Title

Evaluating Research on Data Linkage to Assess Underreporting of Pedestrian and Bicyclist Injury in Police Crash Data

Authors
Doggett, Sarah
Ragland, David R.
Felschundneff, Grace

Publication Date
2018-11-15

Peer reviewed

Evaluating Research on Data Linkage to Assess Underreporting of Pedestrian and Bicyclist Injury in Police Crash Data

Sarah Doggett, Corresponding Author
Graduate Student, Safe Transportation Research and Education Center (SafeTREC)
University of California, Berkeley
2614 Dwight Way #7374
Berkeley, CA 94720-7374
Phone: 510-642-0655, Fax: 510-643-9922
doggett_sarah@berkeley.edu

David R. Ragland, PhD, MPH
Adjunct Professor Emeritus, School of Public Health
Director, Safe Transportation Research and Education Center (SafeTREC)
University of California, Berkeley
2614 Dwight Way #7374
Berkeley, CA 94720-7374
Phone: 510-642-0655, Fax: 510-643-9922
davidr@berkeley.edu

Grace Felschundneff
Senior Editor, Safe Transportation Research and Education Center (SafeTREC)
University of California, Berkeley
2614 Dwight Way #7374
Berkeley, CA 94720-7374
Phone: 510-642-0655, Fax: 510-643-9922
gracefelschundneff@gmail.com

Word Count: 4,426 words + 3 tables/figures (250 words each) = 5,176 words

Submission Date: July 30, 2018

ABSTRACT

Traffic safety decisions are based predominantly on information from police collision reports. However, a number of studies suggest that such reports tend to underrepresent bicycle and pedestrian collisions. Underreporting could lead to inaccurate evaluation of crash rates and may under- or overestimate the effects of road safety countermeasures. This review examined ten studies that used data linkage to explore potential underreporting of pedestrian and/or bicyclist injury in police collision reports. Due to variations in definitions of reporting level, periods of study, and study locations, it was difficult to directly compare the studies. Even among the six studies using the hospital link definition, estimates of reporting levels ranged from 44 to 75 percent for pedestrian crashes, and from 7 to 46 percent for bicycle crashes, suggesting a severe underreporting problem. However, few of the studies provided estimates of the error around their reporting level estimates, and as a result, it is difficult to determine the true level of underreporting. It may be that bicycle and pedestrian crashes appear in both police and hospital datasets but are less likely to be linked. Due to linkage error, link rate can only be used to *estimate* reporting level. Without the *variance* of that estimate, the effect of underreporting on traffic safety analyses cannot be accurately determined. Future studies should include estimates of the error present in their data linkage process for greater accuracy of the underreporting in police data. Datasets should be designed for easier linkage with hospital data and other datasets.

Keywords: Data, Linkage, Pedestrian, Bicycle, Crash, Underreporting

INTRODUCTION

Traffic safety decisions are almost universally based on information from police collision reports. However, many researchers believe that police collision reports have limitations, and fail to include all crashes that occur on the road. For example, a number of studies suggest that police collision reports underrepresent bicycle and pedestrian collisions.

According to the Federal Highway Administration (FHWA), traditional crash data sources are insufficient because they exclude both crashes that take place in non-roadway locations (e.g., parking lots, driveways, and sidewalks) and bicycle crashes and pedestrian injuries that do not involve motor vehicles *(1)*.

Ideally, the degree of bias in police collision reports could be measured and accounted for in analyses *(2)*. This process requires both the reporting level and the uncertainty in the estimate of the reporting level to be known *(3)*. Although researchers have estimated reporting level by linking police collision reports with hospital data, few have determined the level of uncertainty surrounding their estimates.

This paper summarizes existing research using data linkage to study pedestrians and bicyclists, explores potential problems concerning linkage, and offers suggestions on how to improve the data linkage process.

METHODOLOGY AND DEFINITIONS

Searches using Google Scholar and Science Direct were conducted to identify potentially relevant articles, published between 1999 and 2017, for the literature review. Search terms included record linkage, data linkage, crash data linkage, pedestrian underreporting, underreporting, crash injury assessment, and police crash data limitations.

