Queensland Trauma Registry (QTR) data could have been included. However, these data collections would only include a subset of the cases included in the data collections that were included. It was also not expected that they would add a significant amount of extra information about the incident or the injury itself. In addition, the QTR data collection has been discontinued and therefore it was considered unnecessary to explore the feasibility of this data collection in future linkage research.

8.4 Implications for Road Safety

The results of this program of research have important implications for the use of data in road safety. The QRCD has a lot of relevant information and includes all of the Core MDS, Core ODS, and Supplemental variables as well as the vast majority of other recommended data items. In addition, there have been no major changes to QRCD over the past 10 years that would in principle have impacted on the reporting of injuries. This would suggest that while the data may not be entirely accurate or complete, any inaccuracies or incompleteness should be relatively consistent over time. Therefore, while not necessarily being an accurate representation of all road crash injuries, the police data is arguably reliable enough to establish trends in the data, which would allow researchers and decision makers to be confident that any changes in the number of road crash injuries represent actual changes. This is obviously important for having confidence in the data for road safety evaluation and monitoring purposes.

Each of the data collections are able to be accessed by researchers and other external agencies for the purposes of research and/or statistical analysis. They are each available in an electronic unit record format which would allow for the analysis using any common spreadsheet or statistical package. While only NCIS and QRCD are available in a web-based format (making them very high on accessibility), each of the other data collections would be considered at least high on accessibility for road safety research, policy development, and evaluation purposes. However, the process in which access is gained can be time consuming and perhaps could add considerably to the delays described above for data to be available in the first place. These issues can have impact on the recency of published research findings and on the ability for researchers and policy makers to identify emerging problems in a timely manner.

This program of research has highlighted that a reliance on police reported crash data, particularly for serious injuries, is problematic. Firstly, Study 2 showed that using the broad classification of fatal, ‘hospitalised’, and other injury, while complete, is lacking in precision. Specifically, the category of ‘hospitalised’ (which is currently used to define serious injuries and is based on whether police identify the person was taken to hospital) is very broad in its range as determined by more objective measures of severity, such as Survival Risk Ratio (SRR). Also, based on SRR, ‘other injury’ (which is currently used to define non-serious injuries) had a lower median SRR (indicating a greater level of serious injury) than ‘hospitalised’ cases. It seems then that the use of this broad severity measure (which is currently the case) is potentially a misrepresentation of the true seriousness of cases.

In addition, the validity analysis in Study 3 demonstrated that using the police defined measure for the counting of serious injuries is likely resulting in an inaccurate, or at least incomplete, picture of the serious road crash injury problem. This has important implications for the monitoring of road safety improvements, since a serious injury reduction target is included in the current National Road Safety Strategy (Australian Transport Council, 2011). There are a number of ways in which the reporting of serious injuries in police data could be improved. These include more specific training of police in identifying injuries, better liaising between police, ambulance, and hospital staff, as well as improved systems for reporting. While these approaches may tighten the interpretation of hospitalisation or the receipt of medical treatment (by confirming these details with ambulance and hospital) it would still not produce the specific serious injury information that is required (e.g., AIS, ICISS). Also, these options are resource intensive and could be prohibitive given other demands on police officers in investigating road crashes as well as their other police duties. In addition, there may be ethical or legislative constraints for police officers obtaining specific information about patient treatment from ambulance services or hospitals. In the future there may be system and legislative advances to allow for automated ‘cross-checking’ of an injured persons’ status, however under the current system operating across Australia this would not be possible. As a result, data linkage may be a good solution at least for the foreseeable future.

The other major issue with the police data relates to the under-reporting of cases. Study 3 showed that there is a substantial level of under-reporting of road crash injuries to police that is similar to the level of discordance found in other studies (Alsop & Langley, 2001;

Amoros et al., 2006; Boufous et al., 2008). It has also confirmed the pattern of under-reporting found elsewhere in terms of bias towards certain types of road users (i.e., cyclists and motorcyclists). These results could greatly impact on road safety research and policy. An accurate representation of the road crash injury problem in terms of severity and prevalence is essential for: prioritising funding and resources; targeting road safety interventions into areas of higher risk such as in different road user groups or locations (urban/rural); calculating the cost of road crash injuries in order to estimate the burden of road crash injuries in terms of future disability; and calculating the cost/ benefit ratio for evaluating interventions aimed at reducing road crash injuries.

This program of research has also determined that there are some limitations in regards to the use of the health data collections. A major issue relates to the reliable identification of road crash injuries. As a result, it is possible that any estimates of under-reporting to police both overall and for particular road user groups may be over-estimated. This needs to be taken into account in future research and any reporting practices that may rely on these health data sources.

In addition, the health data sources are lacking in key data elements that would be essential for thorough examination of road safety issues and evaluation. For example, the health data collections lack a specific location of where the injury took place, or any information on specific circumstances (e.g., speed, fatigue). They also lack information on other crash or road user characteristics (e.g., road environment, seating position, licence status) outlined in the minimum road crash data requirements (Austroads, 1997; MMUCC, 2012; WHO, 2010). As a result, although injuries not reported to police can be identified, for those cases that do not link to QRCD, information relating to the circumstances (e.g., speeding, location) would remain unknown.

Despite these limitations, the benefits of using these data collections in road safety research appear substantial. The health data collections contain information about road crash cases not reported to police and contain much more detailed and complete information about injury nature and severity. Both of these information gains have distinct benefits for understanding the nature of the road crash injuries and their related costs. While the information about the circumstances of the injuries that are not reported to police may be scarce, there is enough information relating to road users, general location, age, gender, and injury severity to provide a snapshot of those cases the police may be missing.

In terms of augmentation of the police data with injury severity information, it should be noted that any improvements would only apply to those police-reported cases that linked. However, when the profiles of the linked police-reported cases and all police -reported cases were compared in Study 3, there was very little difference. This suggests that research using linked data would not introduce any systematic bias since it still provides a similar pattern of road crash injuries (e.g., mostly drivers and passengers in Major Cities) to using police data alone. However, it would provide greater information about injury treatment and associated outcomes for those cases. This would allow for a more precise and reliable measures of injury severity to be applied to police-reported road crash injuries including confirming the hospitalisation status of an injury, as well as the calculation of length of stay, threat to life, and disability indicators.

In terms of practicalities of conducting data linkage, while it did take a considerable time to gain approval and for the data linkage to be completed, many of these issues were due to this being the first study of its kind in Queensland. Now that agreements are in place and the method has been established, it would be arguably easier and less time consuming to conduct in the future. However, it still may not be feasible to conduct linkage frequently or at least often enough to be part of annual reporting practices, as some aspects of the time taken would still apply (e.g., ethics, custodian approval, manual reviews). Also, issues surrounding resourcing would still be a factor. There are limited numbers of people with the skills required to conduct linkage and as noted in Chapter 4 (see Section 4.5.2) it may continue to be difficult for organisations to source skilled staff for linkage work. In addition, from a research perspective, managing and analysing linked data is complex and requires specific skills and knowledge that would need to be considered if research using linked data were to become routine. As an alternative, data linkage could be performed periodically to check the discordance and biases and make adjustments accordingly. This would be in-line with recommendations made by the World Health Organisation (2010), which suggested conducting linkage studies periodically to assess police classification of injury severity against measures such as the Abbreviated Injury Scale (AIS). WHO (2010) also suggests applying a standard methodology to assess under-reporting in police data and apply conversion factors to police road crash injury data to provide a more accurate estimate.

It is possible that this linkage could be restricted to the police data and those collections that have the most relevance and/or are the most accurate (e.g., only QHAPDC for hospitalised injuries). Specifically, linkage with admitted patients' data could be conducted more routinely to confirm the hospitalisation status of a police-reported road crash injury, which would be a good first step to improving serious injury reporting. Ultimately, data linkage could potentially improve the reporting practices and epidemiological study in road safety. While further research is needed to better quantify the discrepancies, data linkage could be used to develop reliable and valid adjustments to the current reporting arrangements. While it is unlikely that non-fatal injury data will ever be as accurate and reliable as fatal data; data linkage could be used to make substantial improvements. It should be noted however, that there may still be barriers from a custodian and/or agency perspective in terms of concerns surrounding the impact of using linked data in their reporting practices. Results of the interviews as part of Study 1b indicated there was a concern that it would create a break in series and could be difficult to explain the change to the public, as well as data users.

While this program of research was conducted using Queensland data, the results do have national and international implications. As discussed earlier, (see Section 1.2), if the ultimate aim is to create an integrated national data linkage system, as researchers in the area suggest (Holman et al., 2008; Turner, 2008), then it is important to understand the nature of each jurisdiction's information systems and data linkage capabilities. This research has provided a detailed exploration of the data quality and data linkage capabilities in Queensland and therefore informs any future national approach. Also, in light of the National Road Safety Strategy (Australian Transport Council, 2011) emphasising a serious injury reduction target, in addition to fatalities, it is necessary to gain an understanding of current practice and potential for improvement of serious injury definitions and reporting across the different jurisdictions that report nationally. Also, in a recent Victorian Parliamentary Inquiry into Serious Injury (2014), a key recommendation was to conduct data linkage with road crash data in Victoria in to order improve the usefulness of road crash data specifically in terms of serious injury reporting. On an international level, the World Health Organization's Global Status Report on Road Safety (2009) also highlights the need to explore ways to improve current road crash injury data collection in terms of under-reporting and serious injury classification. WHO (2009) recommend that data linkage between police, transport, and health data be explored in jurisdictions around the world to better represent the global burden of road trauma. This program of research has demonstrated the issues with and potential improvements to the current Queensland approach and it is possible that the methodology utilised in this research could be replicated in other Australian states and territories, as well as other countries that have not as yet performed this task.

8.5 Implications for Road Crash Injury Surveillance

The results of this program of research have also shown data quality issues with the health data collections which have implications for the surveillance of road crash injuries. Studies 2 and 3 showed that there are limitations with the health data collections particularly in relation to the identification of the relevant cases. This was not just in terms of the validity of the selections, which varied across data collections, but also in relation to the ease in which these selections can be made. For example, EDIS only has text descriptions to determine whether a case is road crash and often does not include specific reference to any information to enable an understanding of the circumstances or nature of an injury incident (e.g., almost 40% of manually reviewed cases lacked specific information to code road user type). QHAPDC, QISU, and eARF are better in this regard as they have coded variables. However, there were still some cases where information was either missing or unspecified for some key variables (such as place), that could impact on the quality of data selections. In addition, it was also demonstrated that the missing or non-specific information varied according to some key aspects of the injury or injured person (e.g., more 'unspecified' *traffic* cases for males in QHAPDC; more 'unknown' *final assessment* cases for drivers and the very young in eARF; more 'unspecified' *place* cases for males and young people in QISU; and more 'unspecified' road user cases for females and young people in EDIS). It is not clear what the underlying reason for these inconsistencies is. Nonetheless, it is important to note their impact on the conclusions drawn when using these data. The inconsistencies could introduce a bias if used for road crash injury surveillance. The validity issues found in this program of research go beyond the Queensland data collections included. There are also implications in other jurisdictions both within Australia and overseas. Hospital data collections in other jurisdictions use the same coding conventions (e.g., ICD), which are likely to be affected by similar validity issues as those found here. Also, generally, emergency department and ambulance data rely on the use of text for identification of cases, which as demonstrated in this research, also have issues with the identification and coding of injury cases. The use of data linkage to examine data quality of data collections has not often been reported in previous research. While also not the focus of this program of research, it has demonstrated that there are some key quality issues (i.e., in relation to the validity of the selection of cases and classification of road users) and shown the potential utility of using linked data specifically for this purpose.

8.6 Suggestions for Further Research

An important issue requiring further research would be to use data linkage to examine specific road safety issues in detail. For example, the results of this study indicate that there may be a significant under-reporting issue with cyclists and motorcyclists. This could be explored in more detail to ascertain what circumstances may lead to these road users being under-reported. Data linkage could be used in conjunction with other data collection methods (e.g., self-report) to examine the matter in more detail.

Another area of interest could be work-related road crash injuries. Some of the data users interviewed in Study 1b identified a lack of reliable identification of work-related crashes in the current data. In the police data there is no dedicated variable that captures the work-relatedness of a crash. There is a variable that relates to the commercial use of a vehicle, however, it is possible that these vehicles are not always used for work purposes and that 'private' vehicles would also be used for work-related travel (particularly to and from work). There are variables included in the health data collections relating to activity at the time of an injury that could prove useful in determining the work-relatedness of a road crash injury. There is also a variable in QHAPDC that specifies the payment method for an episode which includes 'work cover' (the workers' compensation scheme in Queensland) which could also be an indication. As such, future research could explore the use of linked data to specifically examine work-related crashes. This linkage could possibly go beyond the current data collections and include data from workers' compensation and/or work place health and safety sources. There may be other 'case study' data linkage projects that could also be conducted including examining alcohol-involvement and rural and remote crashes.

While this research has already highlighted quality issues with the health data collections, further research is required to better understand the scope and nature of this problem. The linkage between the health data collections could provide information on other coding or classification errors within these data collections. While some examples of the influences on coding errors have been presented in Study 3 (see Section 7.4.7), further work is required to quantify this more precisely. For example, linkage could provide information on the accuracy of triage coding or the differences between ED injury coding and admitted patient injury coding. While this may not have direct benefit for road safety research, the implications for injury surveillance generally would be of interest. In addition, a comprehensive study comparing medical records with coded data would also be useful. This would provide greater insight into the reasons for the lack of specific information in the data either due to coder error or the lack of information in the medical records. This would assist in understanding whether more effort is required to enhance coding practices or record keeping systems to fundamentally improve the collection of road crash injury information in hospital data.

Another possible future study could be in conducting a cost-benefit analysis for data linkage in road safety. While some of the barriers and benefits have been identified in the current work, more detailed study could be conducted to quantify the costs involved in conducting linkage as well as any cost savings. This may be particularly important if routine linkage were to be conducted in the future.

Further research could also be conducted to refine the selection criteria and coding of the health data collections so as to better quantify the discordance and bias found in the current research. Results have shown that traffic coding in hospital data may not always be accurate and taking into account variables such as place may provide more accurate identification of cases. It is also possible that improvements could be made with more elaborate search and/or data mining tools and/or techniques (e.g., weighting and contingent searching algorithms) that are increasingly being applied in this type of research (Erraguntla, Gopal, Ramachandran, & Mayer, 2012; McKenzie, Scott, Campbell, & McClure, 2009). If the selection and coding could be refined, the discordance and bias estimates calculated using linked data could be applied to the police data as an adjustment for reporting purposes.

8.7 Conclusion

This program of research demonstrated that data linkage is possible in the Queensland context and that there are benefits to road safety research and policy making by conducting periodic linkage. It has shown how data linkage can be used to highlight issues of data quality particularly in relation to defining serious injury and the under-reporting of road crash injuries to police. In addition, it has shown that by linking other data sources with police data, improvements to reporting and the classification of serious injury can be achieved by augmenting these data with hospital data collections. Specifically, police data could be linked to admitted patients' data to confirm the hospitalisation status of a case, AIS and SRR could be mapped to police data cases using hospital data to provide a more precise and/or objective measure of injury severity, and adjustments could be made to reporting on the basis of cases not captured in the police data to better represent certain groups such as cyclists and motorcyclists.

This program of research has also shown that some caution is needed in assuming that the health data collections include all relevant cases and that these cases are always accurately identified. Further research on this issue is required, including the refinement of the methods used to identify cases and classify road users in these data. It is also possible that data linkage in the future could restrict the data collections linked with the police data to those that are relevant to the purpose of use and have the most accurate information.

Overall, the program of research has shown how the methodology applied here could be utilised in other jurisdictions. It has also demonstrated the potential improvements to the understanding of the road safety problem, particularly serious injury, by conducting data linkage. Even if linkage was not performed routinely, further research could be conducted to develop adjustments based on linked data, which could then be applied routinely to current reporting, for a more accurate representation of the road safety problem.

References

Alsop, J., & Langley, J. (2001). Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis & Prevention, 33*(3), 353-359.

Amoros, E., Martin, J.-L., Chiron, M., & Laumon, B. (2007). Road crash casualties: characteristics of police injury severity misclassification. *The Journal of Trauma and Acute Care Surgery, 62*(2), 482-490.

Amoros, E., Martin, J.-L., & Laumon, B. (2006). Under-reporting of road crash casualties in France. *Accident Analysis & Prevention, 38*(4), 627-635.

Aptel, I., Salmi, L. R., Masson, F., Bourdé, A., Henrion, G., & Erny, P. (1999). Road accident statistics: discrepancies between police and hospital data in a French island. *Accident Analysis & Prevention, 31*(1), 101-108.

Aron, A., & Aron, E. (1991). *Statistics for psychology* (Second ed.). Upper Saddle River, New Jersey: Prentice Hall.

Association for the Advancement of Automotive Medicine. (2008). *AAAM Abbreviated Injury Scale 2005 update 2008.* . Barrington, Illinois: Association for the Advancement of Automative Medicine

Australian Bureau of Statistics. (2009). *ABS Data Quality Framework*. Canberra: Australian Bureau of Statistics.

Australian Transport Council. (2011). *National Road Safety Strategy, 2011-2020*: ATSB.

Austroads. (1997). A Minimum Common Dataset for the Reporting of Crashes on Australian Roads: Austroads: Austroads.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health, 25*(5), 464-469.

Berry, J. G., Harrison, J. E., & Bureau, A. T. S. (2008). *Serious injury due to land transport accidents, Australia, 2005-06*: Australian Institute of Health and Welfare and the Department of Infrastructure, Transport, Regional Development and Local Government.

BITRE. (2010). Cost of Road Crashes In Australia 2006. Canberra, Australia: Bureau of Infrastructure, Transport, and Regional Economics.

Boufous, S., Finch, C., Hayen, A., & Williamson, A. (2008). *Data linkage of hospital and Police crash datasets in NSW*: University of New South Wales.

Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., Randall, S. M., & Semmens, J. B. (2012). Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Services Research, 12*(1), 480.

Butchart, A., Peden, M., Matzopoulos, R., Phillips, R., Burrows, S., Bhagwandin, N., . . . Cooper, G. (2001). The South African national non-natural mortality surveillance system: rationale, pilot results and evaluation. *S Afr Med J, 91*(5), 408-417.

Cairney, P. (2005). *The prospects for integrated road safety management in Australia: a national overview*.

Cercarelli, L. R., Rosman, D., & Ryan, G. (1996). Comparison of accident and emergency with police road injury data. *The Journal of Trauma and Acute Care Surgery, 40*(5), 805-809.

Chapman, A., & Rosman, D. (2008). *Measuring road crash injury severity in Western Australia using ICISS methodology*. Paper presented at the Insurance Commision of Western Australia Road Safety Forum and Awards. Perth.

Corrao, G., Bagnardi, V., Vittadini, G., & Favilli, S. (2000). Capture-recapture methods to size alcohol related problems in a population. *Journal of epidemiology and community health, 54*(8), 603-610.

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.(Statistical Data Included). *Injury Prevention, 7*(3), 234.

D'Elia, A., & Newstead, S. V. (2010). *De-identified Linkage of Victorian Injury Data Records: A Feasibility Study*: Monash University Accident Research Centre.

Davie, G., Langley, J., Samaranayaka, A., & Wetherspoon, M. (2008). Accuracy of injury coding under ICD-10-AM for New Zealand public hospital discharges. *Injury Prevention, 14*(5), 319-323.

Davie, G., Samaranayaka, A., Langley, J. D., & Barson, D. (2011). Estimating person-based injury incidence: accuracy of an algorithm to identify readmissions from hospital discharge data. *Injury Prevention, 17*(5), 338-342.

Elsenaar, P., & Abouraad, S. (2005). Road Safety Best Practices. *Global Road Safety Partnership*.

Erraguntla, M., Gopal, B., Ramachandran, S., & Mayer, R. (2012). *Inference of missing ICD 9 codes using text mining and nearest neighbor techniques.* Paper presented at the System Science (HICSS), 2012 45th Hawaii International Conference on.

Farmer, C. M. (2003). Reliability of police-reported information for determining crash and injury severity.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183-1210.

Ferrante, A. (2008). Use of Data-Linkage Methods in Criminal Justice Research: A Commentary on Progress, Problems and Future Possibilities, The. *Current Issues Crim. Just., 20*, 378.

Fox, J., Stahlsmith, L., Remington, P., Tymus, T., & Hargarten, S. (1998). The Wisconsin firearm-related injury surveillance system. *American journal of preventive medicine, 15*(3), 101-108.

German, R. R., Lee, L., Horan, J., Milstein, R., Pertowski, C., & Waller, M. (2001). Updated guidelines for evaluating public health surveillance systems. *MMWR, 50*, 1-35.

Glasson, E. J., & Hussain, R. (2008). Linked data: Opportunities and challenges in disability research. *Journal of Intellectual & Developmental Disability, 33*(4), 285-291. doi: 10.1080/13668250802441409

Goldacre, M., & Glover, J. (2002). *The value of linked data for policy development, strategic planning, clinical practice and public health–an international perspective.* Paper presented at the Symposium on Health Data Linkage.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, 205-220.

Health and Hospitals Network Act, Qld (2011).

Health Records Act, Vic (2001).

Henley, G., & Harrison, J. E. (2009). *Injury severity scaling: A comparison of methods for measurement of injury severity*: Australian Institute of Health and Welfare.

Henley, G., & Harrison, J. E. (2011). *Trends in serious injury due to land transport accidents, Australia*.

Holder, Y., Peden, M., Krug, E., Lund, J., Gururaj, G., & Kobusingye, O. (2001). *Injury surveillance guidelines*: World Health Organization Geneva.

Holman, C. D. A. J., Bass, A. J., Rosman, D. L., Smith, M. B., Semmens, J. B., Glasson, E. J., . . . Stanley, F. J. (2008). A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Australian Health Review, 32*(4), 20-20.

Hook, E. B., & Regal, R. R. (1995). Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic reviews, 17*(2), 243-264.

Hook, E. B., & Regal, R. R. (2000). Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *American journal of epidemiology, 152*(8), 771-779.

Horan, J. M., & Mallonee, S. (2003). Injury surveillance. *Epidemiologic reviews, 25*(1), 24-42.

Hunt, P., Hackman, H., Berenholz, G., McKeown, L., Davis, L., & Ozonoff, V. (2007). Completeness and accuracy of International Classification of Disease (ICD) external cause of injury codes in emergency department electronic data. *Injury Prevention, 13*(6), 422-425.