The abstracts of the articles returned through this search were evaluated to assess their relevancy to the general topic of crash reporting and heath data linkage. Articles with relevant abstracts were read in full to determine whether they reported findings related to pedestrians and/or bicyclists.

Five of the articles explored data linkage in general but mentioned findings specific to pedestrians and/or bicyclists. Two focused exclusively on bicyclists, two exclusively on pedestrians, and one on both pedestrians and bicyclists.

Most of the studies included in the literature review defined underreporting as hospital records without counterparts in police crash reports. However, Janstrup et al. found that there are also police crash reports that do not have counterparts in hospital records *(4)*. This expands the definition of underreporting to crashes that exist only in one dataset. This concept is illustrated in Figure 1.

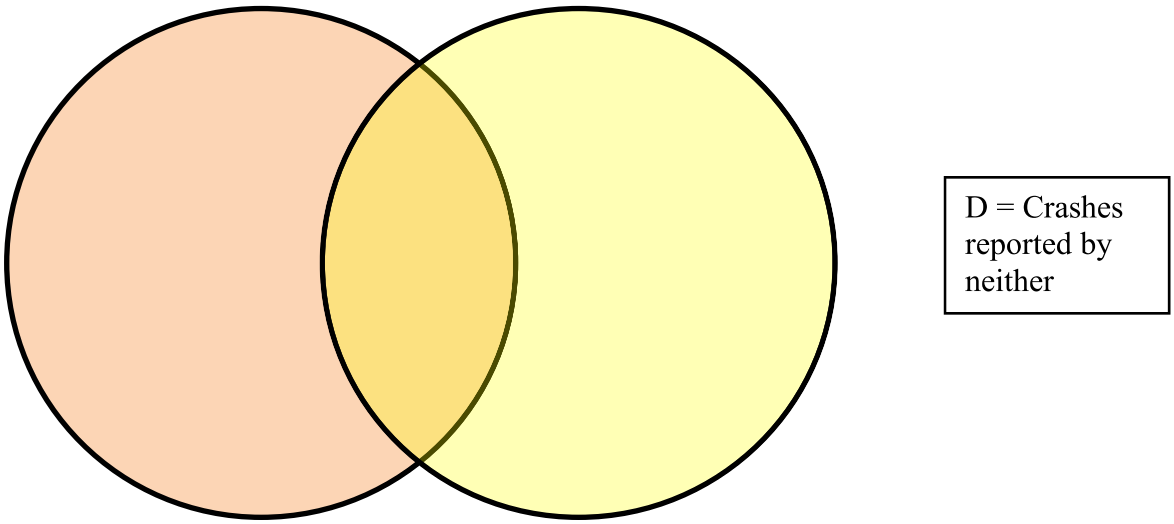

Figure 1 Patterns of Reporting Pedestrian and Bicyclist Crashes by Source

Some crashes may be reported in both datasets but cannot be linked due to errors in data recording—these crashes are *misreported* but cannot not be distinguished from underreported crashes using current linkage techniques. There are also crashes that are *not reported* to either the police or to hospitals (D in Figure 1), due to lack of serious injury or an unwillingness on the part of those involved in the crash to contact the authorities. To identify these unreported crashes, other datasets such as those collected by insurance companies can be analyzed. One of the studies addressed this, and found that relying solely on police and hospital injury data can result in a significant underestimation in the actual number and severity of crashes *(5)*.

Several different definitions of reporting level exist. Using the terminology described in Table 1, these definitions are explained as follows:

- A: Cases only appearing in police crash reports
- B: Cases appearing in both police crash reports and hospital records
- C: Cases only appearing in hospital records
- D: Cases that do not appear in either police crash reports or hospital records

Table 1 Definitions of Reporting Level

Definition of Reporting Level	Formula	Question Addressed
1: Hospital Link Rate	B/(B+C)	Proportion of cases reported in hospital records for which a police crash report is found
2: Police Link Rate	B/(A+B)	Proportion of cases reported in police crash reports for which a hospital record is found
3: Police Ascertainment Rate	(A+B)/(A+B+C)	Proportion of identified cases accounted for in police crash reports
4: Hospital Ascertainment Rate	(B+C)/(A+B+C)	Proportion of identified cases accounted for in hospital records

Several studies used the capture-recapture method to determine reporting level, which is not derived from those shown in Table 1. Originally, this method was used to estimate animal populations based on how many animals were captured and then later recaptured in a least two random samples *(6)*. However, researchers have noted that this method may not be appropriate for estimating reporting levels because some of its underlying assumptions, such as a closed study population and an equal probability for each individual to be captured in each sample, may be violated *(6,4)*.

The definition used can significantly affect the estimated reporting level *(2)*. As an example, assume that 50 cases only appear in police reports (A), 20 cases appear in both police and hospital reports (B), and 30 cases only appear in hospital reports (C). Under this scenario, hospital link rate would be 0.4, police link rate would be 0.29, police ascertainment rate would be 0.7, and hospital ascertainment would be 0.5. While definitions 3 and 4 are theoretically the most correct definitions of reporting level, few studies apply these when estimating reporting level.

LITERATURE REVIEW

Due to the different definitions of reporting level, different periods of study, and different study locations, it is difficult to directly compare the ten studies. As shown in Table 2, six of the ten studies used hospital link rate as their definition of reporting level, while two used capture-recapture, two used police link rate, two used police ascertainment rate, and one used hospital ascertainment rate (this adds up to more than ten because the article by Short and Caufield reported on four different definitions). Even among the six studies using the hospital link definition, estimates of reporting levels ranged from 44 to 75 percent for pedestrian crashes, and from 7 to 46 percent for bicycle crashes.

However, the articles all agreed that police collision reports have limitations and underreport certain types of crashes, especially those involving pedestrians and bicyclists. For example, Stutts and Hunter *(9)* found that police were unlikely to be contacted about pedestrian incidents not involving motor vehicles, although more than two-thirds of serious injuries fell into this category. Sciortino *(7)* found that pedestrian injury victims who were African American, male, or sustained minor injuries were less likely to be included in police crash records. The author suggested that such bias was likely due to the reluctance on the part of pedestrians to summon the police if the police were not initially present at the crash scene. Tarko and Azam *(8)* found that pedestrians were less likely to be included in the linked database if they were struck by vehicles on state roads, at Y intersections, or on divided roadways. However, they were more likely to be included in the database if they were struck while crossing a road instead of walking along its edge or standing outside of the roadway.

Drivers are more likely to be included in crash reports than are cyclists *(9)*. Langley *(10)* found that only 22 percent of bicyclist crashes on public roads could be linked to a crash report—the percentage increased to 54 percent when it included crashes that also involved motor vehicles. Because crash reports are mainly focused on incidents that involve motor vehicles and which occur on public roadways, they likely capture fewer than one-third of bicyclist injury cases serious enough to require medical treatment *(1)*. According to Janstrup et al. *(4)*, the underreporting rate for crashes involving cyclists in Denmark was 14 percent for those resulting in serious injuries, and 7 percent for those resulting in slight injuries.

The literature identified several possible reasons for underreporting. According to Langley *(10)*, motor vehicle crashes listed in hospital data are assumed to have occurred on roadways unless another location is specified, thus hospital derived estimates may overstate the number of motor vehicle crashes on public roads and understate crashes that occur in other locations or those that do not involve motor vehicles. Another possible reason for underreporting is that pedestrian and bicycle crashes are less likely to result in insurance compensation than are motor vehicle crashes. Lujic et al. *(11)* found that those entitled to insurance payouts had higher linkage rates than those who were not, presumably because police reports were required as a condition for receiving compensation. According to Watson et al. *(12)*, it is possible that the severity of injuries resulting from a collision with another vehicle is likely to be more serious, and therefore more likely to be reported.

Several studies have shown that compared with hospital data, police crash reports do not accurately report injury severity. Injury classifications in crash reports are usually based on the KABCO scale, which is less nuanced than the Injury Severity Score that many hospitals use *(13)*. Additionally, KABCO classifications are made at the scene of the crash by officers who typically lack medical training—therefore, less visible but life-threatening injuries such as internal bleeding may be misclassified as non-severe, while more obvious minor injuries such as minor lacerations with profuse bleeding may be misclassified as severe *(13)*. Another problem related to estimation of injury severity, is that the injury classification in police reports is static and does not necessarily reflect subsequent developments *(14)*.