Information Privacy Act, Vic (2000).

Information Privacy Act, Qld (2009).

International Traffic Safety Data and Analysis Group (IRTAD). (2011). Reporting on Serious Road Traffic Casualties: Combining and using different data sources to improve understanding of non-fatal road traffic crashes.

Johnson, R. L., Gabella, B. A., Gerhart, K. A., McCray, J., Menconi, J. C., & Whiteneck, G. G. (1997). Evaluating sources of traumatic spinal cord injury surveillance data in Colorado. *American journal of epidemiology, 146*(3), 266-272.

Klevens, R. M., Fleming, P. L., Li, J., Gaines, C. G., Gallagher, K., Schwarcz, S., . . . Ward, J. W. (2001). The completeness, validity, and timeliness of AIDS surveillance data. *Annals of epidemiology, 11*(7), 443-449.

Langley, J., & Cryer, C. (2012). A consideration of severity is sufficient to focus our prevention efforts. *Injury Prevention, 18*(2), 73-74.

Langley, J., Dow, N., Stephenson, S., & Kypri, K. (2003). Missing cyclists. *Injury Prevention, 9*(4), 376-379. doi: 10.1136/ip.9.4.376

Langley, J., Stephenson, S., & Cryer, C. (2003). Measuring road traffic safety performance: monitoring trends in nonfatal injury. *Traffic Injury Prevention, 4*(4), 291-296.

Langley, J., Stephenson, S., Thorpe, C., & Davie, G. (2006). Accuracy of injury coding under ICD-9 for New Zealand public hospital discharges. *Injury Prevention, 12*(1), 58-61.

Logan, M., & McShane, P. (2006). *Emerging crash trend analysis.* Paper presented at the Proceedings of the Australasian road safety research, policing and education conference.

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. *Australian and New Zealand Journal of Public Health, 32*(1), 28-33.

McDonald, G., Davie, G., & Langley, J. (2009). Validity of Police-Reported Information on Injury Severity for Those Hospitalized from Motor Vehicle Traffic Crashes. *Traffic Injury Prevention, 10*(2), 184 - 190.

McKenzie, K., & McClure, R. J. (2010). Sources of coding discrepancies in injury morbidity data: implications for injury surveillance. *International Journal of Injury Control and Safety Promotion, 17*(1), 53-60.

McKenzie, K., Scott, D. A., Campbell, M., & McClure, R. J. (2009). The use of narrative text for injury surveillance research: A systematic review. *Accident Analysis & Prevention, 42*(2), 354-363.

Meuleners, L. B., Lee, A. H., Cercarelli, L. R., & Legge, M. (2006). Estimating crashes involving heavy vehicles in Western Australia, 1999–2000: a capture–recapture method. *Accident Analysis & Prevention, 38*(1), 170-174.

Miller, T. R., Gibson, R., Zaloshnja, E., Blincoe, L. J., Kindelberger, J., Strashny, A., . . . Sperry, S. (2012). *Underreporting of driver alcohol involvement in United States police and hospital records: capture-recapture estimates.* Paper presented at the Annals of Advances in Automotive Medicine/Annual Scientific Conference.

Mitchell, R., Williamson, A., & O'Connor, R. (2009). The development of an evaluation framework for injury surveillance systems. *BMC Public Health, 9*(1), 260.

National Centre for Classification in Health. (2004). *The International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Australian Modification (ICD-10-CM).: Alphabetic index of procedures (ACHI).* Sydney: National Centre for Classification in Health.

National Collaborative Research Infrastructure Strategy. (2008). *2008 Strategic Roadmap for Australian Research Infrastructure*.

National Highway Traffic Safety Administration. (1998a). *Model minimum uniform crash criteria*. Washington (DC).

National Highway Traffic Safety Administration. (1998b). Traffic Records: A Highway Safety Program Advisory: National Highway Traffic Safety Administration.

Newcombe, H., Kennedy, J., Axford, S., & James, A. (1959). *Automatic linkage of vital records*.

Osler, T., Rutledge, R., Deis, J., & Bedrick, E. (1996). ICISS: an international classification of disease-9 based injury severity score. *The Journal of Trauma and Acute Care Surgery, 41*(3), 380-388.

Private Health Facilities Act, Qld (1999).

Productivity Commission. (2013). Annual Report 20012-13. *Annual Report Series*.

Public Health Act, Qld (2005).

Queensland Health. (2012). *Queensland Hospital Admitted Patient Data Collection (QHAPDC)Manual* Retrieved from http://www.health.qld.gov.au/hsu/pdf/manuals/qhapdc12-13/12_13_QHAPDC_FINAL.pdf.

Rosman, D. L., & Knuiman, M. W. (1994). A comparison of hospital and police road injury data. *Accident Analysis & Prevention, 26*(2), 215-222.

Stephenson, S. C., Langley, J., Henley, G. I., & Harrison, J. E. (2003). *Diagnosis-based injury severity scaling: a method using Australian and New Zealand hospital data coded to ICD-10-AM*: Australian Institute of Health and Welfare.

Strauss, T., & Geadelmann, L. (2009). Evaluation Framework for the Creation and Analysis of Integrated Spatially-referenced Driver-crash Databases.

Strauss, T., & Lentz, J. (2009). Spatial scale of clustering of motor vehicle crash types and appropriate countermeasures.

The Parliament of Victoria Road Safety Committee. (2014). *Inquiry into Serious Injury*. Retrieved from http://www.parliament.vic.gov.au/images/stories/committees/rsc/serious_injury/RSC_-_INQUIRY_INTO_SERIOUS_INJURY.pdf.

Thomas, A. M., Thygerson, S. M., Merrill, R. M., & Cook, L. J. (2012). Identifying work-related motor vehicle crashes in multiple databases. *Traffic Injury Prevention, 13*(4), 348-354.

Toloo, S., FitzGerald, G., Aitken, P., Ting, J., Tippett, V., & Chu, K. (2011). Emergency health services: demand and service delivery models. Monograph 1: literature review and activity trends.

Transport Operation (Road Use Management) Act, Qld § 77A (1995).

Turner, B. (2008). *Review of best practice in road crash database and analysis system design*. Paper presented at the Australasian Road Safety Research Policing Education Conference, 2008, Adelaide, South Australia, Australia.

Ward, H., Lyons, R., Gabbe, B., Thoreau, R., Pinder, L., & Macey, S. (2010). Road Safety Research Report No. 119 Review of Police Road Casualty Injury Severity Classification–A Feasibility Study. London: Department of Transport

Winkler, W. E. (1999). *The state of record linkage and current research problems*. Paper presented at the Statistical Research Division, US Census Bureau.

World Health Organization. (2009). *Global status report on road safety: time for action*: World Health Organization.

World Health Organization. (2010). Data systems: a road safety manual for decision-makers and practitioners *International Journal of Injury Control and Safety Promotion*. Geneva: World Health Organization.

Yorkston, E., Turner, C., Schluter, P., & McClure, R. (2005). Validity and reliability of responses to a self-report home safety survey designed for use in a community-based child injury prevention programme. *International Journal of Injury Control and Safety Promotion*, *12*(3), 193-196.

Young, D., & Grzebieta, R. (2008). *Analysis of the National Coroner's Information System as a data source for fatal crashes.* Paper presented at the Australasian Road Safety Research Policing Education Conference, Adelaide, South Australia, Australia.

Appendix A – Conference Papers

Watson, Angela, McKenzie, Kirsten, & Watson, Barry C. (2011) Priorities for developing and evaluating data quality characteristics of road crash data in Australia. In *Proceedings of Australasian Road Safety Research, Policing and Education Conference 2011*, Perth Convention and Exhibition Centre, Perth, WA

Priorities for developing and evaluating data quality characteristics of road crash data in Australia

Angela Watson[1], Kirsten McKenzie[2], & Barry Watson[1]

[1]Centre for Accident Research and Road Safety-Queensland, Queensland University of Technology

[2]National Centre for Health Information Research & Training, Queensland University of Technology

Abstract

The National Road Safety Strategy 2011-2020 outlines plans to reduce the burden of road trauma via improvements and interventions relating to safe roads, safe speeds, safe vehicles, and safe people. It also highlights that a key aspect in achieving these goals is the availability of comprehensive data on the issue. The use of data is essential so that more in-depth epidemiologic studies of risk can be conducted as well as to allow effective evaluation of road safety interventions and programs. Before utilising data to evaluate the efficacy of prevention programs it is important for a systematic evaluation of the quality of underlying data sources to be undertaken to ensure any trends which are identified reflect true estimates rather than spurious data effects. However, there has been little scientific work specifically focused on establishing core data quality characteristics pertinent to the road safety field and limited work undertaken to develop methods for evaluating data sources according to these core characteristics. There are a variety of data sources in which traffic-related incidents and resulting injuries are recorded, which are collected for a variety of defined purposes. These include police reports, transport safety databases, emergency department data, hospital morbidity data and mortality data to name a few. However, as these data are collected for specific purposes, each of these data sources suffers from some limitations when seeking to gain a complete picture of the problem. Limitations of current data sources include: delays in data being available, lack of accurate and/or specific location information, and an under-reporting of crashes involving particular road user groups such as cyclists. This paper proposes core data quality characteristics that could be used to systematically assess road crash data sources to provide a standardised approach for evaluating data quality in the road safety field. The potential for data linkage to qualitatively and quantitatively improve the quality and comprehensiveness of road crash data is also discussed.

Keywords: Crash data, data linkage.

Introduction

Injuries resulting from transport-related incidents are a significant public health problem world-wide (WHO, 2004). It is predicted, that unless substantial gains are made in the prevention of crashes, transport-related injuries will become the third ranked global burden of disease and injury by 2020. In Australia, approximately 1600 people are killed on our roads each year. On average, the economic cost of fatal crashes is estimated at $3.87 billion, with injury crashes costing $9.61 billion (BTRE, 2009). In order to reduce the burden of transport-related injuries, there is a need to fully understand the nature and contributing circumstances of crashes and the resulting injuries. The National Road Safety Strategy 2011-2020 (ATC, 2011) outlines plans to reduce the burden of road trauma via improvements and interventions relating to safe roads, safe speeds, safe vehicles, and safe people. It also highlights that a key aspect in achieving these goals is the availability of comprehensive data on the issue. The use of data is essential so that more in-depth epidemiologic studies of risk can be conducted as well as enabling effective evaluation of road safety interventions and programs.

Before utilising data to evaluate the efficacy of prevention programs it is important for a systematic evaluation of the quality of underlying data sources to be undertaken to ensure any trends which are identified reflect true estimates rather than spurious data effects. However, there has been little scientific work specifically focused on establishing core data quality characteristics pertinent to the road safety field and limited work undertaken to develop methods for evaluating data sources according to these core characteristics.

There are a variety of data sources in which transport-related incidents and resulting injuries are recorded. These include police reports, emergency department data, hospital morbidity data, and ambulance data. However, as these data are collected for specific purposes, each suffers from some limitations when seeking to gain a complete picture of the problem. It is generally considered that no single data source is sufficient to examine the issue effectively and as a result, there is increasing interest in data linkage as a possible solution.

However, each agency and jurisdiction has different data systems with unique considerations for linkage and use. If the ultimate aim is to create an integrated national data linkage system (as researchers in the area suggest [Austroads, 2005; Holman, et al., 2008; Turner, 2008]), then it is important to understand the nature of each jurisdiction's information systems and data linkage capabilities. Given the lack of standardisation of data sources, legislation, and data linkage progress, work needs to first be undertaken at an individual jurisdiction level to inform a national (and potentially international) approach.

The aim of this paper is to outline core data quality characteristics pertinent to the road safety field that can be used to assess road crash data sources to provide a standardised approach for evaluating data quality in the road safety field. The potential for data linkage to qualitatively and quantitatively improve the quality and comprehensiveness of road crash data will also be discussed.

Framework for assessing data

To determine if a data source is capable of providing good quality information an assessment is required on any limitations of the collection in relation to its capacity to report on injury. It is also necessary to determine how these limitations may affect the accuracy and validity of conclusions that are able to be drawn from the data (Horan & Mallonee, 2003; Mitchell, Williamson, & O'Connor, 2009; WHO, 2001).

There are a variety of frameworks and guidelines with which data related to injury can be assessed, however to date these haven't been systematically defined in regards to the road safety field (e.g., ABS, 2009; Austroads, 1997; Haddon, 1970; Mitchell et al., 2009, NHTSA, 1998; WHO, 2001). For the purposes of this review, data will be discussed in terms of six core quality characteristics: relevance; completeness; accuracy; consistency; timeliness; and accessibility. These six key data quality characteristics or concepts are described below.

Relevance

Relevance is defined as how well the data meets the needs of users in terms of what is measured, and which population is represented. Relevance is important in order to assess whether the data meets the needs of policy-makers and researchers and must be useful for planning and evaluation purposes (ABS, 2009; ATC, 2011). The needs of different data users are diverse, and what one considers 'relevant' may differ from another user's view. This means that within each record, a wide range of data items is usually needed.

Mitchell et al. (2009) discusses the term usefulness, which is a characteristic that also relates to the relevance of a data collection. Usefulness refers to the ability to: (a) identify new and/or emerging injury mechanisms; (b) monitor injury trends over time; and (c) describe key characteristics of the injured population (i.e. WHO's core minimum data set for injury surveillance).

In order to address the issue of relevance, the World Health Organisation's Injury Surveillance Guidelines recommend dividing injury surveillance data into two main categories (core and supplementary) with each of these then subdivided into 'minimum' and 'optional' data. The core minimum data set (core MDS) contains the least amount of data a viable surveillance system can collect on all injuries and the core optional data set (core ODS) involves information that is not necessary to collect but may be collected, if it is seen as useful and feasible to collect. It is also suggested that the core ODS include a narrative or summary of the incident.

Supplementary data includes any additional data that a surveillance system wishes to collect on specific types of injury such as those that are transport-related. In the case of transport-related injuries, information may include details about the circumstances of an incident (e.g., speeding, fatigue) or about other people involved (even if not injured).

Another issue related to relevance is that of representativeness. In other words, to what extent the data collection represents the population of all transport-related injuries or incidents (Mitchell et al., 2009). In order to draw conclusions on the incidence and distribution of transport-related injury, the data collection would need to include all of these injuries regardless of the type of injury, where the injury occurred, or who was injured. Non-representative data may focus prevention efforts on populations that are not truly at risk and could result in a misdirection of resources (Mitchell et al., 2009).

Most data collections do not include all transport-related injuries, instead only including those that fit a particular scope that is relevant for the collection's purpose. For example, hospital admissions data would only include those transport-related injuries that were serious enough to involve admission to hospital. Data collections based on police reported incidents would also not be representative of the entire injury population, as certain transport-related injuries do not fit the definition for inclusion in these collections (e.g., if the injury does not occur on a public road).

Completeness

Strongly related to the issue of relevance is completeness. Completeness refers to the extent to which all relevant cases, all relevant variables, and all data on a relevant variable are included in the data collection (Mitchell et al., 2009). Firstly, data collections would be considered complete if they detect all cases of transport-related injury they intend to detect by definition (sensitivity) and unlikely to detect those injury events they do not intend to detect (specificity). Mitchell et al. (2009) suggest that if between 76% and 100% of the Core MDS and ODS were included in a data collection, it would rate as 'very high'.

Also, not only should the collection include variables relating to the Core MDS and/or Core ODS, these variables should have minimal missing and/or unknown data for them to be considered complete. Mitchell et al. (2009) suggest that a 'high' level of completeness would exist if less than 5% of data within a specific field is missing. In addition to missing or unknown data, a data collection can lack completeness if there are a large number of unspecified or 'other' specified classifications (Mitchell et al., 2009). Incomplete data can be due to a lack of detailed information required to assign a code or classification, a lack of appropriate codes or classifications, lack of time, or lack of skilled coders (Mitchell et al., 2009; NHTSA, 1998). The impact of incomplete data is that the data collection may not provide enough information to allow for adequate data interpretation and could lead to flawed or biased results and therefore decision making.

Accuracy

Accuracy refers to the degree to which data correctly describe the events or persons they were designed to measure (ABS, 2009). Transport-related injury data need to be accurate in several ways, some specific to a location, and others more general. Location information for engineering purposes demands a very high degree of accuracy (within metres), which is frequently not met (Austroads, 2005; Strauss & Lentz, 2009). If location information is not accurate, a problem location might go undetected, and the nature of a location-specific problem might be difficult to determine due to incomplete data.

One of the main indicators of the safety and operation of the road system is the occurrence of transport-related incidents at different levels of severity. Accurate severity information is important for prioritisation of locations, understanding transport-related incident mechanisms, and for evaluating the effectiveness of interventions or countermeasures. Importantly, police are not necessarily in the best position to judge injury severity, at the point of collection of roadside injury information, with injury severity traditionally defined and measured more comprehensively in the clinical setting.

The accuracy of a data collection, and the variable fields within them, is difficult to assess as there is often no real comprehensive or objective data by which to compare the data to a gold standard. However, the literature does suggest that accuracy may be assessed by determining if certain aspects known to enhance the accuracy of data, such as: standardised coding and/or classification (e.g., ICD, AIS); quality control procedures; and the use of technology (GPS), are present (Mitchell et al., 2009; NHTSA, 1998).

Consistency

Consistency of data refers to their ability to reliably monitor transport-related injuries over time, and compare between characteristics within a data set as well as across other relevant data (ABS, 2009). Ideally, the quality of the data should not vary over time, nor should they vary in quality, by the nature of the event/injury, where or when the event/injury occurred, or who was injured or involved. Essentially, users of the data need to be confident that any changes over time or differences between events/individuals are due to actual changes or differences, not simply due to inconsistencies in the data (NHTSA, 1998; WHO, 2001).

Inconsistencies in the data based on the characteristics of the incident or injury can also occur for a variety of reasons. Firstly, reporting policy, work practices, or coding/classification systems may vary by the location of the incident/injury. An incident occurring in a remote location may not be reported, or a lack of resources in some hospitals may lead to less detailed classification. Besides the location of the incident, certain types of incidents/injuries may be less likely to be reported or coded/classified accurately or adequately. For example, a transport-related incident involving illegal behaviour (e.g., unlicensed driving, alcohol) may not be reported to police to avoid prosecution.

One suggested way of enhancing the consistency of a data collection is the use of uniform classification systems (Mitchell et al., 2009; NHTSA, 1998; WHO, 2001). These systems should include a comprehensive set of standard coding/classification guidelines which should be readily available to personnel assigned the duty of recording, classifying or coding data collections. These personnel should also be specifically trained in the procedures and should refer to the guidelines often. Without this training and available material, personnel could base their coding or classification decisions on their own intuitions, opinions, or preconceived notions (CDC, 2001). It is also necessary that any changes to reporting, classification, and recording should be documented in detail (NHTSA, 1998).

Timeliness

Timeliness refers to the delay between the date an event occurs and the date at which the data become available (ABS, 2009). It is suggested that data should become available for use quickly, however the definition of what is 'quick' may vary between agencies and dependent on the purpose for which the data are to be used (Austroads, 2005). It is crucial that agencies are able to respond rapidly to emerging problems, so that the rapid processing of transport-related incident data to make it available is a key concern. For example, Logan and McShane (2006) noted that clusters of crashes could develop quickly, in just a couple of years. Unless the data become available quickly, techniques aimed at detecting emerging clusters will not be effective. Data also needs to be timely for effective evaluations of countermeasures and interventions (NHTSA, 1998). Mitchell et al. (2009) rates the timeliness of the collection, availability, analysis and dissemination as being of high importance for injury data collections. Specifically, they suggest that if data are disseminated within a month the data collection would rate as 'very high'; one to two years as 'high', and more than two years as 'low'. The NHTSA (1998) suggest that it is preferable for data to be available within 90 days. However, they highlight that some supplemental information could wait longer.

The nature of some sources of data means that not all data items can be entered into the database at once; if the data items that have been completed are withheld until each crash record is complete, timeliness will be affected. For example, blood alcohol concentration (BAC) data cannot be entered until results of the toxicology analysis are made available.

Another factor that could influence the timeliness of data availability is related to resourcing. Specifically, an insufficient number of trained personnel to input, code, analyse and/or interpret the data will likely have a negative impact on the timeliness of the data. It is also the case that the roles of the personnel involved, particularly relating to inputting and coding data, are quite diverse (i.e., police officers, nurses), with their priorities directed toward other, arguably more important, tasks (e.g., patient care). This demand on resources can increase the time taken for data to become available.

There are also trade-offs between the timeliness of the data collected and the level of detail recorded regarding a case, as well as the accuracy, completeness and consistency of the data. While the processes that may be in place for coding, recoding, checking, and cleaning of data improve the consistency and accuracy, it may also then increase the time taken for the data to become available, therefore reducing timeliness.

Accessibility

Accessibility relates to the ease with which data can be accessed, which includes ascertaining its availability and suitability for the purpose at hand (ABS, 2009).

The NHTSA (1998) suggests that data should be readily and easily accessible to policy makers, law enforcement, and for use in road safety research and analysis. The NHTSA (1998) further suggest that data should be available electronically, at a unit record level, provided that safeguards are in place to protect confidentiality and privacy. Mitchell et al. (2009) suggest that if data is accessible to users in unit record format from an internet-based interface or data warehouse, it would rates as ‘very high’ on accessibility. While it may be ideal to have free and easily accessible data, there are a number of issues that can limit accessibility.

Major barriers to accessing data relate to confidentiality and privacy. Even when names and addresses are removed, there is still concern that variables such as age and gender in combination with location and temporal variables can lead to the identification of the person/s involved. Information collected and stored by various government agencies are covered by federal and state privacy legislation. These government agencies may also have their own legislation relating to the collecting, storing and access to data. Due to these legislative requirements, there are stringent processes in place in order to access data.

Legislation, policy, and guidelines can be open to interpretation which can complicate the process of negotiating access with different agencies. Therefore, negotiation processes can be protracted where legislation, policy and guidelines are unclear. Even if the process is straightforward, completing the required documentation and having it considered by the relevant authorities can still be quite time consuming.

Another potential barrier to access relates to the concern that data will be misinterpreted or misreported. This is particularly a concern when data custodians are not confident that end-users of the data are aware of the data constraints, limitations and coding conventions. This issue may potentially be overcome by end-users and data custodians communicating better about the nature of the data, including coding information, scope and limitations, as well as by discussing the reporting of data prior to its release or publication.