Table 2 Summary of Data Linkage Articles Relevant to Pedestrian and Bicyclist Safety

Source	Study Period	Study Location	Focus	Definition Used	Findings/Conclusions
Conderino, Fung, Sedlar & Norton, 2017	2009-2015	New York City	General	Hospital Link Rate	• 50% of hospital reports involving a pedestrian crash linked to a police report • 45% of hospital reports involving a bicyclist crash linked to a police report • Sensitivity - 74% • Specificity - 93%
Janstrup et al., 2016	2003-2007	Funen, Denmark	General	Capture-Recapture	• Compared with car occupants, pedestrians are more likely to appear in both police and hospital databases; bicyclists are more likely to appear in either • Only 7% of bicycle crashes resulting in slight injury and only 15% of bicycle crashes resulting in severe injury are reported by the police
Langley, 2003	1995-1999	New Zealand	Bicyclist	Hospital Link Rate	• Only 22% of bicycle crashes on public roads could be linked to a crash report • When limited to bicycle crashes on public roads involving motor vehicles, 54% could be linked to a crash report
Lujic, Finch, Boufous, Hayen & Dunsmuir, 2008	2000-2001	New South Wales	General	Hospital Link Rate	• 69% of road traffic crashes were linked to police records • Drivers were most likely to have their hospital records linked to police records (83%) • 46% of bicyclist crashes and 75% of pedestrian crashes were linked to police records • Authors hypothesized that underreporting for cyclists is due to "ambiguity of…laws and regulations" and the fact that cyclists are "less likely to cause property damage"
Sciortino, Vassar, Radetsky & Knudson, 2005	2000-2001	San Francisco	Pedestrian	Police Ascertainment Rate	• Police reports underestimate the number of pedestrian injuries by 21% (e.g., reporting level is 79%) • African-American pedestrians were less likely than white pedestrians to be linked to a police report • Women were more likely than men to be linked to a police report

Source	Study Period	Study Location	Focus	Definition Used	Findings/Conclusions
Short & Caulfield, 2016	2005-2011 (Police and Hospital Data) 2010-2011 (Injuries Board Data)	Ireland	General	Hospital Link Rate, Police Link Rate, Hospital Ascertainment Rate, Police Ascertainment Rate	• For pedestrian injuries, 28.9% of police records were matched with a hospital record; 44.3% of hospital records were matched with a police record • For bicyclist injuries, 24.8% of police records were matched with a hospital record; 8.2% of hospital records matched with a police record •Police Ascertainment Rate was 73.4% for pedestrians and 26.4% for bicyclists •Hospital Ascertainment Rate was 47.7% for pedestrians and 80.2% for bicyclists • False Positive Rate – 13% • False Negative Rate – 13%
Stutts & Hunter, 1999	1995-1996	Various locations in California, New York, and North Carolina	Pedestrian and Bicyclist	Hospital Link Rate	• 70% of bicyclist injuries reported by hospitals did not involve a motor vehicle • 64% of pedestrian injuries reported by hospitals did not involve a motor vehicle • 31% of bicyclist and 53% of pedestrian injuries occurred in non-roadway locations (e.g., sidewalks, parking lots, trails) • Police crash reports capture less than 33% of serious bicyclist injuries
Tarko & Azam, 2011	2003-2008	Indiana	Pedestrian	Police Link Rate	• Pedestrians struck on a state road, at a Y intersection, or on a divided roadway were less likely to be included in both police and hospital databases • Pedestrians struck while crossing a road, as opposed to walking along or standing outside of the roadway were much more likely to be included in both databases • Authors hypothesized that more severe injuries were more likely to appear in both databases
Tin Tin, Woodward & Ameratunga, 2013	2006-2011	New Zealand	Bicyclist	Capture-Recapture	• Police reports were linked to insurance, hospital, and mortality records • 13% of hospital reported crashes and 64% of hospital reported collisions were linked to police records • 39% of police reported crashes and 43% of police reported collisions were linked to hospital records • When compared with self-reported data from the cyclists, the entire linked dataset had a sensitivity of 63.1% and specificity of 93.5% • The collision-only dataset showed a 40.0% sensitivity and a 99.9% specificity
Watson, Watson & Vallmuur, 2015	2009	Queensland	General	Hospital Link Rate	• The study used discordance rates between police data and hospital records to measure underreporting of crashes to/by the police • The discordance rate was 44% for pedestrians and 93% for bicyclists (e.g., a reporting level of 56% and 7% respectively) • Authors hypothesized that bicyclist injures are not reported to the police because injuries are generally less serious, are less likely to have insurance implications, and are more likely to involve young people, who generally have high discordance rates

POTENTIAL IMPLICATIONS OF STUDY RESULTS

Findings from these ten studies indicate that there is a severe underreporting problem in datasets based only on police collision reports. Underreporting can result in the forecasting of incorrect estimates of crash and fatality rates, and identification of erroneous factors responsible for crash occurrence, thereby making the entire road safety exercise ineffective, according to Ahmed, Sadullah, & Yahya *(15)*. When road safety is evaluated based on data other than the actual number of crashes that occurred, there is a tendency to mistake trends in crash reporting with trends in traffic safety *(3)*. In addition, the authors found that inaccurate crash data can result in improper prioritization of funding and resources, and under- or overestimation of injury severity risk. For example, one study found that estimates based on underreported police-reported crash data minimize the effectiveness of seat belt use in injury severity risk and could have serious policy implications *(16)*.

In cases of crash underreporting, analysis relying on police data may be biased, according to Janstrup et al. *(4)*. Reliance on hospital data may also be problematic, as Watson et al. *(12)* reported that the level of underreporting varied depending on the data with which the police data was linked. Watson found that when hospital data was examined, approximately two thirds of the data were not linked to police data. Similarly, when Short and Caulfield *(5)* added injury claims data to their analysis, the total number of identifiable injuries was found to be more than three times greater than the number identified by police reports, and five times greater than the number identified in hospital records.

Because of the inaccuracy of injury classifications in police collision reports, injury severity cannot be used to match datasets. In addition, by only using injury data from police reports, financial costs to crash victims or to the public for health care associated with a crash cannot be easily and accurately determined *(14)*.

LIMITATIONS OF EXISTING STUDIES

Hauer and Hakkert *(3)* proposed methods to account for underreporting in police crash reports. They argue that the variance of the estimated number of crashes that occur can be calculated if the following factors are known: reported number of crashes, reporting level, and the variance of the reporting level. While the studies included in this literature review have attempted to establish the reporting level associated with various types of crashes, few have reported on the error surrounding their estimates.

In most real-world cases, true match status is unknown and link status is used as a proxy. Under the presence of a perfect matching process, link status and match status would be identical and link rate would be equivalent to the reporting level. However, there are two sources of error in the process of data linkage—false positives and false negatives. False positives are linked records that do not belong to the same person/event *(17)*. False positives are more likely to occur when identifiers are not discriminative and when files are large *(18)*. False negatives, also known as *missed matches*, are records that belong to the same person/event but that are not linked *(17)*. This occurs when records have inaccurate or missing data *(18)*.

Because linkage and analysis processes are often separated due to privacy concerns, researchers using linked datasets may be unable to determine the extent of bias that linkage errors have introduced into their study *(19)*.

It is difficult to predict the direction and magnitude of bias resulting from linkage errors due to the "distribution of errors with respect to variables of interest" which is usually unknown *(19)*. Missed matches reduce sample size and statistical power and can lead to under-underestimates of exposures or outcomes if the linkage is informative *(19)*. Because missed matches do not necessarily occur at random, the linked data may not accurately represent the study population, reducing the viability of the research effort and may introduce bias *(19)*. Selection bias can occur if an individual's presence in the linked dataset is related both to exposure and an outcome of interest *(19)*.