A third possible barrier to access lies with the data systems themselves. Some data sets do not have relevant information in a format that is easily quantifiable. For example, data systems which compile long text descriptions or reports make extraction of specific information about an incident or its location difficult and time consuming. Even in the case of data being held in a suitable format, the software used may be difficult to navigate, except for those who are specifically trained. Data may not be easily extracted and exported into a format conventionally used by those who work with data (i.e. Excel, text delimited, SPSS, or Access).

Police collected data

At present, a primary source of data used for transport-related incidents is police collected road crash data. While the exact nature of these data collections differ from one jurisdiction to another, generally they include all crashes that are reported to police, that occur on a road, and involve a death and/or injury or substantial property damage (e.g., vehicle is towed away). These crash records usually include details relating to the crash, casualty, unit, and controller.

There are potential limitations of police reported data related to the nature of the data source. It is possible that some crashes may not be included because they are not reported to the police. There has been research about the possible limitations of police reported data (Alsop & Langley, 2001; Boufous, Finch, Hayen, & Williamson, 2008; Langley, Dow, Stephenson, & Kypri, 2003). All of these studies found that some transport-related injuries were not recorded by the police, and reporting rates varied according to a number of factors including: age, injury severity, number of vehicles involved, road user type (e.g., cyclists), whether or not a collision occurred, and geographic region. The solution may not necessarily involve any changes to the processes of reporting to police. However, it does highlight that if police data is relied on as the sole data source for understanding transport-related crashes, without the use of other data (i.e., hospital data); there is a risk that certain causes of injuries will not receive the resourcing for intervention that is commensurate with the size of the problem.

Other data sources

There are a number of other sources of transport-related injury information collected in the health sector such as admitted patient data, emergency department data and ambulance data. The data are used for a number of purposes including examination of patterns of morbidity and mortality for population health research, patient tracking through services/departments, and enumeration of diagnostic case mixes health service funding and resource allocation. While the nature of the information collected varies with each collection and across jurisdictions, the data generally include: the time and date of treatment, the nature of the injury, whether the injury was sustained via traffic or a non-traffic event, and some details about the nature of the event (including information about the mode of transport of the injured person, the mode of transport of the counterpart vehicle involved and whether the injured person was a passenger or a driver), and patient outcomes (such as length of stay, mode of separation etc.).

Perhaps the biggest limitation of this sort of data is that only transport-related incidents that involve attendance or admission to hospital, or those in which an ambulance was called are included in the data collections. Some injured persons involved in transport-related incidents may not present at hospital or call an ambulance but instead attend a medical clinic for treatment. It is also possible that an injury resulting from a transport-related incident could be attributed to some other cause, as the information on the cause of an injury can be falsely reported by the patient, poorly documented by the clinical staff and/or incorrectly coded after discharge.

It should also be noted that as the primary purpose of the data collection is not for road safety research, there are other important information pertinent to the road safety field which are not included (e.g., contributing factors such as alcohol involvement, speeding, fatigue etc.). The emphasis in these data-sets is on health-specific information such as the nature of the injury, length of hospital stay and the treatment outcomes. There may be very little, and in some cases no information, regarding the location of the incident.

Based on the various purposes of these data and their potential limitations, it is generally considered that no single data source is sufficient to examine the issue of transport-related incidents and resulting injuries effectively. As a result, there is increasing interest in data linkage as a possible solution to enable a more complete understanding of the issues surrounding transport incidents and the injuries resulting from such incidents.

Data linkage

Data linkage involves the bringing together of two or more different data sources that relate to the same individual or event (NCRIS, 2008). In principle, any datasets that contain information about individuals has the potential to be linked. There are two possible methods of data linkage: deterministic and probabilistic. The deterministic method involves the linking of data sets that share a unique identifier or key, while the probabilistic method matches cases based on certain elements of data that may lead to the identification of an event and/or person. It does this by matching cases based on other identifying variables such as name, DOB, gender, and time and date of event (NCRIS, 2008).

Potential benefits of data linkage

There are a number of suggested benefits of using linked data for research, monitoring and policy development (Austroads, 2005; Glasson & Hussain, 2008; Goldacre, 2002; Holman et al., 2008). It is possible that data linkage can result in improvements to data quality by including more cases or variables and increasing accuracy through the detection and correction of errors. It is also argued that data linkage can be cost-effective. By linking pre-existing data to provide additional information and address research questions, there is less need to collect additional data on an ad-hoc basis which can be time consuming and expensive (Goldacre, 2002). A report by Austroads (2005) suggests that investment in linked data systems for road safety would likely lead to more efficient day-to-day operations and easier access to data for decision makers. It was suggested that the linking of databases will greatly increase the value of data sets by allowing the use of data for a wider range of purposes (Austroads, 2005).

Potential barriers to data linkage

The first major barrier relates to issues of privacy and confidentiality that are outlined previously. In order to conduct a record linkage project, a researcher needs to obtain approval from multiple data custodians and human research ethics committees. The time and effort involved in this process may discourage the frequent conduct of record linkage studies. It may also be necessary to involve an appropriate third party (or possibly one of the data custodians) in the data linkage process, as access to the identifying information required for data linkage is more restricted, if not prohibited, for researchers. It is important to note, however, that processes in order to provide linked data to researchers, while safe-guarding privacy, have been established in other Australian jurisdictions as well as overseas.

Another potential barrier is the linkage process itself. The deterministic method is the most accurate method; however it involves a unique identifier being matched across data sets. Unfortunately, in the case of the data sources discussed previously, though information in different data sets may relate to the same incident, person or case, there is no system of unique identifiers across all data sets. Also, in the case of the police data, the unique identifier is often assigned to an event (i.e., the crash), while the unique identifiers within health data sets are typically assigned to a patient.

As such, the probabilistic method is required for linkage of these datasets in the absence of a shared unique identifier. However, this method relies on having specific and accurate information on the relevant linkage variables in both data sets. This method requires that enough data points can be chosen for matching purposes so that no two events or individuals will be confused, leading to a lack of specificity. Conversely, if the data matching criteria is too specific, there is a potential for an individual to not be matched despite them actually being present in both data sets (i.e. lack of sensitivity). So although this method has been utilised in the past in other jurisdictions, a limitation is that the formats used with different data sets may not be compatible, resulting in an inability for some of the data sets to communicate with each other or leading to errors in matching.

Previous data linkage research

In terms of transport-related incidents and injuries, a variety of data linkage projects have been conducted (e.g., Alsop & Langley, 2001; Boufous, et al., 2008; Cercarelli, Rosman, & Ryan, 1996; Langley, et al., 2003). Alsop and Langley (2001) used probabilistic linkage of police and hospital records in New Zealand. They found that less than two-thirds of all hospitalised traffic crash casualties were recorded in the police data. They also found that this varied based on the number of vehicles involved, the geographical location, age and injury severity. Langley, et al. (2003) conducted probabilistic linkage between hospital records and police records to specifically examine the potential under-reporting of cyclist injuries in New Zealand. The results showed that only 22% of cyclists that crashed on a public road could be linked to the police records. Of the crashes that involved a motor vehicle 54% were recorded by police. They also found that age, ethnicity, and injury severity predicted whether a hospitalised cycle crash was more likely to be recorded in the police data. Within Australia, Cercarelli, et al. (1996) linked police reports, hospital admissions and accident and emergency (A&E) department data. The researchers found that around 50% of attendances at the A&E were recorded by police, and that around 50% of cases recorded by police as being admitted to hospital were actually admitted. The researchers outline that while the discrepancy between the data sets does represent an under-reporting of cases, it also suggests that differences in coding systems may also lead to cases not being linked. Another Australian study conducted in NSW by Boufous, et al. (2008) linked hospital admissions data (Inpatient Statistics Collection [ISC]) with the Traffic Accident Data System (TADS). Using probabilistic linkage, the researchers matched 56.2% of hospitalisations coded as being as a result of traffic crash with a record in TADS. The researchers also found that the linkage rate varied according to age (i.e., lower linkage rate for younger age groups), road user type (e.g., lower linkage rate for cyclists), severity (i.e., higher linkage rates with increased severity) and geographical location.

While these studies highlight the issues of under-reporting and bias within police data systems, the barriers and limitations of data linkage were not explored either at all, or in any depth, in any of the studies conducted to date. Also, many of these studies involved the ad-hoc linkage of data as opposed to routine data linkage. It is likely that routine data linkage may involve issues (e.g., changes to data systems, inter-agency agreements) that ad-hoc project based data linkage does not and vice versa. Each jurisdiction has different data systems with unique considerations for linkage and use. If the ultimate aim, as researchers in the area suggest (Austroads, 2005; Holman, et al., 2008; Turner, 2008), is to create an integrated national data linkage system, then it is important to understand the nature of each State and Territory's information systems and data linkage capabilities.

Research priorities

In order to improve the quality, comprehensiveness, and usefulness of transport-related injury data, there are a number of suggested priorities for future research, including: scoping existing data collections in order to assess their completeness, consistency, accuracy, accessibility and relevance; determining the barriers to and facilitators of linking transport-related injury data; and assessing whether linked data provide qualitative and quantitative advantage over non-linked data. These priorities could be addressed by: discussions with data custodians, users, and other key stakeholders; reviewing legislative and policy documents; and analysis of sample data from current traffic injury data sources. While it is important to establish whether data linkage is feasible, it is also necessary to establish whether the benefits that would be derived from linked data would be sufficient to offset the likely costs. This could be achieved by piloting data linkage (including a comparison of linked data with non-linked data) and conducting cost-benefit analysis for both routine and ad- hoc data linkage.

Summary

Data is vital to informing policies and interventions designed to reduce the burden of road trauma. This paper proposes core data quality characteristics to enable the systematic assessment of road crash data sources to provide a standardised approach for evaluating data quality in the road safety field. It is possible that linkage of key data collections has the potential to overcome the limitations of single data sources and maximize the collective benefit of data relating to road trauma. However further research needs to establish whether road safety data linkage is feasible within each jurisdiction (given differences in data linkage capabilities across jurisdictions) and whether linked data provide advantage over non-linked data, both qualitatively and quantitatively.

References

Alsop, J. and Langley, J. (2001). Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis and Prevention*, 33, p.353-359.

Australian Bureau of Statistics (2009). *ABS Data Quality Framework, May 2009*. Australian Bureau of Statistics: Canberra. http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/1520.0Main%20Features1May%202009?opendocument&tabname=Summary&prodno=1520.0&issue=May%202009&num=&view=

Australian Transport Council. (2011). *The Draft National Road Safety Strategy*. Australian Transport Safety Bureau: Canberra. http://www.infrastructure.gov.au/roads/safety/national_road_safety_strategy/files/Draft_National_Road_Safety_Strategy_ext.pdf

Austroads (1997). *A Minimum Common Dataset for the Reporting of Crashes on Australian Roads.* Report No. AP-126/97, Austroads: Sydney.

Austroads (2005). *The Prospects for Integrated Road Safety Management in Australia: A National Overview.* Report No. AP-R280/05, Austroads: Sydney.

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. *Journal of Safety Research,* 39, p.65-72.

Bureau of Infrastructure, Transport and Regional Economics [BITRE] (2009). *Road crash costs in Australia 2006*, Report 118, Canberra, November.

Cercarelli, L., Rosman, D., and Ryan, G. (1996). Comparison of accident and emergency with police road injury data. *The Journal of Trauma*, 40(5), p.805-809.

Connelly, L. and Supangan, R. (2006). The economic costs of road traffic crashes: Australia, states and territories. Accident Analysis and Prevention, 38, p.1087-1093.

Goldacre, M. (2002). The value of linked data for policy development, strategic planning, clinical practice and public health: An international perspective. *Symposium on Health Data Linkage: Its value for Australian health policy* *development and policy relevant research,* March 2002, Potts Point, Sydney, New South Wales.

Glasson, E.J., and Hussain, R. (2008). Linked data: Opportunities and challenges in disability research. Journal of Intellectual and Developmental Disability, 33(4), p.285-291.

Haddon, W. Jr. (1970). A logical framework for categorizing highway safety phenomena and activity. Paper presented at the 10th International study Week in Traffic and Safety Engineering, Rotterdam, 7-11 September.

Holman, C.D., Bass, A.J., Rosman, D.L., Smith, M.B., Semmens, J.B., Glassson, et al. (2008). A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Australian Health Review*, 32(4), p. 766-777.

Horan, J.M., and Mallonee, S. (2003) Injury surveillance. *Epidemiology Review*, 25, p. 24-42.

Langley, J., Dow, N., Stephenson, S., Kypri, K. (2003). Missing Cyclists. *Injury Prevention*, 9, p. 376-379.

Logan, M. and McShane, P. (2006). Emerging crash trend analysis. Proceedings of the Australasian Road Safety Research, Policing and Education Conference, Brisbane, October 2006.

Mitchell, R., Williamson, A., and O'Connor, R. (2009). Development of an evaluation framework for injury surveillance systems. *BMC Public Health*, 9, p.260.

NHTSA (1998). *Traffic Records: A Highway Safety Program Advisory.* National Highway Traffic Safety Administration, http://www.nhtsa.dot.gov/people/perform/pdfs/Advisory.pdf

NCRIS (2008). *Strategic Roadmap for Australian Research Infrastructure*. NCRIS, Department of Innovation, Industry, Science and Research: Canberra. http://ncris.innovation.gov.au/Documents/2008_Roadmap.pdf

Strauss, T. and Lentz, J. (2009). *Spatial Scale of Clustering of Motor Vehicle Crash Types and Appropriate Countermeasures*. MTC Project 2007-10, Midwest Transportation Consortium, Iowa State University Institute for Transportation: Ames.

Turner, B. (2008). Review of best practice in road crash database and analysis system design. *Proceedings of the Australasian Road Safety Research, Policing and Education Conference,* Adelaide, November 2008.

WHO (2004). *The world health report*. World Health Organisation, Geneva, 2004.

WHO (2001). *Injury Surveillance Guidelines*. World Health Organisation, Geneva, 2001.

How Accurate Is The Identification Of Serious Traffic Injuries By Police? The Concordance Between Police And Hospital Reported Traffic Injuries

Watson[a], A., Watson[a], B., & Vallmuur[a], K.

[a] Centre for Accident Research & Road Safety – Queensland (CARRS-Q), Queensland University of Technology (QUT) Kelvin Grove, Queensland

Abstract

Police reported crash data are the primary source of crash information in most jurisdictions. However, the definition of serious injury within police-reported data is not consistent across jurisdictions and may not be accurate. With the Australian National Road Safety Strategy targeting the reduction of serious injuries, there is a greater need to assess the accuracy of the methods used to identify these injuries. A possible source of more accurate information relating to injury severity is hospital data. While other studies have compared police and hospital data to highlight the under-reporting in police-reported data, little attention has been given to the accuracy of the methods used by police to identify serious injuries. The current study aimed to assess how accurate the identification of serious injuries is in police-reported crash data, by comparing the profiles of transport-related injuries in the Queensland Road Crash Database with an aligned sample of data from the Queensland Hospital Admitted Patients Data Collection. Results showed that, while a similar number of traffic injuries were recorded in both data sets, the profile of these injuries was different based on gender, age, location, and road user. The results suggest that the 'hospitalisation' severity category used by police may not reflect true hospitalisations in all cases. Further, it highlights the wide variety of severity levels within 'hospitalised' cases that are not captured by the current police-reported definitions. While a data linkage study is required to confirm these results, they highlight that a reliance on police-reported serious traffic injury data alone could result in inaccurate estimates of the impact and cost of crashes and lead to a misallocation of valuable resources.

Introduction

Police reported crash data are the primary source of crash information in most jurisdictions. However, the definition of serious injury within police-reported data is not consistent across jurisdictions and may not be accurate. With the Australian National Road Safety Strategy (ATC, 2011) targeting the reduction of serious injuries, which was not previously the case, there is a greater need to assess the accuracy of the methods used to identify these injuries. Accurate severity information is important for prioritisation of intervention locations, understanding transport-related incident mechanisms, evaluating the effectiveness of interventions or countermeasures, and the calculation of the cost of crashes. In most Australian jurisdictions, the current classification of severity, and ultimately serious injury, by police is primarily based on process rather than a clinical assessment per se. Injury severity (with the exception of a fatality) is classified based on the extent of medical intervention (i.e., requiring medical treatment, taken or admitted to hospital). In Queensland, this classification is as follows: fatality; hospitalisation (taken to hospital); medical treatment; minor injury; and property damage only. Studies in other jurisdictions (e.g., New Zealand, USA) have shown that categories like these do not always correspond with objective measures relating to threat to life. Fatal cases and those with an absence of injury are generally accurately classified; however, the non-fatal injuries are more likely to be misclassified based on more objective severity measures (Farmer, 2003; McDonald, Davie, & Langley, 2009).

Arguably, it would be more accurate if the severity of an injury was based on clinical information (i.e., the nature of the injury) and involved some sort of assessment of threat to life or permanent disability. However, collecting this clinical information at the roadside particularly by police may not be ideal. Police do not have the training or expertise to record information on the nature of an injury or injuries with the required level of accuracy. Also, the consistency of the recorded information from case to case could be questionable (Ward, Lyons, Gabbe, Thoreau, Pinder, & Macey, 2010).

A possible source of more accurate information relating to injury severity is hospital data. While other studies have compared police and hospital data to highlight the under-reporting in police-reported data, little attention has been given to the accuracy of the methods used by police to identify serious injuries. The current study aimed to, in addition to highlighting the possible under-reporting of crashes to police, assess how accurate the identification of serious injuries is in police-reported crash data. It aimed to do this by comparing the profiles of traffic-related ('hospitalised') injuries in the Queensland Road Crash Database and identified traffic-related injuries in the Queensland Hospital Admitted Patients Data Collection.

Methods

Ethics approval was obtained from the Queensland University of Technology's Human Research Ethics Committee (#1100001065). A Public Health Act agreement was completed by the researcher and signed by Queensland Health. The Queensland Road Crash Database (QRCD) data was provided following approval (via designated form) from the Manager of the Data Analysis Unit at the Department of Transport and Main Roads. Queensland Hospital Admitted Patients Data Collection (QHAPDC) data was provided by the Manager of the Health Statistics Centre at Queensland Health.

Data sources

Queensland Road Crash Database (QRCD)

The QRCD stores information relating to all police reported crashes in Queensland since 1986. The definition of a police reported crash is:

> "a crash that has been reported to the police which resulted from the movement of at least one road vehicle on a road and involving death or injury to any person, or property damage to the value of:
>
> - $2500 to property other than vehicles (after 1 December 1999)
> - $2500 damage to vehicle and property (after 1 December 1991 and prior to 1 December 1999)
> - value of property damage is greater than $1000 (prior to December 1991) or;
> - at least one vehicle was towed away." Department of Transport and Main Roads (2010)

A crash will be excluded from the database, even if it complies with the above definition, if the incident involved deliberate intent (e.g., assault, suicide) or is not attributable to vehicle movement.

Queensland Hospital Admitted Patient Data Collection (QHAPDC)

QHAPDC contains data on all patients separated (an inclusive term meaning discharged, died, transferred or statistically separated) from any hospital permitted to admit patients, including public psychiatric hospitals.

Data specifications

Cases for each data collection were selected based on their alignment with the Queensland Road Crash Data definition of a traffic-related injury (i.e., occurred on a public road and involved a moving vehicle). Where possible, other exclusions based on the definition outlined in Queensland Road Crash Data were also applied (e.g., intentional acts, pedestrian colliding with a railway train). In order to conduct analyses, the following variables were used for each data set:

Age was coded into 5 year age groups (with the exception 85+).

Gender (1 = Male; 2 = Female). Some data sets refer to sex rather than gender, however, gender will be the term used throughout.

Severity of injury was measured by three variables: *Broad severity*, *Abbreviated Injury Scale, and Survival Risk Ratios*.

4. *Broad severity* was coded into three levels (fatality; hospitalisation; other injury). These categories are the basis for how severity is generally captured across jurisdictions. It should be noted that for the purposes of this categorisation, hospitalisation will be treated as 'taken to hospital' as defined by the QRCD.
5. *The Abbreviated Injury Scale (AIS)* is a body-region based coding system developed by the Association for the Advancement of Automotive Medicine (AAAM, 2008). A single injury is classified on a scale from 1-6 (1 = minor; 2 = moderate; 3 = serious; 4 = severe; 5 = critical; and 6 = maximum). If there is not enough information to assign a value, a code of 9 (not specified) is applied. For the purposes of this study, the AIS score was mapped to principal diagnosis International Classification of Diseases (ICD-10-AM) codes in the data

(NCCH, 2008). A tool for mapping ICD codes to AIS score was sourced from the European Center for Injury Prevention.

6. *Survival Risk Ratios (SRR)*, assigned to a single injury, provide an estimate of the probability of death and is based on ICD-10-AM coding, ranging from 0 (no chance of survival) to 1 (100% chance of survival). SRRs were mapped to principal diagnosis ICD codes as used by Stephenson, Henley, Harrison, and Langley (2003). It should be noted that it was not possible to calculate ICISS (ICD Injury Severity Score), which a more comprehensive assessment of injury severity than SRR alone. This was because, to calculate ICISS information on all the injuries a patient suffers requires the calculation of the multiplication of SRRs for each injury and each data set only provided the principal diagnosis.

In order to specifically explore issues of serious injury definitions, three classifications of *serious injuries* were derived:

4. SRRs equal to or less than 0.941 were coded as serious with all other values coded as non-serious. This criterion was based on the work of Cryer and Langely (2006).
5. All those with an AIS of 3 or greater were classified as serious, the rest as non-serious.
6. All those coded as ‘hospitalised’ and fatal were classified as serious, the rest as non-serious.

Accessibility/Remoteness Index of Australia (ARIA+) broadly classifies geographic areas based on their distance from the five nearest major population centres (National Centre for Social Applications of GIS, 2009). ARIA+ is categorised into five groups (1 = Major Cities; 2 = Inner Regional; 3 = Outer Regional; 4 = Remote; 5 = Very Remote). Some of the data sets included ARIA+ classifications, while others provided postcode. In cases where postcode was provided without ARIA+, postcodes were mapped to ARIA+ using data from the Australian Bureau of Statistics. Some postcodes map to multiple ARIA+ categories, so in these cases the postcode is assigned to the ARIA+ category that has the largest proportion of the population.