When data for key variables are missing, cases that are linked are less likely to be representative of the study population because they have fewer common values such as unusual zip codes *(17)*. Unfortunately, missing data is common in data linkage. The individual datasets used for linking—police, hospital, insurance, traffic—are generally not developed for research purposes. The intended use of a dataset largely determines its collection method *(20)*. Therefore, the contents and details of their attributes, and the application of various datasets for purposes other than those for which they were designed may result in decreased data quality. Additionally, many variables that would aid in correctly linking the datasets are often redacted for privacy purposes (i.e., name, address, social security number). Data may be inaccurate due to misspellings or lack of information (i.e., staff might guess the age of an unconscious patient and set the birth month and day to '01') *(21)*. Finally, the method of data collection can influence data quality because errors are more likely to occur when data is copied from handwritten forms or transcribed from conversations than when the information is entered directly into a database *(18)*.

While there are statistical methods for managing data that are completely or partially missing at random, it is likely that missing data in crash and hospital data linkages are usually missing *not* at random, which introduces systemic bias into the analysis of the linked dataset *(17)*. Studies of linked databases may be significantly impacted by linkage errors, especially if certain types of people or events are more or less likely to have the outcome of interest *(17)*. Often, the bias introduced by these errors cannot be determined because the errors themselves are unknown.

Linkage error can be estimated by using gold standard data, which is data with a known true match status. These data are rarely available in the real world, although sometimes synthetic datasets are used instead. To estimate linkage error, gold standard data is matched using the same process as the study datasets. This allows link status to be compared with match status and for the calculation of metrics such as sensitivity and specificity.

Sensitivity, or the true positive rate, is the proportion of matches that are correctly identified as links. Specificity, or the true negative rate, is the proportion of non-matches that are correctly identified as non-links. High sensitivity reduces the number of false negatives, while high specificity reduces the number of false positives. Often, there is a tradeoff between sensitivity and specificity.

Although sensitivity and specificity can be affected by the method of data linkage used, they are much more dependent on the quality of the data that is being linked. Data quality is affected by the accuracy, completeness, consistency, timeliness, accessibility, and believability of the variables used to link the databases *(21)*.

Of the ten studies included in the present review, only two report the sensitivity and specificity of the data linkage process. Both showed a higher specificity than sensitivity, which indicates that there are more false negatives than false positives in their datasets. One reported the false positive and false negative rates, which are similar though not interchangeable metrics.

Few of the studies provide an estimate of the error around their reporting level estimates, thus it is possible that bicycle and pedestrian crashes are not as underreported as these studies suggest. It may be that bicycle and pedestrian crashes appear in both police and hospital datasets but are less likely to be linked. Because of linkage error, link rate can only be used to *estimate* reporting level. Without knowledge of the *variance* of that estimate, the effect of underreporting on traffic safety analyses cannot be accurately determined.

SUGGESTIONS FOR IMPROVEMENT

Prior to linking data, researchers should carefully examine the individual datasets to ensure data integrity and completeness. Bohensky *(17)* recommends that direct, unique identifiers be included within datasets to reduce bias from incorrectly entered data or missing variables. Unfortunately, privacy concerns may prevent this method from being implemented in the United States. In the absence of direct identifiers, Bohensky *(17)* suggests the use of financial incentives to encourage data custodians to improve data quality and consistency. Definitions of variables used for linking, such as injury severity, should be standardized at the national or international level so that datasets from different sources can be reliably linked. Data linkage studies need to develop a clear and systematic method of reporting their methodology so that they can be easily repeatable by other researchers *(17)*.

Additionally, data linkers must develop indicators to describe the linkage errors present in a linked dataset *(17)*. Harron suggests three approaches to evaluating linkage quality *(19)*. The first approach is use of a gold standard dataset to directly measure missed and false matches—while this is the most accurate approach, it is difficult to apply this methodology in practice as gold standard data are rarely available. The second approach is to compare the characteristics of linked and unlinked data to determine potential sources of bias—this approach requires a linkage design in which all records in at least one file are expected to link, as well as provision of characteristics of the unlinked records to the researchers. The third approach is to conduct a sensitivity analysis to evaluate how sensitive the results are in response to changes in the linkage method—this approach requires that researchers have access to match weights for each linked record. Match weights do not reveal any sensitive information and thus are more likely to be shared with researchers. However, the sensitivity analysis may be difficult to interpret as the effects of missed matches and false matches cannot be distinguished from each other *(19)*.