Road user was coded into five categories (1 = Driver, 2 = Motorcyclist (including pillions), 3 = Cyclist (including pillions), 4 = Pedestrian; 5 = Passenger).

Queensland Road Crash Database (QRCD)

By definition, all injury cases in the QRCD for 2009 were included. However, for the purposes of comparison with QHAPDC, only fatalities and hospitalisations were used.

The coding of variables was as follows:

Age was provided in years, and was coded into 5 year age groups (with the exception of 85+).

Gender was retained as coded (1 = Female; 2 = Male).

Broad severity was coded from the variable *casualty severity* (1= fatality; 2 = hospitalisation; 3 = medical treatment; 4 = minor injury), with ‘medical treatment’ and ‘minor injury’ collapsed into the ‘other injury’ category.

AIS and *SRR*, was coded using the *injury description* variable. This variable, while a text description, is recorded in a standard form that is the same as those of the ICD-10-AM principal diagnosis descriptions. This allowed a principal diagnosis ICD-10-AM code to be mapped to each injury description. These ICD codes were then mapped to the AIS and a SRR using processes outlined previously.

ARIA+ was an already coded variable in the data, so was retained in its original form. *ARIA+* in this case relates to the location of the crash.

Road user was categorised using the variable *casualty road user type*. The original variable coding was retained from this variable with the exception of ‘motorcycle pillions’ and ‘bicycle pillions’. These two classifications were put into the ‘motorcyclist’ and ‘cyclist’ categories respectively.

Queensland Hospital Admitted Patient Data Collection (QHAPDC)

To select traffic-related injuries for 2009 for comparison to QRCD, the first step involved selecting cases that were coded as being land transport-related. For the QHAPDC collection this included cases with an ICD-10-AM external cause code from V00-V89. Using the fourth character in the ICD-10-AM external cause code to identify whether an incident was traffic or non-traffic, 43,991 (67.8%) of land transport cases were classified as traffic. Other exclusions were also made due to cases not fitting the definition of a road crash. Specifically, when the injury resulted from a pedestrian colliding with a pedestrian conveyance (V00) or a railway train (V05) it was not included. Also, all transfers, as identified by *separation mode* were excluded to partly eliminate multiple counts of cases.

Variables were selected, created and/or recoded as follows:

Age was provided in 5 year age groups (with the exception of 85+).

Gender was retained as coded (1 = Female; 2 = Male).

Broad severity was defined using the *mode of separation* variable, with those coded as ‘died in hospital’ categorised as a fatality and all other cases categorised as ‘hospitalised’.

AIS and *SRR,* was coded using the principal diagnosis ICD-10-AM codes. These ICD codes were then mapped to the AIS and a SRR using processes outlined previously.

ARIA+ was an already coded variable in the data, so was retained in its original form. *ARIA+* in this case relates to the location of the hospital.

Road user was categorised using the second and fourth characters of the ICD-10-AM external cause code.

Data analysis

Data was imported from csv into SPSS 19 for coding and analysis. Comparisons were made using Chi-square tests of independence. Due to the large sample size, a more stringent alpha of .001 was adopted. Also, Cramer's V (ϕ_c) was calculated in order to provide an estimate of effect size to give a clearer idea of the meaningfulness of any statistical significance found. As suggested by Aron and Aron (1991), a Cramer's V of less than .10 was considered to be a small effect size, between .10 and .30 moderate, and more than .30 a large effect size. Post-hoc analyses were also undertaken using an adjusted standardised residual statistic. This statistic can be used to identify those cells with observed frequencies significantly higher or lower than expected. With an alpha level set at .001, adjusted standard residuals outside -3.10 and +3.10 were considered significant (Haberman, 1978).

Results

Overall, in 2009, QHAPDC had 6,725 compared to 7,003 cases in QRCD. In terms of the profile of cases, compared to the QRCD, the QHAPDC had a statistically significantly greater proportion of males, motorcyclists, and cyclists included in the data collection. QHAPDC also had a higher proportion of younger people (14 and younger) [$\chi^2(17) = 125.69$, $p < .001$, $\phi_c = .10$] and a lower proportion of cases in remote or very remote areas compared to QRCD (see Figure 1 and Table 1).

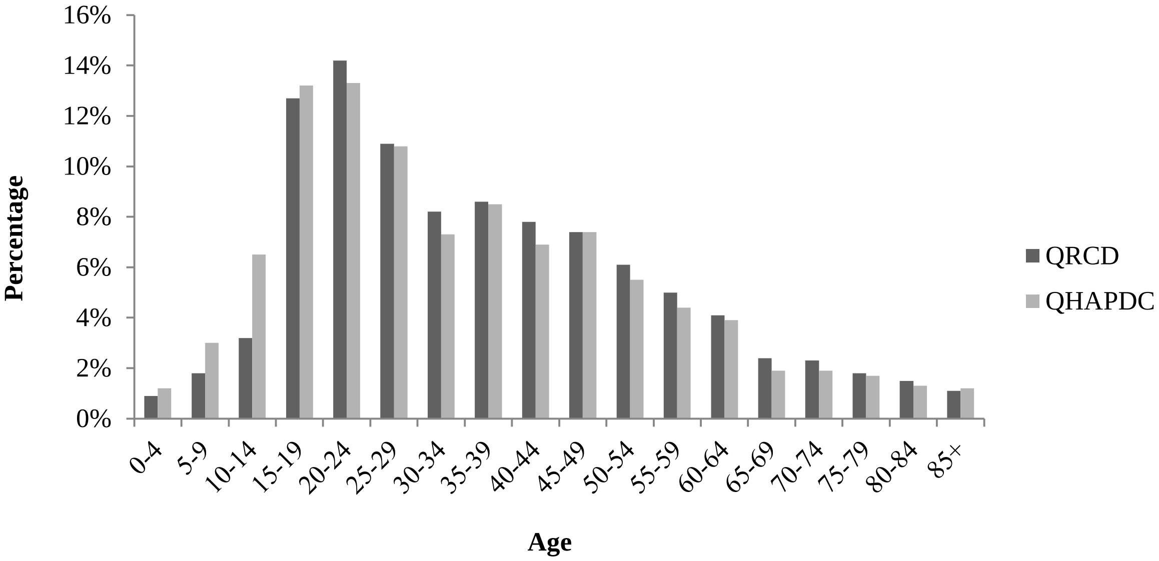


Figure 1. Age distribution of QRCD and QHAPDC for 2009

Table 1. Demographic characteristics by data source for QRCD and QHAPDC 2009

Variable	Level	Data source QRCD n (%)	 QHAPDC n (%)	Significance test
Gender	**Male**	4,039 (57.7)	4,646 (69.1)[1]	
	Female	2,960 (42.3)	2,079 (30.9)	$\chi^2(1) = 191.06$, $p < .001$, $\phi_c = .12$
ARIA+	**Major Cities**	3,611 (51.6)	3,753 (55.8)	
	Inner Regional	1,644 (23.5)	1,745 (25.9)	
	Outer Regional	1,320 (18.9)	1,063 (15.8)	
	Remote	246 (3.5)	116 (1.7)[1]	
	Very Remote	181 (2.6)	48 (0.7)[1]	$\chi^2(4) = 151.87$, $p < .001$, $\phi_c = .11$
Road user	**Driver**	3,723 (53.2)	1,904 (29.5)	
	Motorcyclist	1,015 (14.5)[1]	2,024 (31.4)[1]	
	Cyclist	362 (5.2)[1]	1,067 (16.5)[1]	
	Pedestrian	464 (6.6)	435 (6.7)	
	Passenger	1,439 (20.5)	1,021 (15.8)	$\chi^2(4) = 162.62$, $p < .001$, $\phi_c = .11$

[1] Standardised residuals outside +/- 3.10

In terms of broad severity, not surprisingly, QRCD had a greater proportion of fatalities compared to QHAPDC. Based on AIS, QHAPDC had greater proportion of moderate injuries; however, there was no difference on SRR in terms of the proportion of serious vs. non-serious (see Table 2). However, it should be noted that much greater proportion of the QRCD were unable to be classified, due to the missing injury description data, for either AIS or SRR compared to QHAPDC.

Table 2. Severity profile by data source for QRCD and QHAPDC 2009

Variable	Level	Data source QRCD n (%)	 QHAPDC n (%)	Significance test
Broad severity	**Fatality**	331 (4.7) [1]	71 (1.1)[1]	
	Hospitalisation	6,672 (95.3)	6,654 (98.9)	$\chi^2(1) = 162.62$, $p < .001$, $\phi_c = .11$
Unspecified injury	**Yes**	5,602 (86.5)[1]	31 (0.5) [1]	
	No	1,401 (19.3)	6,694 (99.5)	$\chi^2(1) = 8968.61$, $p < .001$, $\phi_c = .81$
AIS	**Minor**	633 (45.2)	2,037 (34.8)	
	Moderate	424 (30.3)	2,789 (47.7)[1]	
	Serious	342 (24.4)	900 (15.4)	
	Severe	0 (0.0)	89 (1.5)	
	Critical	1 (0.1)	21 (0.4)	
	Maximum	1 (0.1)	16 (0.3)	$\chi^2(5) = 190.46$, $p < .001$, $\phi_c = .16$
SRR	**Serious (< 0.942)**	177 (12.7)	921 (13.8)	

Non-serious (> 0.941)	1,218 (87.3)	5,733 (86.2)	$\chi^2(1) = 1.13$, p = .288, ϕ_c = .01

[1] Standardised residuals outside +/- 3.10

Due to the substantial amount of missing and unspecified data (injury description) in QRCD which was used to calculate AIS and SRR, an analysis was conducted to see if there was any bias based on the broad severity measure. It should be noted that this was conducted on all 2009 cases, including the other injury category.

There was a statistically significant difference in the proportion of unspecified injury descriptions by broad severity [$\chi^2(2) = 1036.9$, $p < .001$, $\phi_c = .23$]. Specifically, the injury description was more likely than expected to be unspecified for hospitalisations and less likely than expected to be unspecified for fatalities (see Table 3).

Table 3. Unspecified injury description by broad severity for QRCD 2009

	Injury description	
	Specified **n (%)**	**Unspecified** **n (%)**
Fatality	300 (90.6)	31 (9.4)[1]
Hospitalisation	1,101 (16.5)[1]	5,571 (83.5)[1]
Other injury	2,755 (22.9)	9,260 (77.1)

[1] Standardised residuals outside +/- 3.10

Table 4 shows the proportion of serious injuries in QRCD based on Broad Severity, AIS, and SRR classification criteria. There were a much larger proportion of serious injuries classified when using the broad severity criteria compared to both AIS and SRR. A total of 38 cases were classified as serious using all three criteria. While the SRR and AIS proportions are quite similar, interestingly, only 40 cases were coded as serious under both AIS and SRR criteria.

Table 4. The number and proportion of serious and non-serious injuries based on the three different severity measure criteria, QRCD 2009

	Broad severity (Fatal and Hospitalised)	**AIS (score of 3 or above)**	**SRR (0.941 or less)**
Serious	7,003 (36.8%)	355 (8.6%)	387 (9.3%)
Non-serious	12,015 (63.2%)	3,788 (91.4%)	3,762 (90.7%)

Table 5 shows the proportion of serious injuries in QHAPDC based on Broad Severity, AIS, and SRR classification criteria. Due to the nature of the data collection (all cases hospitalised or fatality), based on broad severity, all cases are classified as serious. The proportion of serious cases based on AIS was higher than the proportion of serious based on SRR. There were 488 cases coded as serious under both AIS and SRR criteria.

Table 5. The number and proportion of serious and non-serious injuries based on the three different severity measure criteria, QHAPDC 2009

	Broad severity (Fatal and Hospitalised)	AIS (score of 3 or above)	SRR (0.941 or less)
Serious	6,725 (100.0%)	1,026 (17.5%)	921 (13.8%)
Non-serious	0 (100.0%)	4,826 (82.5%)	5,773 (86.2%)

To further explore the broad severity classification, the median of SRRs were calculated for each broad severity category for each data collection. Table 6 shows, for QRCD, that the median SRR was lowest (more severe) for fatalities. Surprisingly, the median SRR for other injury was lower than that of hospitalisations, suggesting that other injuries (medical treatment and minor injuries) are more severe than those cases taken to hospital. This table also shows that the range of severities (as measured by SRR) was quite wide within each broad severity category.

Table 6. Median and range SRR for each broad severity category, QRCD 2009

	Median SRR	Range (min – max)
Fatality	0.940	0.746 – 1.000
Hospitalisation	0.985	0.500 – 1.000
Other injury	0.954	0.554 – 1.000

Table 7 shows, for QHAPDC, that the median SRR was lower (more severe) for fatalities compared to hospitalised cases. The range of severities (as measured by SRR) was quite wide for both fatalities and hospitalisations.

Table 7. Median and range SRR for each broad severity category, QHAPDC 2009

	Median SRR	Range (min – max)
Fatality	0.867	0.306 – 0.996
Hospitalisation	0.991	0.306 – 1.000

Discussion

In terms of overall numbers, the difference between QRCD and QHAPDC was minimal. However, when the profiles were compared, there were significant differences between QRCD and QHAPDC. Specifically, QHAPDC had a greater proportion of males, younger people (aged 0-14), motorcyclists, and cyclists compared to QRCD. These differences provide some evidence of under-reporting for QRCD and that this under-reporting has a bias towards certain injured persons. This under-reporting, specifically including these motorcyclists and cyclists, has been demonstrated in other research in the area (Alsop & Langley, 2001; Cryer et al., 2001; Langley, Dow, et al., 2003).

However, it is also possible that some of the differences are not due to under-reporting, but instead due to misclassification of traffic-related injuries in QHAPDC and/or the lack of precision in the technique for selecting traffic injury cases. It is not clear at this stage how valid QHAPDC coding is in terms of identifying traffic cases and road users. The primary purpose of this data is not for this type of classification, so it is possible that the accuracy of the coding could be compromised. It is also possible that the classification of hospitalised in QRCD is also incorrect. Further research, using data linkage, may quantify the extent of misclassification versus under-reporting.

In addition to the above differences, QHAPDC had a lower proportion of Remote and Very Remote cases based on ARIA+ compared to QRCD. This result is perhaps not surprising considering the classification basis for each collection. QHAPDC ARIA+ relates to the location of the hospital, whereas QRCD ARIA+ relates to the location of the crash. It is likely that even when a crash occurs in a Remote or Very Remote location, the injured person would not necessarily be treated in a hospital in a Remote or Very Remote location due to lack of facilities. Also, excluding transfer cases would select out many cases from facilities in Remote and Very Remote locations, as the patient would likely be transferred to a facility in a less remote location. Ultimately, these differences would bias this comparison somewhat. This bias may have been reduced by selecting out the transfers from the final hospital not the initial hospital (using *Admission Source*). However, this technique can introduce other issues with completeness and reliability and was also not available to the researcher for this study.

For severity, there was no difference between the collections in terms of the proportion classified as serious based on Survival Risk Ratio (SRR). However, QRCD had a greater proportion of fatalities and serious or worse AIS classification compared to QHAPDC. The difference between the collections in terms of fatalities is not surprising as there would be a considerable number of fatalities that are not admitted to hospital (i.e., died at scene, died in transit, and died on arrival). Generally, the differences in severity between QRCD and QHAPDC should be treated with caution. QRCD had a considerably greater proportion (87% vs. 0.5%) of missing/unspecified injury descriptions which were used to determine AIS and SRR. There was also a bias in the amount of missing and unspecified injury descriptions in QRCD in terms of broad severity. Specifically, it was found that the injury description was less likely to have complete information when the case was hospitalised. It is possible that police may be less likely to complete the injury description field in cases where other parties (e.g., ambulance officers or hospital staff) are involved (as would be the case with a hospitalised case), as the police officer would defer to medical staff expertise and may think they would better capture that information in other data sources. It is also possible that in cases where the injured person is taken to hospital, that the police officer may not have the opportunity to assess the injury due the person being treated at the time or having already left the scene by the time the officer arrives. The incompleteness and inconsistency of the information required for determining objective severity measures provides further evidence that using police data alone for determining severity is problematic.

For both data collections, the ranges of severity values were quite varied. The AIS, SRR, and broad severity classification of serious injury do not correspond. It appears that using police data with a measure relating to be taken to hospital may not be indicative of serious injury. There is a broad range of injury types and SRRs within this category, and the category of ‘other injuries’ actually had a lower median SRR (more severe) than the hospitalised category. However, even based on a definition that is restricted to those admitted to hospital (as is the case in QHAPDC) it still may not be specific enough, as the range of SRRs within this category was quite wide.

Conclusion

Both the possible under-reporting in combination with the lack of precision with assigning severity found in this study make it difficult to accurately determine the cost and impact of serious injury crashes. A more precise measure of serious injury would be preferred over current practice as it is more closely related to threat to life and therefore more directly corresponding to the outcomes being measured when cost and impact is determined. Unfortunately, due to the large amount of missing information in police data, and the questionable accuracy of what is there, relying on police data alone to determine the prevalence and nature of serious injury crashes could be misleading. The inclusion of other data sources, such as hospital data, in the determination of serious injury crash impact has the potential to address the shortcomings of current approaches. However, these data collections often lack other information, which is included in police data, which are needed to determine the nature and circumstances of crashes (e.g., alcohol involvement, speed). As a result, data linkage (combining the data collections when they have individuals in common) is increasingly becoming a popular alternative to using individual data collections. Further research is required however, to assess the possibilities of data linkage, including its feasibility in the context of road safety.

References

AAAM. (2008). *AAAM Abbreviated Injury Scale 2005 update 2008*. Barrington Illinois: Association for the Advancement of Automotive Medicine; 2008.

Alsop, J. and J. Langley (2001). Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis & Prevention* ***33****(3)*, 353-359.

Aron A. and Aron E. (1991). *Statistics for psychology*. Second edition. Upper Saddle River, New Jersey: Prentice Hall.

ATC (2011). *The National Road Safety Strategy 2011–2020*. Australian Transport Council, 2011. <http://www.atcouncil.gov.au/documents/files/NRSS_2011_2020_15Aug11.pdf> (accessed 2nd April, 2013).

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.(Statistical Data Included). *Injury Prevention, 7*(3), 234.

Farmer, C. M. (2003). Reliability of police-reported information for determining crash and injury severity. *Traffic Injury Prevention, 4*, 38–44.

Langley, J. D., Dow, N., Stephenson, S., & Kypri, K. (2003). Missing cyclists. *Injury Prevention, 9*(4), 376-379.

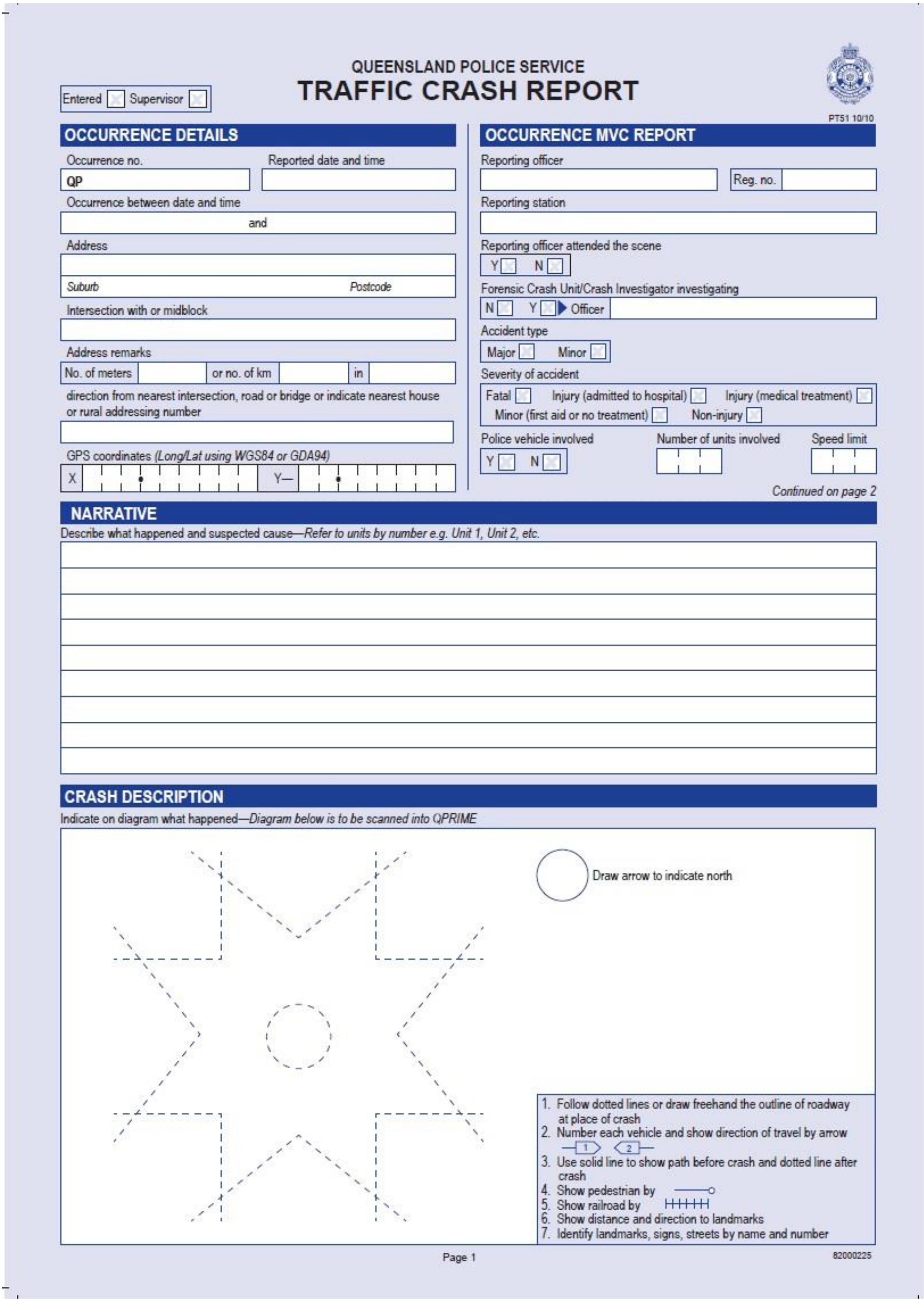
McDonald, G., Davie, G. and Langley, J. (2009). Validity of police-reported information on injury severity for those hospitalized from motor vehicle traffic crashes. *Traffic Injury Prevention, 10*, 184–190.