Further research must be conducted to identify the populations most likely to suffer from selection bias during data linkage, in addition to determining the effects that different linkage processes have on reducing such bias *(17)*.

Additionally, researchers must use a consistent definition of reporting level so that results can be compared. While hospital and police ascertainment rates are the most theoretically accurate definitions of reporting level, researchers may consider use of hospital link rate instead because it is the most common definition. Finally, linkage with other datasets should be explored. For example, the addition of linked traffic data could help account for exposure in crash safety analysis.

CONCLUSIONS

Research has argued that police collision reports tend to underrepresent bicycle and pedestrian crashes, especially when motor vehicles are not involved. To account for this bias when making traffic safety decisions based on police data, the estimated reporting level and the uncertainty of the estimated reporting level must be known. Ten studies using data linkage to explore pedestrian and/or bicyclist safety were evaluated and summarized. There may be other relevant studies that could be reviewed in a more extensive literature review. Due to different definitions of reporting level, periods of study, and study locations, it was difficult to directly compare the studies. Even among the six studies using the hospital link definition, estimates of reporting levels ranged from 44 to 75 percent for pedestrian crashes, and from 7 to 46 percent for bicycle crashes.

These results indicate a severe underreporting problem in police collision reports, which could lead to inaccurate estimates of crash rates and could under- or overestimate the effects of road safety countermeasures.

However, few of the studies provided an estimate of the error around their reporting level estimates. Therefore, it is possible that bicycle and pedestrian crashes are not as underreported as these studies suggest. It may be that bicycle and pedestrian crashes appear in both police and hospital datasets but are less likely to be linked. Because of linkage error, link rate can only be used to *estimate* reporting level. Without knowledge of the *variance* of that estimate, the effect of underreporting on traffic safety analyses cannot be accurately determined.

Future studies must include estimates of the error present in their data linkage process so that the level of underreporting in police data can accurately be measured and accounted for. Additionally, datasets should be designed so that they can more easily be linked. This could involve standardizing the definition of common fields in each dataset, but could also involve the introduction of some type of individual identifier so that records can be linked automatically. Finally, linkage with other datasets should be explored.

The authors confirm contribution to the paper as follows: study conception and design: S. Doggett, D. Ragland; data collection: S. Doggett; analysis and interpretation of results: S. Doggett, D. Ragland; draft manuscript preparation: S. Doggett. G. Felschundneff. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Stutts, Jane C., and William W. Hunter. 1999. *Injuries to Pedestrians and Bicyclists: An Analysis Based on Hospital Emergency Department Data*. Final FHWA-RD-99-078. https://www.fhwa.dot.gov/publications/research/safety/pedbike/99078/index.cfm. Accessed March 18, 2018.
2. Elvik, Rune, and Anne Mysen. 1999. Incomplete Accident Reporting: Meta-Analysis of Studies Made in 13 Countries. *Transportation Research Record: Journal of the Transportation Research Board* 1665 (January): 133–40. https://doi.org/10.3141/1665-18.
3. Hauer, E, and A S Hakkert. 1988. Extent and Some Implications of Incomplete Accident Reporting. *Transportation Research Record* Methods for Evaluating Highway Improvements (1185): 1–10.
4. Janstrup, Kira H., Sigal Kaplan, Tove Hels, Jens Lauritsen, and Carlo G. Prato. 2016. Understanding Traffic Crash Under-Reporting: Linking Police and Medical Records to Individual and Crash Characteristics. *Traffic Injury Prevention* 17 (6): 580–84. https://doi.org/10.1080/15389588.2015.1128533.
5. Short, Jack, and Brian Caulfield. 2016. Record Linkage for Road Traffic Injuries in Ireland Using Police Hospital and Injury Claims Data. *Journal of Safety Research* 58 (Supplement C): 1–14. https://doi.org/10.1016/j.jsr.2016.05.002.
6. Tin Tin, Sandar, Alistair Woodward, and Shanthi Ameratunga. 2013. Completeness and Accuracy of Crash Outcome Data in a Cohort of Cyclists: A Validation Study. *BMC Public Health* 13 (1). https://doi.org/10.1186/1471-2458-13-420.
7. Sciortino, Stanley, Mary Vassar, Michael Radetsky, and M. Margaret Knudson. 2005. San Francisco Pedestrian Injury Surveillance: Mapping, under-Reporting, and Injury Severity in Police and Hospital Records. *Accident Analysis & Prevention* 37 (6): 1102–13. https://doi.org/10.1016/j.aap.2005.06.010.