NCCH. (2008). *The International Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification (ICD-10-AM)*. 8th edition. NCCH: Sydney.

Stephenson, S., Henley, G., Harrison, J., and Langley, J. (2003). Diagnosis-based injury severity scaling. *Injury Research and Statistics Series, number 20*. Adelaide: AIHW (AIHW catalogue No INJCAT 59), 2003.

Ward, H., Lyons, R., Gabbe, B., Thoreau, R., Pinder, L., and Macey, S. (2010). *Road Safety Research Report No. 119 Review of Police Road Casualty Injury Severity Classification–A Feasibility Study*. Department of Transport: London.

Appendix B - PT51 Crash Reporting Form



Entered ☐ Supervisor ☐

QUEENSLAND POLICE SERVICE

TRAFFIC CRASH REPORT

PT51 10/10

OCCURRENCE DETAILS

Occurrence no. QP

Reported date and time

Occurrence between date and time ______ and ______

Address

Suburb *Postcode*

Intersection with or midblock

Address remarks

No. of meters		or no. of km		in	

direction from nearest intersection, road or bridge or indicate nearest house or rural addressing number

GPS coordinates *(Long/Lat using WGS84 or GDA94)*

X ___.______ Y— ___.______

OCCURRENCE MVC REPORT

Reporting officer

Reg. no.

Reporting station

Reporting officer attended the scene

Y ☐ N ☐

Forensic Crash Unit/Crash Investigator investigating

N ☐ Y ☐ ▶ Officer

Accident type

Major ☐ Minor ☐

Severity of accident

Fatal ☐ Injury (admitted to hospital) ☐ Injury (medical treatment) ☐ Minor (first aid or no treatment) ☐ Non-injury ☐

Police vehicle involved

Y ☐ N ☐

Number of units involved

Speed limit

Continued on page 2

NARRATIVE

Describe what happened and suspected cause—*Refer to units by number e.g. Unit 1, Unit 2, etc.*

CRASH DESCRIPTION

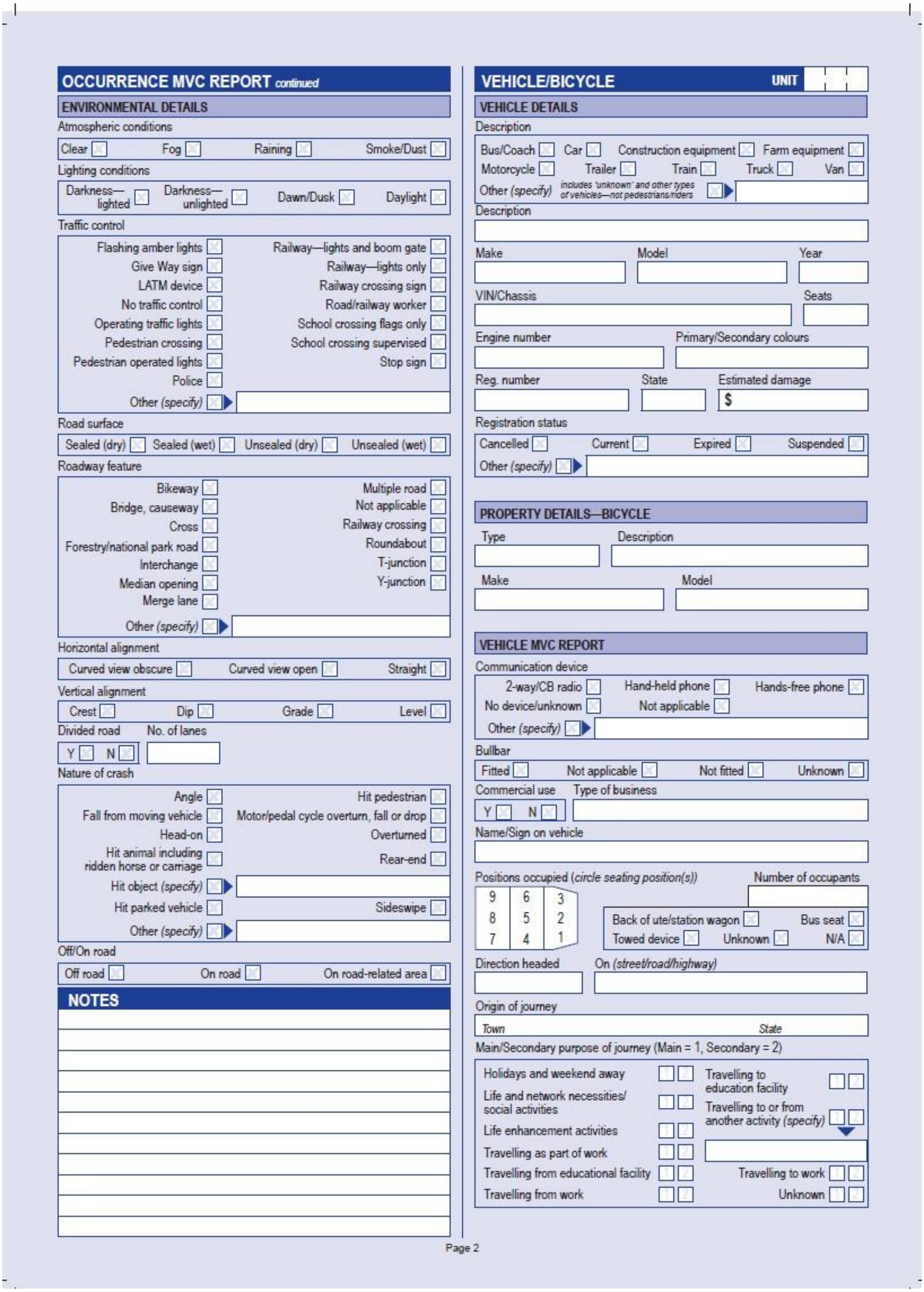
Indicate on diagram what happened—*Diagram below is to be scanned into QPRIME*

○ Draw arrow to indicate north

1. Follow dotted lines or draw freehand the outline of roadway at place of crash
2. Number each vehicle and show direction of travel by arrow
3. Use solid line to show path before crash and dotted line after crash
4. Show pedestrian by ———○
5. Show railroad by ┼┼┼┼┼
6. Show distance and direction to landmarks
7. Identify landmarks, signs, streets by name and number

Page 1

82000225



OCCURRENCE MVC REPORT *continued*

ENVIRONMENTAL DETAILS

Atmospheric conditions

Clear ☐ Fog ☐ Raining ☐ Smoke/Dust ☐

Lighting conditions

Darkness—lighted ☐ Darkness—unlighted ☐ Dawn/Dusk ☐ Daylight ☐

Traffic control

Flashing amber lights ☐ Railway—lights and boom gate ☐
Give Way sign ☐ Railway—lights only ☐
LATM device ☐ Railway crossing sign ☐
No traffic control ☐ Road/railway worker ☐
Operating traffic lights ☐ School crossing flags only ☐
Pedestrian crossing ☐ School crossing supervised ☐
Pedestrian operated lights ☐ Stop sign ☐
Police ☐
Other *(specify)* ☐ ▶

Road surface

Sealed (dry) ☐ Sealed (wet) ☐ Unsealed (dry) ☐ Unsealed (wet) ☐

Roadway feature

Bikeway ☐ Multiple road ☐
Bridge, causeway ☐ Not applicable ☐
Cross ☐ Railway crossing ☐
Forestry/national park road ☐ Roundabout ☐
Interchange ☐ T-junction ☐
Median opening ☐ Y-junction ☐
Merge lane ☐
Other *(specify)* ☐ ▶

Horizontal alignment

Curved view obscure ☐ Curved view open ☐ Straight ☐

Vertical alignment

Crest ☐ Dip ☐ Grade ☐ Level ☐

Divided road No. of lanes

Y ☐ N ☐

Nature of crash

Angle ☐ Hit pedestrian ☐
Fall from moving vehicle ☐ Motor/pedal cycle overturn, fall or drop ☐
Head-on ☐ Overturned ☐
Hit animal including ridden horse or carriage ☐ Rear-end ☐
Hit object *(specify)* ☐ ▶
Hit parked vehicle ☐ Sideswipe ☐
Other *(specify)* ☐ ▶

Off/On road

Off road ☐ On road ☐ On road-related area ☐

NOTES

VEHICLE/BICYCLE UNIT

VEHICLE DETAILS

Description

Bus/Coach ☐ Car ☐ Construction equipment ☐ Farm equipment ☐
Motorcycle ☐ Trailer ☐ Train ☐ Truck ☐ Van ☐
Other *(specify)* *includes 'unknown' and other types of vehicles—not pedestrians/riders* ☐ ▶

Description

Make Model Year

VIN/Chassis Seats

Engine number Primary/Secondary colours

Reg. number State Estimated damage $

Registration status

Cancelled ☐ Current ☐ Expired ☐ Suspended ☐
Other *(specify)* ☐ ▶

PROPERTY DETAILS—BICYCLE

Type Description

Make Model

VEHICLE MVC REPORT

Communication device

2-way/CB radio ☐ Hand-held phone ☐ Hands-free phone ☐
No device/unknown ☐ Not applicable ☐
Other *(specify)* ☐ ▶

Bullbar

Fitted ☐ Not applicable ☐ Not fitted ☐ Unknown ☐

Commercial use Type of business

Y ☐ N ☐

Name/Sign on vehicle

Positions occupied *(circle seating position(s))* Number of occupants

9	6	3
8	5	2
7	4	1

Back of ute/station wagon ☐ Bus seat ☐
Towed device ☐ Unknown ☐ N/A ☐

Direction headed On *(street/road/highway)*

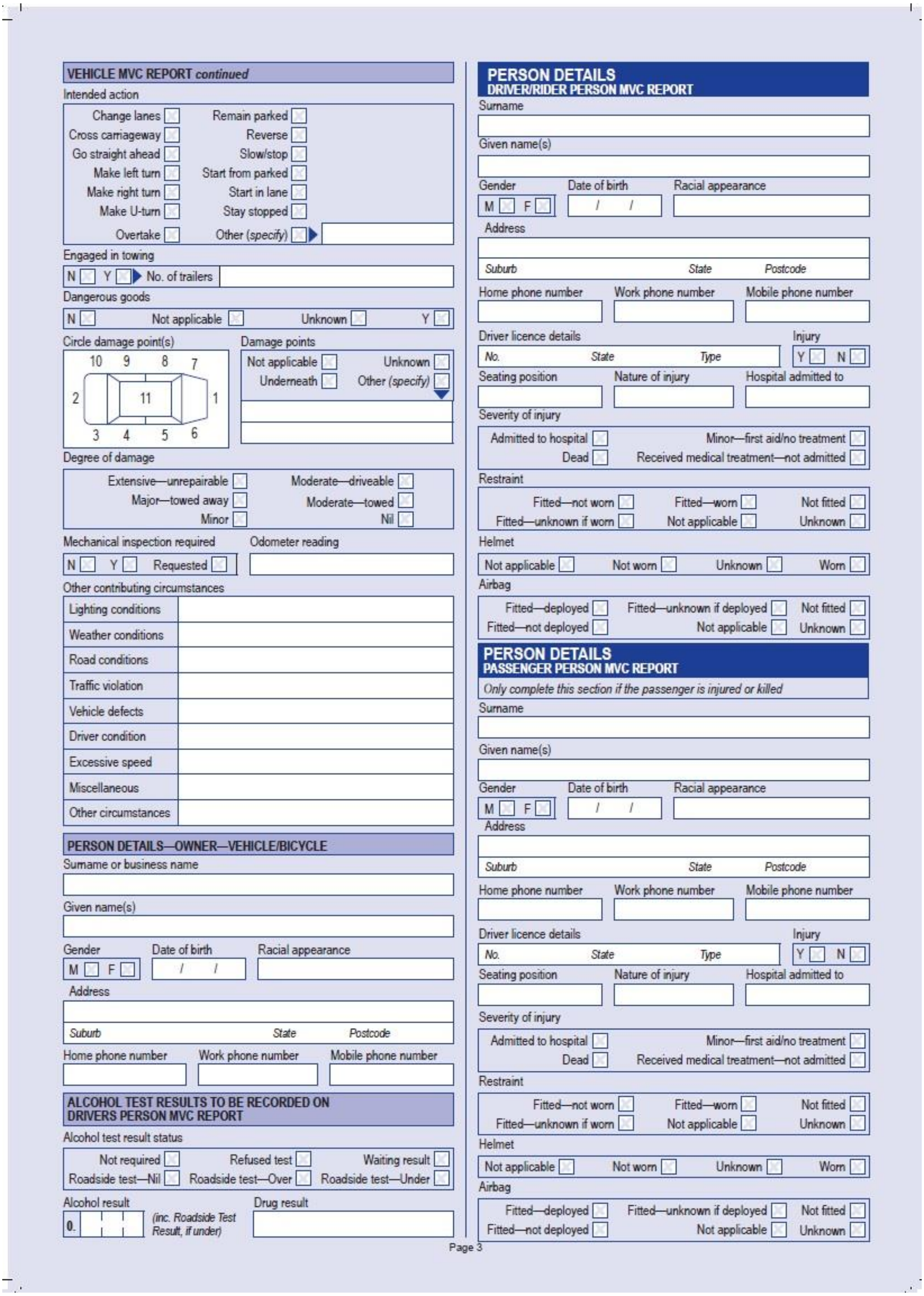
Origin of journey

Town *State*

Main/Secondary purpose of journey (Main = 1, Secondary = 2)

Holidays and weekend away ☐☐ Travelling to education facility ☐☐
Life and network necessities/social activities ☐☐ Travelling to or from another activity *(specify)* ☐☐ ▼
Life enhancement activities ☐☐
Travelling as part of work ☐☐
Travelling from educational facility ☐☐ Travelling to work ☐☐
Travelling from work ☐☐ Unknown ☐☐

Page 2



VEHICLE MVC REPORT *continued*

Intended action

Change lanes □ Remain parked □
Cross carriageway □ Reverse □
Go straight ahead □ Slow/stop □
Make left turn □ Start from parked □
Make right turn □ Start in lane □
Make U-turn □ Stay stopped □
Overtake □ Other *(specify)* □ ▶

Engaged in towing

N □ Y □ ▶ No. of trailers

Dangerous goods

N □ Not applicable □ Unknown □ Y □

Circle damage point(s)

10 9 8 7
2 11 1
3 4 5 6

Damage points

Not applicable □ Unknown □
Underneath □ Other *(specify)* □

Degree of damage

Extensive—unrepairable □ Moderate—driveable □
Major—towed away □ Moderate—towed □
Minor □ Nil □

Mechanical inspection required

N □ Y □ Requested □

Odometer reading

Other contributing circumstances

Lighting conditions	
Weather conditions	
Road conditions	
Traffic violation	
Vehicle defects	
Driver condition	
Excessive speed	
Miscellaneous	
Other circumstances	

PERSON DETAILS—OWNER—VEHICLE/BICYCLE

Surname or business name

Given name(s)

Gender Date of birth Racial appearance
M □ F □ / /

Address

Suburb *State* *Postcode*

Home phone number Work phone number Mobile phone number

ALCOHOL TEST RESULTS TO BE RECORDED ON DRIVERS PERSON MVC REPORT

Alcohol test result status

Not required □ Refused test □ Waiting result □
Roadside test—Nil □ Roadside test—Over □ Roadside test—Under □

Alcohol result

0. *(inc. Roadside Test Result, if under)*

Drug result

PERSON DETAILS
DRIVER/RIDER PERSON MVC REPORT

Surname

Given name(s)

Gender Date of birth Racial appearance
M □ F □ / /

Address

Suburb *State* *Postcode*

Home phone number Work phone number Mobile phone number

Driver licence details Injury
No. *State* *Type* Y □ N □

Seating position Nature of injury Hospital admitted to

Severity of injury

Admitted to hospital □ Minor—first aid/no treatment □
Dead □ Received medical treatment—not admitted □

Restraint

Fitted—not worn □ Fitted—worn □ Not fitted □
Fitted—unknown if worn □ Not applicable □ Unknown □

Helmet

Not applicable □ Not worn □ Unknown □ Worn □

Airbag

Fitted—deployed □ Fitted—unknown if deployed □ Not fitted □
Fitted—not deployed □ Not applicable □ Unknown □

PERSON DETAILS
PASSENGER PERSON MVC REPORT

Only complete this section if the passenger is injured or killed

Surname

Given name(s)

Gender Date of birth Racial appearance
M □ F □ / /

Address

Suburb *State* *Postcode*

Home phone number Work phone number Mobile phone number

Driver licence details Injury
No. *State* *Type* Y □ N □

Seating position Nature of injury Hospital admitted to

Severity of injury

Admitted to hospital □ Minor—first aid/no treatment □
Dead □ Received medical treatment—not admitted □

Restraint

Fitted—not worn □ Fitted—worn □ Not fitted □
Fitted—unknown if worn □ Not applicable □ Unknown □

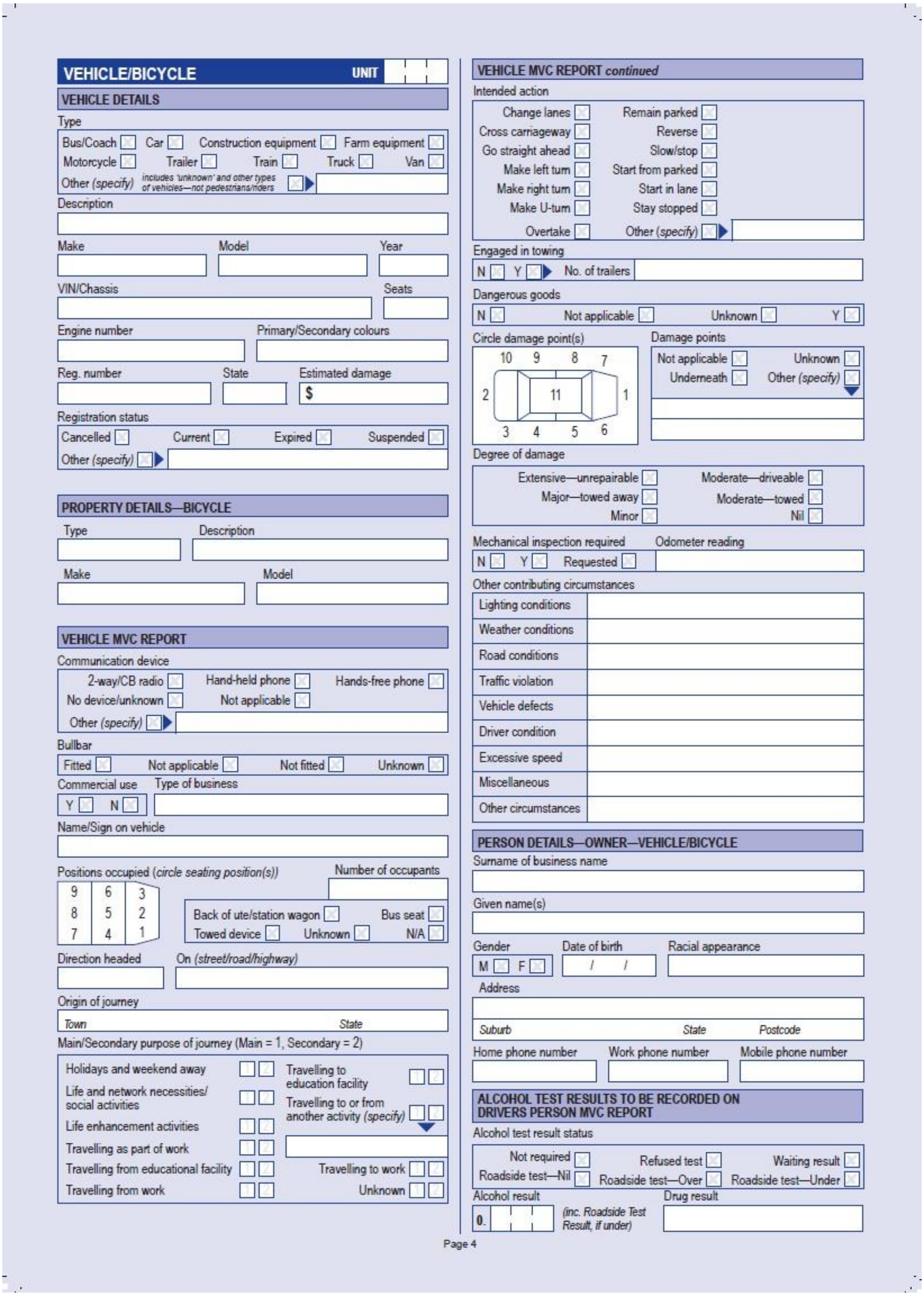
Helmet

Not applicable □ Not worn □ Unknown □ Worn □

Airbag

Fitted—deployed □ Fitted—unknown if deployed □ Not fitted □
Fitted—not deployed □ Not applicable □ Unknown □

Page 3



VEHICLE/BICYCLE

UNIT

VEHICLE DETAILS

Type

Bus/Coach □ Car □ Construction equipment □ Farm equipment □
Motorcycle □ Trailer □ Train □ Truck □ Van □
Other *(specify)* *includes 'unknown' and other types of vehicles—not pedestrians/riders* □ ▶

Description

Make | Model | Year

VIN/Chassis | Seats

Engine number | Primary/Secondary colours

Reg. number | State | Estimated damage $

Registration status

Cancelled □ Current □ Expired □ Suspended □
Other *(specify)* □ ▶

PROPERTY DETAILS—BICYCLE

Type | Description

Make | Model

VEHICLE MVC REPORT

Communication device

2-way/CB radio □ Hand-held phone □ Hands-free phone □
No device/unknown □ Not applicable □
Other *(specify)* □ ▶

Bullbar

Fitted □ Not applicable □ Not fitted □ Unknown □

Commercial use | Type of business

Y □ N □

Name/Sign on vehicle

Positions occupied *(circle seating position(s))* | Number of occupants

9 6 3
8 5 2
7 4 1

Back of ute/station wagon □ Bus seat □
Towed device □ Unknown □ N/A □

Direction headed | On *(street/road/highway)*

Origin of journey

Town *State*

Main/Secondary purpose of journey (Main = 1, Secondary = 2)

Holidays and weekend away □□
Life and network necessities/ social activities □□
Life enhancement activities □□
Travelling as part of work □□
Travelling from educational facility □□
Travelling from work □□
Travelling to education facility □□
Travelling to or from another activity *(specify)* □□
Travelling to work □□
Unknown □□

VEHICLE MVC REPORT *continued*

Intended action

Change lanes □ Remain parked □
Cross carriageway □ Reverse □
Go straight ahead □ Slow/stop □
Make left turn □ Start from parked □
Make right turn □ Start in lane □
Make U-turn □ Stay stopped □
Overtake □ Other *(specify)* □ ▶

Engaged in towing

N □ Y □ ▶ No. of trailers

Dangerous goods

N □ Not applicable □ Unknown □ Y □

Circle damage point(s)

10 9 8 7
2 11 1
3 4 5 6

Damage points

Not applicable □ Unknown □
Underneath □ Other *(specify)* □

Degree of damage

Extensive—unrepairable □ Moderate—driveable □
Major—towed away □ Moderate—towed □
Minor □ Nil □

Mechanical inspection required | Odometer reading

N □ Y □ Requested □

Other contributing circumstances

Lighting conditions	
Weather conditions	
Road conditions	
Traffic violation	
Vehicle defects	
Driver condition	
Excessive speed	
Miscellaneous	
Other circumstances	

PERSON DETAILS—OWNER—VEHICLE/BICYCLE

Surname of business name

Given name(s)

Gender | Date of birth | Racial appearance

M □ F □ / /

Address

Suburb *State* *Postcode*

Home phone number | Work phone number | Mobile phone number

ALCOHOL TEST RESULTS TO BE RECORDED ON DRIVERS PERSON MVC REPORT

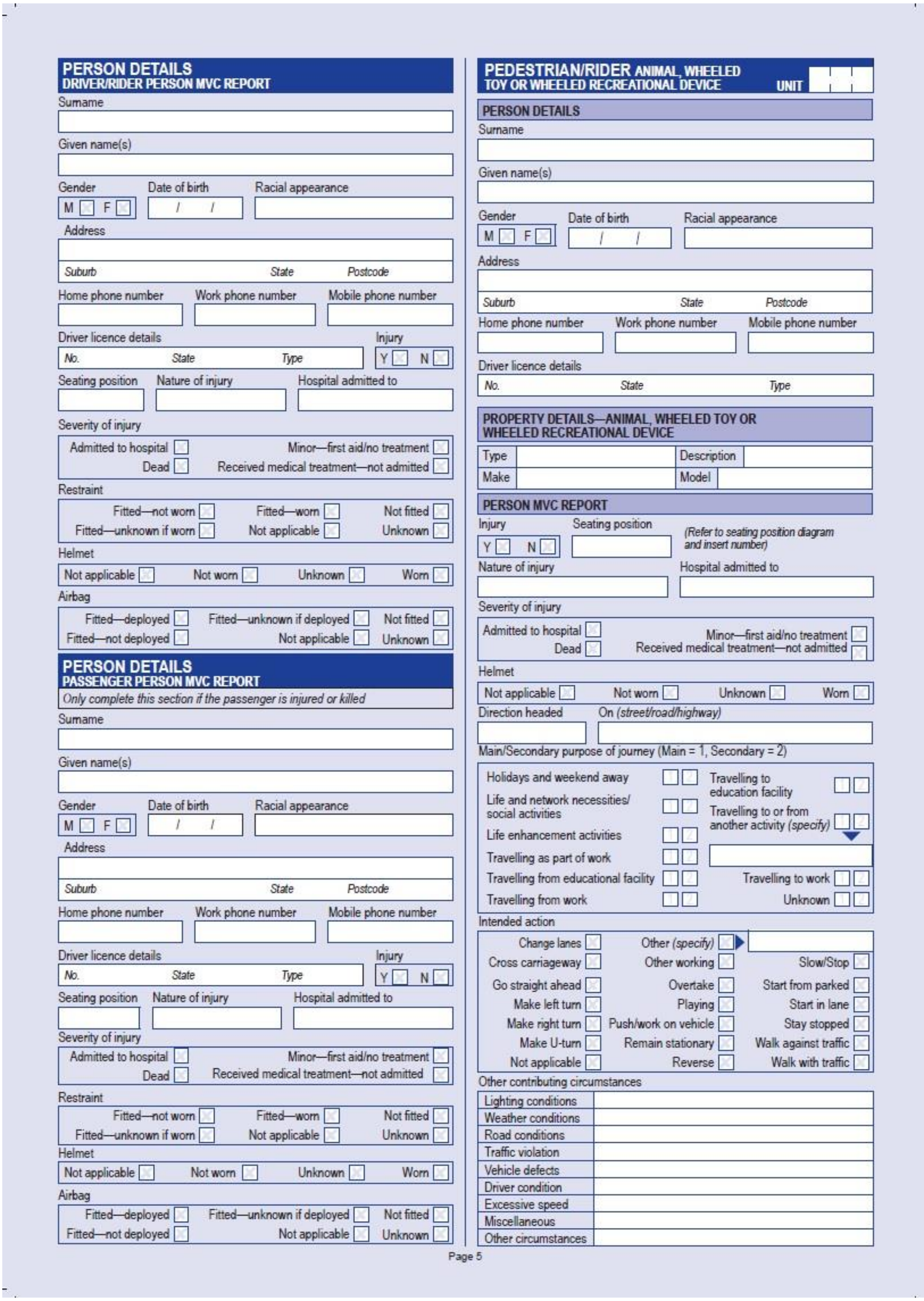
Alcohol test result status

Not required □ Refused test □ Waiting result □
Roadside test—Nil □ Roadside test—Over □ Roadside test—Under □

Alcohol result | Drug result

0. *(inc. Roadside Test Result, if under)*

Page 4



PERSON DETAILS
DRIVER/RIDER PERSON MVC REPORT

Surname

Given name(s)

Gender M ☒ F ☒ Date of birth / / Racial appearance

Address

Suburb *State* *Postcode*

Home phone number Work phone number Mobile phone number

Driver licence details *No.* *State* *Type* Injury Y ☒ N ☒

Seating position Nature of injury Hospital admitted to

Severity of injury

Admitted to hospital ☒ Minor—first aid/no treatment ☒

Dead ☒ Received medical treatment—not admitted ☒

Restraint

Fitted—not worn ☒ Fitted—worn ☒ Not fitted ☒

Fitted—unknown if worn ☒ Not applicable ☒ Unknown ☒

Helmet

Not applicable ☒ Not worn ☒ Unknown ☒ Worn ☒

Airbag

Fitted—deployed ☒ Fitted—unknown if deployed ☒ Not fitted ☒

Fitted—not deployed ☒ Not applicable ☒ Unknown ☒

PERSON DETAILS
PASSENGER PERSON MVC REPORT

Only complete this section if the passenger is injured or killed

Surname

Given name(s)

Gender M ☒ F ☒ Date of birth / / Racial appearance

Address

Suburb *State* *Postcode*

Home phone number Work phone number Mobile phone number

Driver licence details *No.* *State* *Type* Injury Y ☒ N ☒

Seating position Nature of injury Hospital admitted to

Severity of injury

Admitted to hospital ☒ Minor—first aid/no treatment ☒

Dead ☒ Received medical treatment—not admitted ☒

Restraint

Fitted—not worn ☒ Fitted—worn ☒ Not fitted ☒

Fitted—unknown if worn ☒ Not applicable ☒ Unknown ☒

Helmet

Not applicable ☒ Not worn ☒ Unknown ☒ Worn ☒

Airbag

Fitted—deployed ☒ Fitted—unknown if deployed ☒ Not fitted ☒

Fitted—not deployed ☒ Not applicable ☒ Unknown ☒

PEDESTRIAN/RIDER ANIMAL, WHEELED TOY OR WHEELED RECREATIONAL DEVICE

UNIT

PERSON DETAILS

Surname

Given name(s)

Gender M ☒ F ☒ Date of birth / / Racial appearance

Address

Suburb *State* *Postcode*

Home phone number Work phone number Mobile phone number

Driver licence details

No. *State* *Type*

PROPERTY DETAILS—ANIMAL, WHEELED TOY OR WHEELED RECREATIONAL DEVICE

Type		Description	
Make		Model	

PERSON MVC REPORT

Injury Y ☒ N ☒ Seating position *(Refer to seating position diagram and insert number)*

Nature of injury Hospital admitted to

Severity of injury

Admitted to hospital ☒ Minor—first aid/no treatment ☒

Dead ☒ Received medical treatment—not admitted ☒

Helmet

Not applicable ☒ Not worn ☒ Unknown ☒ Worn ☒

Direction headed On *(street/road/highway)*

Main/Secondary purpose of journey (Main = 1, Secondary = 2)

Holidays and weekend away □□

Life and network necessities/ social activities □□

Life enhancement activities □□

Travelling as part of work □□

Travelling from educational facility □□

Travelling from work □□

Travelling to education facility □□

Travelling to or from another activity *(specify)* □□

Travelling to work □□

Unknown □□

Intended action

Change lanes ☒ Other *(specify)* ☒

Cross carriageway ☒ Other working ☒ Slow/Stop ☒

Go straight ahead ☒ Overtake ☒ Start from parked ☒

Make left turn ☒ Playing ☒ Start in lane ☒

Make right turn ☒ Push/work on vehicle ☒ Stay stopped ☒

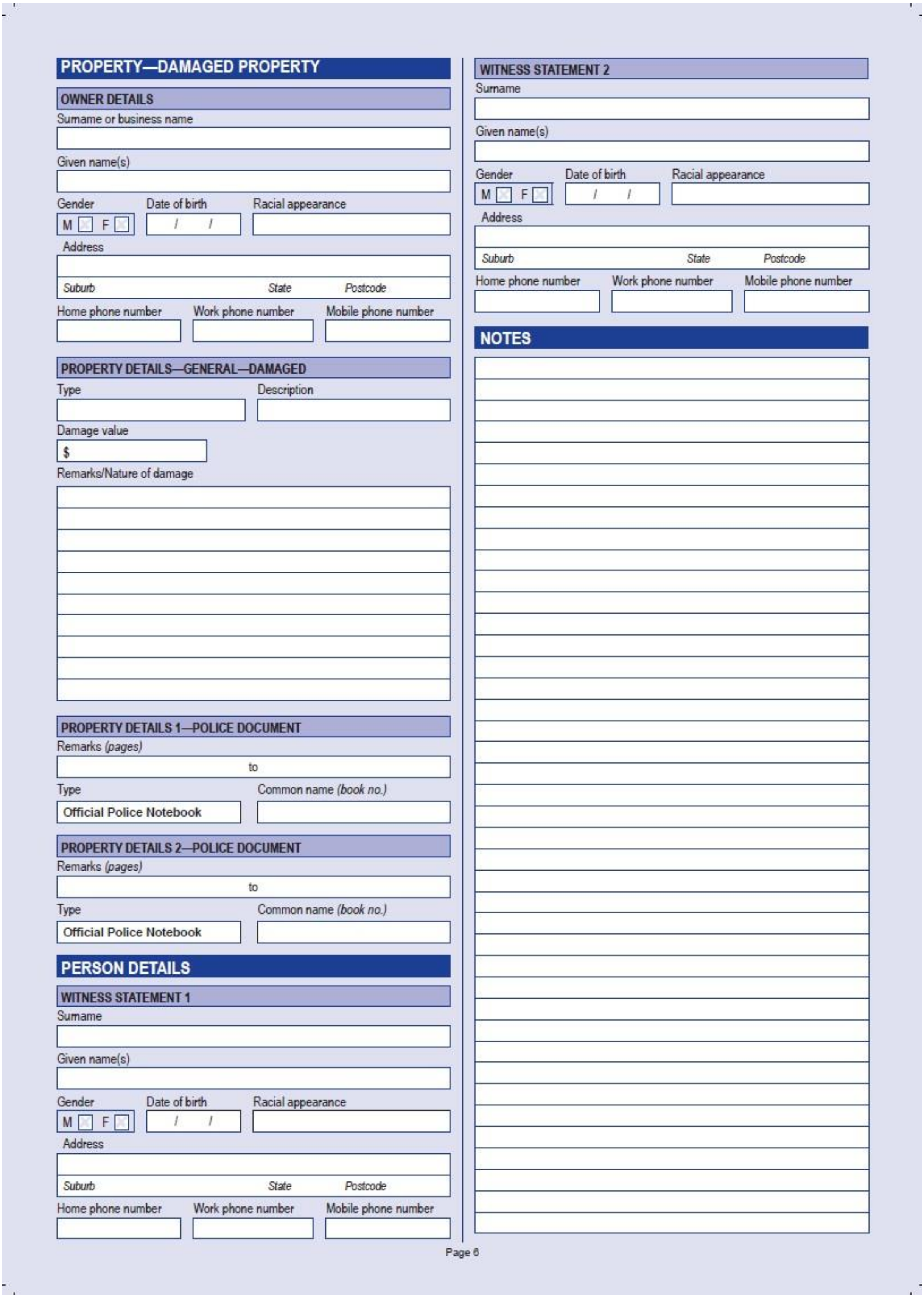
Make U-turn ☒ Remain stationary ☒ Walk against traffic ☒

Not applicable ☒ Reverse ☒ Walk with traffic ☒

Other contributing circumstances

Lighting conditions	
Weather conditions	
Road conditions	
Traffic violation	
Vehicle defects	
Driver condition	
Excessive speed	
Miscellaneous	
Other circumstances	

Page 5



PROPERTY—DAMAGED PROPERTY

OWNER DETAILS

Surname or business name

Given name(s)

Gender M ☒ F ☒

Date of birth / /

Racial appearance

Address

Suburb *State* *Postcode*

Home phone number

Work phone number

Mobile phone number

PROPERTY DETAILS—GENERAL—DAMAGED

Type

Description

Damage value

$

Remarks/Nature of damage

PROPERTY DETAILS 1—POLICE DOCUMENT

Remarks *(pages)*

to

Type

Official Police Notebook

Common name *(book no.)*

PROPERTY DETAILS 2—POLICE DOCUMENT

Remarks *(pages)*

to

Type

Official Police Notebook

Common name *(book no.)*

PERSON DETAILS

WITNESS STATEMENT 1

Surname

Given name(s)

Gender M ☒ F ☒

Date of birth / /

Racial appearance

Address

Suburb *State* *Postcode*

Home phone number

Work phone number

Mobile phone number

WITNESS STATEMENT 2

Surname

Given name(s)

Gender M ☒ F ☒

Date of birth / /

Racial appearance

Address

Suburb *State* *Postcode*

Home phone number

Work phone number

Mobile phone number

NOTES

Appendix C – Crash Request Form

Department of Transport and Main Roads - Road Crash, Registration, Licensing and Infringement Data Request Form

*Please use **BLOCK LETTERS** if handwritten.*

Contact Details

Office Use Only

Request Number: rq ……………………..

Priority: ……………………………………....

Link Number: rq ..……..........……...……..

Due Date: ……….…………………………

Name:

Email

Phone:

Alternate phone:

Fax:

Organisation

Please tick appropriate box(es): □ Road Crash Data □ Registration/Licensing/Infringement Data

Request Information

When do you require this data? *Note: Normal turnaround time is at least 5 working days; complex requests will take longer. If data is required before this time, please state the date (& time if appropriate) you require it. If your requested timeframe is not achievable we will contact you to negotiate a timeframe. **requests marked as "URGENT" or "ASAP" will be automatically allocated a 5 working day turnaround***

Is this **updating previous data supplied**? If possible, please provide the **request number** and/or approximate **date** that the previous data was supplied. Also, if available, please **attach** the data.

How do you plan to **use** this data? For example: *presentation, research paper, ministerial.*

Time range

□ Previous 5 full years of data □ Previous 12 full months of data □ Year to date

Other time range / comments, how would you like it broken down? *Example: year, month*

<u>Geographical area</u>

□ All of Queensland	□ Police Region	□ Queensland Transport Region	□ Road/Hwy
□ Local Government Area	□ Police District	□ Main Roads District	□ Road/Hwy section
□ Statistical Local Area	□ Police Division		□ Intersection

Geographic details and comments. ***Note:*** *Registration, licensing and infringement data are not available for some areas such as, Road/Hwy, Road/Hwy section and Intersection.*

Statistical Data Required

Road Crash Data: *(examples of possible characteristics)*

Crashes	**Casualties**	**Units**	**Unit controllers**	**Contributing circumstances**
□ Severity □ Crash nature □ Roadway feature □ Traffic control □ Speed limit □ Roadway surface □ Atmospheric condition □ Lighting □ Horizontal alignment □ Vertical alignment □ DCA code □ DCA group □ Time of day □ Day of week	□ Severity □ Road user type □ Road user type – unit group □ Age □ Gender □ Helmet use □ Restraint use □ Seating position	□ Unit type □ Intended action □ Overall damage □ Main damage point □ Towing □ Number of occupants □ Dangerous goods □ Defective □ Registration status □ Type of use (business or private)	□ Road user type □ Age □ Gender □ Licence type □ State licensed in	□ Contributing circumstances □ Contributing factors (circumstance groupings)

<u>Registration Licensing and Infringement Data</u>: *(examples of possible characteristics)*

Registration	**Licensing**	**Infringement**	**Recreational Vessels**
□ New Business □ Transfers □ Vehicles by body type □ Make □ Model □ Gross Vehicle Mass □ Purpose Of Use	□ Age □ Gender □ Class □ Level	□ Category □ Description □ Code	□ Length □ Draft □ Body Type □ Registration Category □ Powered by

<u>Data request comments and details</u>:

Please send this form to:

Data Analysis, Department of Transport and Main Roads

Email: <u>DataAnalysis@tmr.qld.gov.au</u>

Fax: (07) 3066 2410

The Department of Transport and Main Roads is collecting the information on this form for the purposes of providing you with road crash, registration, licensing and infringement data. Your personal details will not be disclosed to any other third party without your consent unless required or authorised to do so by law.

Appendix D – Interview Schedules

Interview Questions – expert data users

General

Which sources of transport related injury data have you had experience with (have accessed or tried to access)?

Relevance

For what purpose/s do you use these data?

What sort of research questions?

Epidemiological/Risk

Longitudinal

Prevalence/surveillance

Evaluation

How well do the data identify new or emerging issues/problems in traffic incidents/crashes/injuries?

Generally?

Specific incident types or road user groups?

Adequacy

How well do the data describe key characteristics of the traffic incidents and the injuries involved?

For example the WHO and Austroads guidelines suggest the following as core minimum:

- a unique person/event identifier;
- age of the injured person;
- sex of the injured person;
- location the injury occurred;

- mechanism or cause;
- external cause of injury;
- date of injury;
- time of injury;
- severity of injury; and
- nature of the injury.

What do you believe is core information?

What else could be included?

Road user types

Vehicle information

Contributing circumstances

Controller information (not necessarily injured)

Is there anything that could be excluded?

What additional information is available about the incident/injured parties?

What incidents/events are not included in the data collection?

By definition?

Due to error/not reported?

Quantifiable?

Across data sets

Are the data able to identify risk groups and factors?

How important is validity in data vs. reliability?

Completeness

Is there missing/unknown data?

Is the missing data quantifiable?

Which variables are commonly missing and why?

- a unique person/event identifier;
- age of the injured person;
- sex of the injured person;
- location the injury occurred; - detail
- mechanism or cause;
- external cause of injury;
- date of injury;
- time of injury;
- severity of injury; and
- nature of the injury.
- Road user types
- Vehicle information
- Contributing circumstances
- Controller information (not necessarily injured)

Reliability

Is there any misclassification?

Is the misclassification quantifiable?

Impossibilities

Which variables are often subject to misclassification and why?

What are the data checking/cleaning/auditing processes, if any?

How well do the data allow the monitoring of traffic incidents/crashes/injuries over time?

Does the nature and quality of information recorded vary depending on the type/nature of the incident/injury?

Location?

Road user group?

Severity?

Any other factors?

What variables do data sets you work with have in common with each other?

Are they coded/recorded the same way?

What do you know is not consistent?

Are the data coded using any national or international standards?

How could reliability and consistency both within and between data sets be improved?

Timeliness

What are the impacts on research with delays in data being available?

If some data was available sooner, what would you like to see at a minimum?

Access and sharing

What are the processes, including any ethical processes, in order for access to data to be granted?

What are the barriers/facilitators?

How long does it usually take?

How could these processes be improved?

What is the nature of common requests you make for data?

What is the nature of requests that have been denied?

Why were they denied?

Is there an example in which data has been requested and permission granted, however subsequent request of the same nature been denied?

Data analysis

In what form is the data usually provided to you?

Is the form the data are provided able to be analysed without manipulation?

If manipulation is needed, what form?

What documentation is available to assist in data analysis and interpretation?

How helpful are they?

What improvements, if any, could be made with the way in which data is provided?

Data linkage

Have you been involved in any linkage projects?

What was the nature of the project/s?

How was the linkage done?

What do you believe the perceived barriers are to linkage?

What improvements would be needed to make linkage more feasible?

Interview Questions – data custodians

The following questions are asked in relation to the (name of data source)

Relevance

What is the primary purpose of the data collection?

What are the other purposes, if any?

How well do the data identify new or emerging issues/problems in traffic incidents/crashes/injuries?

Generally?

Specific incident types or road user groups?

What are the years covered by the database?

Adequacy

How well do the data describe key characteristics of the traffic incidents and the injuries involved?

For example the WHO and Austroads guidelines suggest the following as core minimum:

- a unique person/event identifier;
- age of the injured person;
- sex of the injured person;
- location the injury occurred;
- mechanism or cause;
- external cause of injury;
- date of injury;
- time of injury;
- severity of injury; and
- nature of the injury.

What do you believe is core information?

What else could be included?

Road user types

Vehicle information

Contributing circumstances

Controller information (not necessarily injured)

What would be involved in adding this information to the data?

e.g., New variable field

Process (if any)?

Barriers?

Is there anything you think should be excluded?

What additional information is available about the incident/injured parties?

What incidents/events are not included in the data collection?

By definition?

Due to error?

Are the data able to identify risk groups and factors?

Data collection processes

When are data collected?

Where are data collected?

Who collects the data?

In what form are data collected?

Is there a standard form?

Tick boxes vs. Free text

How are the data collated?

What is the process of data from event to inclusion in the data set?