8. Tarko, Andrew, and Md. Shafiul Azam. 2011. Pedestrian Injury Analysis with Consideration of the Selectivity Bias in Linked Police-Hospital Data. *Accident Analysis & Prevention* 43 (5): 1689–95. https://doi.org/10.1016/j.aap.2011.03.027.
9. Conderino, Sarah, Lawrence Fung, Slavenka Sedlar, and Jennifer M. Norton. 2017. Linkage of Traffic Crash and Hospitalization Records with Limited Identifiers for Enhanced Public Health Surveillance. *Accident Analysis & Prevention* 101 (April): 117–23. https://doi.org/10.1016/j.aap.2017.02.011.
10. Langley, J D. 2003. Missing Cyclists. *Injury Prevention* 9 (4): 376–79. https://doi.org/10.1136/ip.9.4.376.
11. Lujic, Sanja, Caroline Finch, Soufiane Boufous, Andrew Hayen, and William Dunsmuir. 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. https://doi.org/10.1111/j.1753-6405.2008.00162.x.
12. Watson, Angela, Barry Watson, and Kirsten Vallmuur. 2015. Estimating Under-Reporting of Road Crash Injuries to Police Using Multiple Linked Data Collections. *Accident Analysis & Prevention* 83 (October): 18–25. https://doi.org/10.1016/j.aap.2015.06.011.
13. California Department of Public Health. 2015. *Exploratory Analysis of Injury Classification of Crash Victims Using Crash-Medical Linked Data in California*. California Department of Public Health. https://archive.cdph.ca.gov/programs/Documents/Exploratory%20Analysis%20of%20Injury%20Classification%20of%20Crash%20Victims%20Using%20Crash-Medical%20Linked%20Data_Final.pdf. Accessed March 18, 2018.
14. Washington Area Bicyclist Association. 2015. *Modernizing the Collection, Integration and Disclosure of Crash Data: Policy Recommendations for the District of Columbia.* Washington Area Bicyclist Association. http://www.waba.org/wp-content/uploads/2016/01/DC-Crash-Data-Policy-Paper-July-2015.pdf. Accessed March 24, 2018.
15. Ahmed, Ashar, Ahmad Farhan Mohd Sadullah, and Ahmad Shukri Yahya. 2017. Errors in Accident Data, Its Types, Causes and Methods of Rectification-Analysis of the Literature. *Accident Analysis & Prevention*, July. https://doi.org/10.1016/j.aap.2017.07.018.
16. Abay, Kibrom A. 2015. Investigating the Nature and Impact of Reporting Bias in Road Crash Data. *Transportation Research Part A: Policy and Practice*, 71 (January): 31-45. https://doi.org/10.1016/j.tra.2014.11.002
17. Bohensky, Megan. 2016. Bias in Data Linkage Studies. In *Methodological Developments in Data Linkage*, First, 63–82. John Wiley and Sons.
18. Harron, Katie, Harvey Goldstein, and Chris Dibben. 2016. Introduction. In *Methodological Developments in Data Linkage*, First, 63–82. John Wiley and Sons.
19. Harron, Katie, James C Doidge, Hannah E Knight, Ruth E Gilbert, Harvey Goldstein, David A Cromwell, and Jan H van der Meulen. 2017. A Guide to Evaluating Linkage Quality for the Analysis of Linked Data. *International Journal of Epidemiology* 46 (5): 1699–1710. https://doi.org/10.1093/ije/dyx177.
20. Imprialou, Marianna, and Mohammed Quddus. 2017. Crash Data Quality for Road Safety Research: Current State and Future Directions. *Accident Analysis & Prevention*. https://doi.org/10.1016/j.aap.2017.02.022.
21. Christen, Peter. 2012. Data Pre-Processing. In *Data Matching*, by Peter Christen, 39–67. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31164-2_3.