Are there any modifications to the data during this process?

Is the data coded according to national/international standards? (e.g., ICD-10)

Who completes the coding?

Who are the funders?

Who owns the data?

Completeness

How much missing/unknown data?

Specific fields in which it’s missing:

- a unique person/event identifier;
- age of the injured person;
- sex of the injured person;

- location the injury occurred; - detail
- mechanism or cause;
- external cause of injury;
- date of injury;
- time of injury;
- severity of injury; and
- nature of the injury.
- Road user types
- Vehicle information
- Contributing circumstances
- Controller information (not necessarily injured)

Why is it missing/unknown?

Reliability

Is there any misclassification?

If so, of what nature/which variables?

What are the data checking/cleaning/auditing processes, if any?

How well do the data allow the monitoring of traffic incidents/crashes/injuries over time?

Is this stable/consistent? E.g., any changes in the last 10 years or planned in the future

Consistency

Does the nature and quality of information recorded vary depending on the type/nature of the incident/injury?

Location?

Road user group?

Severity?

Any other factors?

What information/coding does this data set have in common with any other traffic or injury data sets in Queensland/interstate/internationally (if anything)?

What do you know is not consistent?

Are the data coded using any national or international standards?

Timeliness

What is the delay between an event occurring and it being available in the data set?

Are there processes in place to manage delays?

Are data (parts) able to be released/is released in stages?

Access and sharing

Is there any legislation relating the storing, reporting or access to data? Including those relating to privacy?

What procedures are in place to deal with privacy issues?

How does this impact on the release of data?

What are the levels of access? Who has access?

Are data routinely shared with any other agencies/organisations?

> If so, on what basis?
>
> And in what form?

What are the processes, including any ethical processes in order for access to data to be granted?

What is the nature of common requests?

What is the nature of requests that are unable to be granted?

Why are they not granted?

How are requests managed?

What are considered appropriate persons/use for data?

Is there an example in which data has been requested and permission granted, however the data was used in a manner that your organisation was unhappy with?

> Without saying who was involved, can you give some details of what your organisation was unhappy about, and how it was dealt with? (e.g., misinterpretation)

Data analysis

In what form is data stored?

What software/programs/language is used?

How is it extracted and in what form?

Is the form the data are provided able to be analysed without manipulation?

If manipulation is needed, what form?

What documentation is available to assist in data analysis and interpretation? E.g., glossary/definitions/coding keys

Data linkage

Are there any current linkage processes, if so how is it achieved?

What are the perceived barriers to linkage?

Data linkage experts

What do you think is the best practice model of data linkage?

(including things such as governance; the role of custodians, researchers, data linkers; whether linkage keys are kept; whether data sources are consolidated as part of linkage; method used (deterministic vs. probabilistic) etc.)

Can you describe some of the difficulties you have experienced in linking data?

What are the benefits you see with using linked data in research?

Has the quality of the linked data been examined by the data linkage unit/data custodians/researchers?

If quality was examined, how was this done and what were the results?

If you were to give advice to a new data linkage centre what would it be?

Appendix E – Data Collection Variable Fields

Table E.1: Queensland Road Crash Database (QRCD) variables

Identifying variables	Event/crash	Crash number Crash date
	Individuals	CRN (Drivers/Riders) Name (Controllers/causalities) Address (Controllers/causalities) DOB (Controllers/causalities)
Crash variables	Nature/circumstances	Crash severity Crash Nature Crash Speed Limit Crash Horizontal Alignment Crash Vertical Alignment Crash Roadway Feature Crash Traffic Control Crash Lighting Condition Crash Atmospheric Condition Crash Surface Condition Crash DCA Code Crash DCA Description Crash DCA Group Number of Units Involved Number of Casualties Circumstance Code Circumstance Description
	Temporal	Crash Day of Week Crash Month Crash Year Crash Time
	Location	Crash SLA Crash LGR Crash Police Region Crash Police Division Crash Police District Crash Transport Region Crash Main Roads District Crash ARIA+ Crash Longitude Crash Latitude Crash Street Crash Intersecting Street (If applicable)

Unit variables	Unit type Unit Intended Action Unit Headed Direction Unit Overall Damage Unit Main Damage Point Unit Number of Occupants Unit Type of Business Unit Origin State Unit Origin Town Unit Street ID Vehicle State Registered Vehicle Make Vehicle Model Vehicle Body Type Unit GVM (If applicable – trucks utes etc.)
Controller variables	Controller Licence Type Controller BAC Controller Age Controller Gender
Casualty variables	Casualty Severity Injury Description Casualty Age Casualty Gender Casualty Road User Type Casualty Unit Type Casualty Restraint Use Casualty Helmet Use Casualty Seating Position

Table E.2: Queensland Hospital Admitted Patients Data Collection (QHAPDC) variables

Identifying information	UR number Facility number Name DOB Address
Case information	Statistical Division of usual residence Hospital locality (ARIA+) Age Sex Day of week Month Year

	Length of stay
	Mode of discharge
	Diagnosis string
	Procedure string
	External cause string
	Place string
	Activity string
	Compensable status

Table E.3: Queensland Injury Surveillance Unit (QISU) variables

Identifying information	UR number
	Facility number
	Postcode
Case information	Age
	Sex
	Day of week
	Time of presentation
	Month
	Year
	Length of stay
	Presenting problem
	Hospital name
	Mode of separation
	Injury text description
	External cause
	Place
	Activity
	Intent
	Diagnosis codes
	Triage score
	Mechanism and major injury factor
	Nature of injury
	Postcode of usual residence

Table E.4: Emergency Department Information System (EDIS) variables

Identifying information	UR number
	Facility number
	Name
	DOB
	Address
	Arrival date
	Arrival time
	Arrival day*
Case information	Present postcode

	Present suburb
	Campus code
	Age
	Gender
	Mode of arrival
	Departure destination
	Departure status
	Presenting problem
	Presenting problem nurses assessment
	Diagnosis ICD code primary
	Diagnosis description primary
	Triage priority
	Presenting complaint code

Appendix F – Pull Out Supplement for Chapter 5 Methodology

Table F.1: Data selection criteria and coding for each data collection in Study 2 (Chapter 5)

	General	QRCD	QHAPDC	eARF	QISU	EDIS	NCIS
Selection of road crashes		All casualties	All acute admissions with ICD-10-AM External Cause Codes from V00-V89 and fourth character of ‘traffic’	*Case nature* (Bicycle Collision ; Motor Vehicle Collision; Motorcycle Collision; Pedestrian Collision) and *location type* (street; public transport; vehicle)	*External definition* (Motor vehicle – driver; Motor vehicle – passenger; Motorcycle – driver; Motorcycle – passenger; Pedal cyclist or pedal cyclist passenger; Pedestrian) and *type of place* (street/highway)	*Presenting problem* keyword search (e.g., car, motorbike, pedestrian) without exclusion terms (e.g., off-road, track)	*Primary mechanism* (blunt force), *secondary mechanism* (transport incident), *object* (not air or water), *context* (land transport traffic injury event), and *intent code* (unintentional)
Age	5 year age groups (with the exception 85+).	Provided in single years re-coded into 5 year age groups	Retained as coded	Coded from *date of birth*	Provided in single years re-coded into 5 year age groups	Provided in single years re-coded into 5 year age groups	Provided in single years re-coded into 5 year age groups
Gender	1 = Female; 2 = Male	Retained as coded	Recoded to 1 = Female; 2 = Male	Retained as coded	Retained as coded	Retained as coded	Retained as coded
Severity	1. *Broad severity* (fatality; hospitalisation;	1. *Casualty severity* (1= fatality; 2 =	1. *Mode of separation* (‘died in	Not able to be coded	1. *Mode of separation* (died in ED, dead on	1. *Departure status* (died in ED, dead on	Not determined as all cases fatalities

	other injury). *2. The Abbreviated Injury Scale (AIS)* (1 = minor; 2 = moderate; 3 = serious; 4 = severe; 5 = critical; and 6 = maximum). Mapped to principal diagnosis ICD-10 codes in the data (when available). *3. Survival Risk Ratios (SRR)* - estimate of the probability of death from 0 (no chance of survival) to 1 (100% chance of survival). SRRs were mapped to principal diagnosis ICD codes.	hospitalisation; 3 = medical treatment; 4 = minor injury), with 'medical treatment' and 'minor injury' collapsed into 'other injury' 2 and 3. *AIS* and *SRR* were coded using the *injury description* variable. Principal diagnosis ICD-10-AM code mapped to each injury description. ICD codes then mapped to AIS and a SRR	hospital' = fatality; all other cases = hospitalised) 2 and 3. *Principal diagnosis* ICD-10-AM codes mapped to the *AIS* and a *SRR*		arrival = fatality; all other cases = hospitalisation) 2 and 3. *Principal diagnosis* ICD-10-AM codes mapped to the *AIS* and a *SRR*	arrival = fatality; all other cases = hospitalisation) 2 and 3. *Principal diagnosis* ICD-10-AM codes mapped to the *AIS* and a *SRR*	

ARIA+	ARIA+ (1 = Major Cities; 2 = Inner Regional; 3 = Outer Regional; 4 = Remote; 5 = Very Remote).	Retained as coded	Retained as coded	Postcode mapped to ARIA+ using data from the Australian Bureau of Statistics (2013).	Postcode mapped to ARIA+ using data from the Australian Bureau of Statistics (2013).	Postcode mapped to ARIA+ using data from the Australian Bureau of Statistics (2013).	Postcode mapped to ARIA+ using data from the Australian Bureau of Statistics (2013).
Road user	1 = Driver, 2 = Motorcyclist, 3 = Cyclist, 4 = Pedestrian; 5 = Car passenger	*Casualty road user type*. Coding was retained with the exception of ‘motorcycle pillions’ and ‘bicycle pillions’ recoded into ‘motorcyclist’ and ‘cyclist’ respectively.	Second and fourth characters of the ICD-10-AM *external cause code*.	Combination of *case nature*, *vehicle type*, and *comments*	*External code* (motor vehicle – driver = driver; motorcycle – driver and motorcycle – passenger = motorcyclist; pedal cyclist or pedal cyclist passenger = cyclist; pedestrian = pedestrian; motor vehicle passenger = passenger)	*Presenting problem* text search (e.g., driver = driver; motorcycle, MCA, MBA = motorcyclist; bicycle, PBS, PBA = cyclist; passenger = passenger; none of the keywords = unspecified)	*Mode of transport* and *user code* (e.g., *user code* = driver, rider or operator and *mode of transport* = light transport; heavy transport; and special all-terrain vehicle coded as driver)

Appendix G - Road Crash Search Terms

Inclusion terms

- MVC
- MVA
- MBA
- MBC
- MOTORCYCLE
- DRIVER
- BICYC
- CYCLIST
- PEDESTRIAN
- CAR
- BIKE
- PBA
- DRIVING
- TRUCK
- TRANSPORT
- TAXI
- BUS
- RTC
- SEATBELT
- KM
- VEHIC

Exclusion terms

- DOOR
- OFF ROAD
- HOUSE
- YARD
- QUAD BIKE
- ASSAULT
- PROPERTY
- GARAGE
- DRIVEWAY/DRIVE WAY
- TRACK
- PATH
- TRAIL
- MOTORCROSS/MOTOR CROSS
- DIRT
- JUMP
- SCREW
- CARPARK/CAR PARK

Appendix H – Relationships between Independent Variables

QHAPDC and QRCD

Table H.1: Road user type by gender for QHAPDC and QRCD

		Gender	
		Female	Male
Road user	Driver	2,574 (52.2%)	3,053 (35.8%)
	Motorcyclist	399 (8.1%)	2,640 (31.0%)
	Bicyclist	258 (5.2%)	1,171 (13.7%)
	Pedestrian	339 (6.9%)	560 (6.6%)
	Passenger	1,362 (27.6%)	1,094 (12.8%)
		$\chi^2(4) = 1511.58$, $p < .001$, $\phi_c = .34$	

Table H.2: ARIA+ by gender for QHAPDC and QRCD

		Gender	
		Female	Male
ARIA+	Major Cities	2,782 (55.2%)	4,581 (52.8%)
	Inner Regional	1,218 (24.2%)	2,171 (25.0%)
	Outer Regional	849 (16.8%)	1,531 (17.6%)
	Remote	120 (2.4%)	242 (2.8%)
	Very Remote	70 (1.4%)	159 (1.8%)
		$\chi^2(4) = 11.31$, $p = .023$	

Table H.3: Age groups by gender for QHAPDC and QRCD

		Gender	
		Female	Male
Road user	0 – 16	394 (7.9%)	733 (8.5%)
	17 – 24	712 (14.4%)	1,063 (12.4%)
	25 – 59	2,879 (58.1%)	5,558 (64.7%)
	60+	971 (19.6%)	1,239 (14.4%)
		$\chi^2(3) = 84.33$, $p < .001$, $\phi_c = .08$	

Table H.4: Road user by ARIA+ for QHAPDC and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Road user	Driver	2,952 (40.8%)	1,388 (41.9%)	1,025 (43.9%)	156 (44.4%)	106 (48.8%)
	Motorcyclist	1,540 (21.3%)	894 (27.0%)	507 (21.7%)	63 (17.9%)	35 (16.1%)
	Bicyclist	896 (12.4%)	302 (9.1%)	204 (8.7%)	22 (6.3%)	4 (1.8%)
	Pedestrian	621 (8.6%)	140 (4.2%)	133 (5.7%)	3 (0.9%)	2 (0.9%)
	Passenger	1,232 (17.0%)	587 (17.7%)	464 (19.9%)	107 (30.5%)	70 (32.3%)
		$\chi^2(16) = 270.19$, $p < .001$, $\phi_c = .07$				

Table H.5: Age by ARIA+ for QHAPDC and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Age group	0 – 16	543 (7.5%)	290 (8.7%)	253 (10.7%)	28 (7.8%)	14 (6.1%)
	17 – 24	913 (12.6%)	503 (15.1%)	291 (12.3%)	42 (11.7%)	26 (11.4%)
	25 – 59	4,648 (64.0%)	1,974 (59.1%)	1,450 (61.5%)	232 (64.4%)	132 (57.9%)
	60+	1,156 (15.9%)	575 (17.2%)	365 (15.5%)	58 (16.1%)	56 (24.6%)
		$\chi^2(12) = 60.99$, $p < .001$, $\phi_c = .04$				

Table H.6: Age by road user for QHAPDC and QRCD

		Road user				
		Driver	Motorcyclist	Cyclist	Pedestrian	Passenger
Age group	0 – 16	10 (0.2%)	134 (4.4%)	357 (25.1%)	173 (19.6%)	443 (18.5%)
	17 – 24	627 (11.3%)	282 (9.3%)	188 (13.2%)	145 (16.5%)	500 (20.9%)
	25 – 59	3,780 (68.1%)	2,305 (76.0%)	681 (47.8%)	398 (45.2%)	1,119 (46.8%)
	60+	1,135 (20.4%)	313 (10.3%)	199 (14.0%)	165 (18.7%)	330 (13.8%)
		$\chi^2(12) = 2009.18$, $p < .001$, $\phi_c = .23$				

Table H.7: Road user type by broad severity for QHAPDC and QRCD

		Broad severity	
		Fatality	Hospitalisation
Road user	Driver	177 (44.6%)	5,450 (41.7%)
	Motorcyclist	75 (18.9%)	2,964 (22.7%)
	Bicyclist	11 (2.8%)	1,418 (10.9%)
	Pedestrian	52 (13.1%)	847 (6.5%)
	Passenger	82 (20.7%)	2,378 (18.2%)
		$\chi^2(4) = 53.42, p < .001, \phi_c = .06$	

Table H.8: ARIA+ by broad severity for QHAPDC and QRCD

		Broad severity	
		Fatality	Hospitalisation
ARIA+	Major Cities	138 (34.3%)	7,226 (54.2%)
	Inner Regional	121 (30.1%)	3,268 (24.5%)
	Outer Regional	101 (25.1%)	2,282 (17.1%)
	Remote	29 (7.2%)	333 (2.5%)
	Very Remote	13 (3.2%)	216 (1.6%)
		$\chi^2(4) = 87.08, p < .001, \phi_c = .08$	

Table H.9: Age groups by broad severity for QHAPDC and QRCD

		Broad severity	
		Fatality	Hospitalisation
Road user	0 – 16	23 (6.0%)	1,105 (8.4%)
	17 – 24	43 (11.2%)	1,732 (13.2%)
	25 – 59	238 (61.8%)	8,199 (62.3%)
	60+	81 (21.0%)	2,129 (16.2%)
		$\chi^2(3) = 9.20, p = .027$	

Table H.10: Gender by broad severity for QHAPDC and QRCD

		Broad severity	
		Fatality	Hospitalisation
Gender	Female	109 (27.2%)	4,930 (37.0%)
	Male	292 (72.8%)	8,393 (63.0%)
		$\chi^2(1) = 16.16, p < .001, \phi_c = .03$	

Table H.11: Road user type by seriousness for QHAPDC and QRCD

		Seriousness	
		Serious	Non-serious
Road user	Driver	377 (35.8%)	2,162 (32.0%)
	Motorcyclist	239 (22.7%)	1,978 (29.2%)
	Bicyclist	131 (12.5%)	1,030 (15.2%)
	Pedestrian	112 (10.6%)	452 (6.7%)
	Passenger	193 (18.3%)	1,144 (16.9%)
		$\chi^2(4) = 43.51$, $p < .001$, $\phi_c = .08$	

Table H.12: ARIA+ by seriousness for QHAPDC and QRCD

		Seriousness	
		Serious	Non-serious
ARIA+	Major Cities	691 (62.9%)	3,697 (52.9%)
	Inner Regional	227 (20.7%)	1,850 (26.4%)
	Outer Regional	147 (13.4%)	1,189 (17.0%)
	Remote	20 (1.8%)	166 (2.4%)
	Very Remote	13 (1.2%)	89 (1.3%)
		$\chi^2(4) = 38.90$, $p < .001$, $\phi_c = .07$	

Table H.13: Age groups by seriousness for QHAPDC and QRCD

		Seriousness	
		Serious	Non-serious
Road user	0 – 16	75 (7.0%)	740 (10.7%)
	17 – 24	144 (13.4%)	930 (13.5%)
	25 – 59	620 (57.6%)	4,434 (61.3%)
	60+	238 (22.1%)	1,008 (14.6%)
		$\chi^2(3) = 48.63$, $p < .001$, $\phi_c = .08$	

Table H.14: Gender by seriousness for QHAPDC and QRCD

		Seriousness	
		Serious	Non-serious
Gender	Female	306 (27.9%)	2,299 (32.9%)
	Male	792 (72.1%)	4,691 (67.1%)
		$\chi^2(1) = 43.51$, $p = .001$	

Table H.15: Broad severity by seriousness for QHAPDC and QRCD

		Seriousness	
		Serious	Non-serious
Broad severity	Fatal	113 (10.3%)	258 (3.7%)
	Hospitalisation	985 (89.7%)	6,733 (96.3%)
		$\chi^2(1) = 94.49$, $p < .001$, $\phi_c = .11$	

eARF and QRCD

Table H.16: Road user type by gender for eARF and QRCD

		Gender	
		Female	Male
Road user	Driver	6,913 (59.7%)	6,753 (54.5%)
	Motorcyclist	555 (4.8%)	1,903 (15.4%)
	Bicyclist	243 (2.1%)	944 (7.6%)
	Pedestrian	384 (3.3%)	515 (4.2%)
	Passenger	3,478 (30.1%)	2,277 (18.4%)
		$\chi^2(4) = 1398.49$, $p < .001$, $\phi_c = .24$	

Table H.17: ARIA+ by gender for eARF and QRCD

		Gender	
		Female	Male
ARIA+	Major Cities	8,537 (57.2%)	8,125 (52.6%)
	Inner Regional	3,527 (23.6%)	3,679 (23.8%)
	Outer Regional	2,472 (16.6%)	2,985 (19.3%)
	Remote	247 (1.7%)	409 (2.6%)
	Very Remote	153 (1.0%)	254 (1.6%)
		$\chi^2(4) = 117.96$, $p < .001$, $\phi_c = .06$	

Table H.18: Age groups by gender for eARF and QRCD

		Gender	
		Female	Male
Road user	0 – 16	1,076 (7.3%)	1,052 (7.0%)
	17 – 24	4,028 (27.4%)	4,015 (26.6%)
	25 – 59	7,661 (52.2%)	8,179 (54.3%)
	60+	1,916 (13.1%)	1,822 (12.1%)
		$\chi^2(3) = 14.56$, $p = .002$	

There was a relationship between road user and ARIA+ [$\chi^2(16) = 283.06$, $p < .001$, $\phi_c = .05$].

Table H.19: Road user by ARIA+ for eARF and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Road user	Driver	7,860 (58.3%)	3,099 (56.5%)	2,219 (53.3%)	308 (53.7%)	187 (53.6%)
	Motorcyclist	1,274 (9.4%)	625 (11.4%)	507 (12.2%)	41 (7.1%)	20 (5.7%)
	Bicyclist	769 (5.7%)	210 (3.8%)	201 (4.8%)	10 (1.7%)	1 (0.3%)
	Pedestrian	611 (4.5%)	156 (2.8%)	123 (3.0%)	8 (1.4%)	2 (0.6%)
	Passenger	2,978 (22.1%)	1,393 (25.4%)	1,117 (26.8%)	207 (36.1%)	139 (39.8%)
		$\chi^2(16) = 283.06$, $p < .001$, $\phi_c = .05$				

There was a relationship between age and ARIA+ [$\chi^2(12) = 157.26$, $p < .001$, $\phi_c = .07$].

Table H.20: Age by ARIA+ for eARF and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Age group	0 – 16	1,065 (6.5%)	527 (7.5%)	475 (9.0%)	43 (6.7%)	19 (4.9%)
	17 – 24	4,397 (26.8%)	1,920 (27.3%)	1,423 (27.0%)	199 (30.9%)	108 (27.6%)
	25 – 59	9,083 (55.4%)	3,496 (49.7%)	2,710 (51.3%)	332 (51.5%)	205 (52.4%)
	60+	1,836 (11.2%)	1,097 (15.6%)	671 (12.7%)	71 (11.0%)	59 (15.1%)
		$\chi^2(12) = 157.26$, $p < .001$, $\phi_c = .07$				

Table H.21: Age by road user for eARF and QRCD

		Road user				
		Driver	Motorcyclist	Cyclist	Pedestrian	Passenger
Age group	0 – 16	16 (0.4%)	62 (2.6%)	197 (16.9%)	171 (19.3%)	1,144 (20.5%)
	17 – 24	3,517 (25.9%)	486 (20.1%)	224 (19.2%)	261 (29.4%)	1,836 (32.9%)
	25 – 59	8,178 (60.3%)	1,679 (69.5%)	648 (55.5%)	332 (37.4%)	1,974 (35.4%)
	60+	1,808 (13.3%)	190 (7.9%)	99 (8.5%)	124 (14.0%)	628 (11.3%)
		$\chi^2(12) = 3663.75$, $p < .001$, $\phi_c = .22$				

QISU and QRCD

Table H.22: Road user type by gender for QISU and QRCD

		Gender	
		Female	Male
Road user	Driver	5,982 (60.9%)	5,985 (52.1%)
	Motorcyclist	346 (3.5%)	1,901 (16.6%)
	Bicyclist	251 (2.6%)	1,094 (9.5%)
	Pedestrian	397 (4.0%)	557 (4.9%)
	Passenger	2,848 (29.0%)	1,940 (16.9%)
		$\chi^2(4) = 1685.38$, $p < .001$, $\phi_c = .28$	

Table H.23: ARIA+ by gender for QISU and QRCD

		Gender	
		Female	Male
ARIA+	Major Cities	5,859 (59.9%)	6,252 (54.8%)
	Inner Regional	2,088 (21.3%)	2,499 (21.9%)
	Outer Regional	1,497 (15.3%)	2,003 (17.6%)
	Remote	192 (2.0%)	351 (3.1%)
	Very Remote	150 (1.5%)	307 (2.7%)
		$\chi^2(4) = 99.09$, $p < .001$, $\phi_c = .07$	

Table H.24: Age groups by gender for QISU and QRCD

		Gender	
		Female	Male
Road user	0 – 16	807 (8.3%)	950 (8.3%)
	17 – 24	2,654 (27.2%)	3,042 (26.7%)
	25 – 59	5,162 (52.9%)	6,209 (54.5%)
	60+	1,131 (11.6%)	1,185 (10.4%)
		$\chi^2(3) = 9.80, p = .020$	

Table H.25: Road user by ARIA+ for QISU and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Road user	Driver	6,956 (57.2%)	2,610 (56.6%)	1,878 (53.4%)	292 (53.7%)	205 (44.9%)
	Motorcyclist	1,117 (9.2%)	562 (12.2%)	449 (12.8%)	40 (7.4%)	62 (13.6%)
	Bicyclist	870 (7.1%)	217 (4.7%)	210 (6.0%)	9 (1.7%)	32 (7.0%)
	Pedestrian	659 (5.4%)	147 (3.2%)	125 (3.6%)	8 (1.5%)	9 (2.0%)
	Passenger	2,567 (21.1%)	1,074 (23.3%)	853 (24.3%)	195 (35.8%)	149 (32.6%)
		$\chi^2(16) = 274.53, p < .001, \phi_c = .06$				

Table H.26: Age by ARIA+ for QISU and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Age group	0 – 16	980 (8.2%)	348 (7.6%)	335 (9.7%)	37 (6.9%)	45 (9.9%)
	17 – 24	3,151 (26.2%)	1,258 (27.5%)	948 (27.4%)	171 (31.8%)	138 (30.5%)
	25 – 59	6,696 (55.7%)	2,341 (51.3%)	1,785 (51.5%)	276 (51.3%)	219 (48.3%)
	60+	1,190 (9.9%)	620 (13.6%)	397 (11.5%)	54 (10.0%)	51 (11.3%)
		$\chi^2(12) = 85.47, p < .001, \phi_c = .06$				

Table H.27: Age by road user for QISU and QRCD

		Road user				
		Driver	Motorcyclist	Cyclist	Pedestrian	Passenger
Age group	0 – 16	24 (0.2%)	48 (2.1%)	376 (28.1%)	218 (23.1%)	1,093 (23.4%)
	17 – 24	3,091 (25.9%)	512 (22.9%)	274 (20.5%)	270 (28.6%)	1,550 (33.2%)
	25 – 59	7,326 (61.3%)	1,552 (69.4%)	606 (45.3%)	331 (35.0%)	1,556 (33.3%)
	60+	1,512 (12.6%)	123 (5.5%)	82 (6.1%)	126 (13.3%)	473 (10.1%)
		$\chi^2(12) = 4109.24$, $p < .001$, $\phi_c = .26$				

Table H.28: Road user type by broad severity for QISU and QRCD

		Broad severity		
		Fatality	Hospitalisation	Other injury
Road user	Driver	152 (45.5%)	3,679 (52.6%)	8,140 (57.8%)
	Motorcyclist	63 (18.9%)	1,025 (14.7%)	1,166 (8.3%)
	Bicyclist	8 (2.4%)	402 (5.8%)	938 (6.7%)
	Pedestrian	40 (12.0%)	452 (6.5%)	463 (3.3%)
	Passenger	71 (21.3%)	1,432 (20.5%)	3,367 (23.9%)
		$\chi^2(8) = 419.47$, $p < .001$, $\phi_c = .10$		

Table H.29: ARIA+ by broad severity for QISU and QRCD

		Broad severity		
		Fatality	Hospitalisation	Other injury
ARIA+	Major Cities	96 (28.8%)	3,615 (51.9%)	8,458 (60.4%)
	Inner Regional	106 (31.8%)	1,633 (23.5%)	2,871 (20.5%)
	Outer Regional	89 (26.7%)	1,300 (18.7%)	2,126 (15.2%)
	Remote	29 (8.7%)	221 (3.2%)	294 (2.1%)
	Very Remote	13 (3.9%)	190 (2.7%)	254 (1.8%)
		$\chi^2(8) = 290.84$, $p < .001$, $\phi_c = .08$		

Table H.30: Age groups by broad severity for QISU and QRCD

		Broad severity		
		Fatality	Hospitalisation	Other injury
Road user	0 – 16	21 (6.3%)	473 (6.8%)	1,265 (9.1%)
	17 – 24	78 (23.4%)	1,870 (26.8%)	3,749 (27.1%)
	25 – 59	179 (53.6%)	3,729 (53.5%)	7,463 (53.9%)
	60+	56 (16.8%)	901 (12.9%)	1,359 (9.8%)
		$\chi^2(6) = 85.77$, $p < .001$, $\phi_c = .05$		

Table H.31: Gender by broad severity for QISU and QRCD

		Broad severity		
		Fatality	Hospitalisation	Other injury
Gender	Female	90 (27.0%)	2,972 (42.5%)	6,762 (48.4%)
	Male	243 (73.0%)	4,015 (57.5%)	7,219 (51.6%)
		$\chi^2(2) = 113.34$, $p < .001$, $\phi_c = .07$		

Table H.32: Road user type by seriousness for QISU and QRCD

		Seriousness	
		Serious	Non-serious
Road user	Driver	325 (61.9%)	2,754 (46.4%)
	Motorcyclist	43 (8.2%)	732 (12.3%)
	Bicyclist	26 (5.0%)	674 (11.4%)
	Pedestrian	36 (6.9%)	307 (5.2%)
	Passenger	95 (18.1%)	1,469 (24.7%)
		$\chi^2(4) = 60.89$, $p < .001$, $\phi_c = .10$	

Table H.33: ARIA+ by seriousness for QISU and QRCD

		Seriousness	
		Serious	Non-serious
ARIA+	Major Cities	233 (45.2%)	3,249 (55.6%)
	Inner Regional	133 (25.8%)	1,258 (21.5%)
	Outer Regional	112 (21.7%)	991 (16.9%)
	Remote	18 (3.5%)	135 (2.3%)
	Very Remote	20 (3.9%)	215 (3.7%)
		$\chi^2(4) = 290.84$, $p < .001$, $\phi_c = .08$	

Table H.34: Age groups by seriousness for QISU and QRCD

		Seriousness	
		Serious	Non-serious
Road user	0 – 16	60 (11.6%)	832 (14.2%)
	17 – 24	138 (26.6%)	1,615 (27.6%)
	25 – 59	247 (47.6%)	2,905 (49.7%)
	60+	74 (14.3%)	495 (8.5%)
		$\chi^2(3) = 20.92$, $p < .001$, $\phi_c = .06$	

Table H.35: Gender by seriousness for QISU and QRCD

		Seriousness	
		Serious	Non-serious
Gender	Female	219 (41.9%)	2,514 (42.5%)
	Male	304 (58.1%)	3,395 (57.5%)
		$\chi^2(1) = 0.09$, $p = .766$	

Table H.36: Broad severity by seriousness for QISU and QRCD

		Seriousness	
		Serious	Non-serious
Broad severity	Fatal	58 (11.0%)	245 (4.1%)
	Hospitalisation	188 (35.8%)	1,221 (20.6%)
	Other injury	279 (53.1%)	4,470 (75.3%)
		$\chi^2(1) = 94.49$, $p < .001$, $\phi_c = .11$	

EDIS and QRCD

Table H.37: Road user type by gender for EDIS and QRCD

		Gender	
		Female	Male
Road user	Driver	3,026 (47.5%)	3,314 (27.6%)
	Motorcyclist	612 (9.6%)	4,110 (36.2%)
	Bicyclist	518 (8.1%)	2,369 (20.9%)
	Pedestrian	262 (4.1%)	379 (3.3%)
	Passenger	1,948 (30.6%)	1,363 (12.0%)
		$\chi^2(4) = 2715.31$, $p < .001$, $\phi_c = .39$	

Table H.38: ARIA+ by gender for EDIS and QRCD

		Gender	
		Female	Male
ARIA+	Major Cities	5,490 (55.2%)	8,165 (52.2%)
	Inner Regional	2,683 (27.0%)	4,415 (28.2%)
	Outer Regional	1,499 (15.1%)	2,472 (15.8%)
	Remote	113 (1.1%)	222 (1.4%)
	Very Remote	167 (1.7%)	369 (2.4%)
		$\chi^2(4) = 32.91$, $p < .001$, $\phi_c = .04$	

Table H.39: Age groups by gender for EDIS and QRCD

		Gender	
		Female	Male
Road user	0 – 16	1,026 (9.9%)	1,858 (11.4%)
	17 – 24	3,034 (29.3%)	4,715 (29.0%)
	25 – 59	5,060 (48.9%)	8,338 (51.3%)
	60+	1,228 (11.9%)	1,345 (8.3%)
		$\chi^2(3) = 105.20$, $p < .001$, $\phi_c = .06$	

Table H.40: Road user by ARIA+ for EDIS and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Road user	Driver	3,242 (36.0%)	1,564 (32.9%)	996 (35.8%)	136 (45.9%)	126 (34.8%)
	Motorcyclist	2,187 (24.3%)	1,468 (30.9%)	751 (27.0%)	51 (17.2%)	113 (31.2%)
	Bicyclist	1,605 (17.8%)	767 (16.2%)	401 (14.4%)	10 (3.4%)	32 (8.8%)
	Pedestrian	421 (4.7%)	111 (2.3%)	95 (3.4%)	2 (0.7%)	3 (0.8%)
	Passenger	1,557 (17.3%)	838 (17.6%)	538 (19.3%)	97 (32.8%)	88 (24.3%)
		$\chi^2(16) = 256.40$, $p < .001$, $\phi_c = .06$				

Table H.41: Age by ARIA+ for EDIS and QRCD

		ARIA+				
		Major Cities	Inner Regional	Outer Regional	Remote	Very Remote
Age group	0 – 16	1,276 (9.3%)	943 (13.3%)	504 (12.7%)	19 (5.7%)	70 (13.1%)
	17 – 24	3,981 (29.2%)	2,104 (29.6%)	1,084 (27.3%)	113 (33.9%)	148 (27.7%)
	25 – 59	7,195 (52.7%)	3,261 (45.9%)	1,994 (50.3%)	171 (51.4%)	265 (49.5%)
	60+	1,200 (8.8%)	789 (11.1%)	386 (9.7%)	30 (9.0%)	52 (9.7%)
		$\chi^2(12) = 165.89$, $p < .001$, $\phi_c = .06$				

Table H.42: Age by road user for EDIS and QRCD

		Road user				
		Driver	Motorcyclist	Cyclist	Pedestrian	Passenger
Age group	0 – 16	18 (0.3%)	436 (9.2%)	978 (33.9%)	114 (17.8%)	583 (17.7%)
	17 – 24	1,785 (29.0%)	1,332 (28.2%)	687 (23.8%)	199 (31.0%)	1,229 (37.2%)
	25 – 59	3,514 (57.0%)	2,762 (58.5%)	1,058 (36.7%)	223 (34.8%)	1,136 (34.4%)
	60+	843 (13.7%)	191 (4.0%)	163 (5.6%)	105 (16.4%)	352 (10.7%)
		$\chi^2(12) = 2865.94$, $p < .001$, $\phi_c = .23$				

Table H.43: Road user type by broad severity for EDIS and QRCD

		Broad severity	
		Fatality	Hospitalisation
Road user	Driver	153 (44.9%)	6,007 (34.6%)
	Motorcyclist	67 (19.6%)	4,655 (26.8%)
	Bicyclist	9 (2.6%)	2,878 (16.6%)
	Pedestrian	41 (12.0%)	600 (3.5%)
	Passenger	71 (20.8%)	3,244 (18.7%)
		$\chi^2(4) = 2865.94$, $p < .001$, $\phi_c = .23$	

Table H.44: ARIA+ by broad severity for EDIS and QRCD

		Broad severity	
		Fatality	Hospitalisation
ARIA+	Major Cities	102 (29.3%)	13,555 (53.7%)
	Inner Regional	110 (31.6%)	6,989 (27.7%)
	Outer Regional	94 (27.0%)	3,881 (15.4%)
	Remote	29 (8.3%)	306 (1.2%)
	Very Remote	13 (3.7%)	523 (2.1%)
		$\chi^2(4) = 207.69$, $p < .001$, $\phi_c = .09$	

Table H.45: Age groups by broad severity for EDIS and QRCD

		Broad severity	
		Fatality	Hospitalisation
Road user	0 – 16	25 (7.1%)	2,860 (10.9%)
	17 – 24	80 (22.9%)	7,671 (29.2%)
	25 – 59	185 (52.9%)	13,215 (50.3%)
	60+	60 (17.1%)	2,513 (9.6%)
		$\chi^2(3) = 30.19$, $p < .001$, $\phi_c = .03$	

Table H.46: Gender by broad severity for EDIS and QRCD

		Broad severity	
		Fatality	Hospitalisation
Gender	Female	95 (27.2%)	10,260 (39.1%)
	Male	254 (72.8%)	16,009 (60.9%)
		$\chi^2(1) = 20.30$, $p < .001$, $\phi_c = .03$	

Table H.32: Road user type by seriousness for EDIS and QRCD

		Seriousness	
		Serious	Non-serious
Road user	Driver	155 (22.9%)	2,883 (26.1%)
	Motorcyclist	246 (36.4%)	3,440 (31.1%)
	Bicyclist	112 (16.6%)	2,450 (22.2%)
	Pedestrian	46 (6.8%)	250 (2.3%)
	Passenger	117 (17.3%)	2,032 (18.4%)
		$\chi^2(4) = 69.72$, $p < .001$, $\phi_c = .08$	

Table H.33: ARIA+ by seriousness for EDIS and QRCD

		Seriousness	
		Serious	Non-serious
ARIA+	Major Cities	527 (50.7%)	9,856 (53.7%)
	Inner Regional	285 (27.4%)	5,342 (29.1%)
	Outer Regional	181 (17.4%)	2,656 (14.5%)
	Remote	20 (1.9%)	136 (0.7%)
	Very Remote	26 (2.5%)	369 (2.0%)
		$\chi^2(4) = 26.74$, $p < .001$, $\phi_c = .04$	

Table H.34: Age groups by seriousness for EDIS and QRCD

		Seriousness	
		Serious	Non-serious
Road user	0 – 16	71 (6.1%)	2,460 (12.8%)
	17 – 24	317 (27.4%)	5,822 (30.3%)
	25 – 59	587 (50.7%)	9,389 (48.8%)
	60+	183 (15.8%)	1,552 (8.1%)
		$\chi^2(3) = 119.57$, $p < .001$, $\phi_c = .08$	

Table H.35: Gender by seriousness for EDIS and QRCD

		Seriousness	
		Serious	Non-serious
Gender	Female	354 (30.6%)	7,411 (38.6%)
	Male	803 (69.4%)	11,808 (61.4%)
		$\chi^2(1) = 29.35$, $p < .001$, $\phi_c = .04$	

Table H.36: Broad severity by seriousness for EDIS and QRCD

		Seriousness	
		Serious	Non-serious
Broad severity	Fatal	71 (22.4%)	246 (77.6%)
	Hospitalisation	1,087 (5.4%)	18,977 (94.6%)
		$\chi^2(1) = 167.90$, $p < .001$, $\phi_c = .09$	

Appendix I – Pull Out Supplement for Chapter 7 Methodology

Table I.1: Data selection criteria and coding for each data collection in Study 3 (Chapter 7)

	General	QRCD	QHAPDC	eARF	QISU	EDIS

Selection of road crashes		All casualties	All acute admissions with ICD-10-AM External Cause Codes from V00-V89 and fourth character of 'traffic'	*Case nature* (Bicycle Collision ; Motor Vehicle Collision; Motorcycle Collision; Pedestrian Collision) and *location type* (street; public transport; vehicle)	*External definition* (Motor vehicle – driver;
					Motor vehicle – passenger;
					Motorcycle – driver;
					Motorcycle – passenger;
					Pedal cyclist or pedal cyclist passenger;
					Pedestrian) and *type of place* (street/highway)
						Presenting problem keyword search (e.g., car, motorbike, pedestrian) without exclusion terms (e.g., off-road, track)

Age	5 year age groups (with the exception 85+).	Provided in single years re-coded into 5 year age groups	Retained as coded	Coded from *date of birth*	Provided in single years re-coded into 5 year age groups	Provided in single years re-coded into 5 year age groups
Gender	1 = Female; 2 = Male	Retained as coded	Recoded to 1 = Female; 2 = Male	Retained as coded	Retained as coded	Retained as coded
Severity	1. *Broad severity* (fatality; hospitalisation; other injury). *2. The Abbreviated Injury Scale (AIS)* (1 = minor; 2 =	1. *Casualty severity* (1= fatality; 2 = hospitalisation; 3 = medical treatment; 4 = minor injury), with 'medical treatment' and	1. *Mode of separation* ('died in hospital' = fatality; all other cases = hospitalised) 2 and 3. *Principal*	Not able to be coded	1. *Mode of separation* (died in ED, dead on arrival = fatality; all other cases = hospitalisation) 2 and 3. *Principal*	1. *Departure status* (died in ED, dead on arrival = fatality; all other cases = hospitalisation) 2 and 3. *Principal diagnosis* ICD-10-

	moderate; 3 = serious; 4 = severe; 5 = critical; and 6 = maximum). Mapped to principal diagnosis ICD-10 codes in the data (when available). *3. Survival Risk Ratios (SRR)* - estimate of the probability of death from 0 (no chance of survival) to 1 (100% chance of survival). SRRs were mapped to principal diagnosis ICD codes.	'minor injury' collapsed into 'other injury' 2 and 3. *AIS* and *SRR* were coded using the *injury description* variable. Principal diagnosis ICD-10-AM code mapped to each injury description. ICD codes then mapped to AIS and a SRR	*diagnosis* ICD-10-AM codes mapped to the *AIS* and a *SRR*		*diagnosis* ICD-10-AM codes mapped to the *AIS* and a *SRR*	AM codes mapped to the *AIS* and a *SRR*
ARIA+	ARIA+ (1 = Major Cities; 2 = Inner Regional; 3 = Outer Regional; 4 = Remote; 5 = Very Remote).	Retained as coded	Retained as coded	Postcode mapped to ARIA+ using data from the Australian Bureau of Statistics (2013).	Postcode mapped to ARIA+ using data from the Australian Bureau of Statistics (2013).	Postcode mapped to ARIA+ using data from the Australian Bureau of Statistics (2013).
Road user	1 = Driver, 2 = Motorcyclist, 3 = Cyclist, 4 = Pedestrian; 5 = Car passenger	*Casualty road user type*. Coding was retained with the exception of 'motorcycle pillions' and	Second and fourth characters of the ICD-10-AM *external cause code*.	Combination of *case nature*, *vehicle type*, and *comments*	*External code* (motor vehicle – driver = driver; motorcycle – driver and motorcycle – passenger =	*Presenting problem* text search (e.g., driver = driver; motorcycle, MCA, MBA = motorcyclist;

		‘bicycle pillions’ recoded into ‘motorcyclist’ and ‘cyclist’ respectively.			motorcyclist; pedal cyclist or pedal cyclist passenger = cyclist; pedestrian = pedestrian; motor vehicle passenger = passenger)	bicycle, PBS, PBA = cyclist; passenger = passenger; none of the keywords = unspecified)
Collision	0 = no collision 1 = collision	All cases with a *crash nature* of: angle; rear-end; head-on; sideswipe; and hit pedestrian.	Non-collisions were all cases with an *external cause code* of V17, V18, V28, V38, V48, V58, V68, and V78. Collisions were all other cases.	Not able to be coded	Collisions were all cases with a mechanism of : contact with moving object or contact with a person	Not able to be coded

Appendix J – Summary of completeness and consistency Chapter 5

Table J.1: Summary of completeness and consistency for variables in each data collection

	QRCD	QHAPDC	eARF	QISU	EDIS
Injury description/nature of injury	73.4% unknown/unspecified < likely for males, unknown gender; cyclists and pedestrians; fatalities		23.9% unknown/unspecified > likely for unknown gender; 0-4 years; drivers		
Traffic		14.1% unspecified > likely for males < likely for Major Cities and fatalities			
Place		33.3% unspecified > likely for 0-14 years, motorcyclists, and cyclists		13.4% unspecified > likely for males, motorcyclists < likely for 0-9 years and 50-84 years	
Activity		75.2% > likely for males, 0-4 years, 65+ years, drivers, passengers, and pedestrians		32.0% unspecified > likely for females, drivers, pedestrians, passengers, and Inner Regional < likely for 5-14 years	
Road user (presenting problem)					41.7% missing/unspecified > females < 5-19 years