5-2022

Data linkage for crash-injury outcome assessment.

Aryan Hosseinzadeh
University of Louisville

DATA LINKAGE FOR CRASH-INJURY OUTCOME ASSESSMENT

Aryan Hosseinzadeh

M.S., Transportation Planning, Tehran Polytechnique, Iran, 2016

B.S., Civil and Environmental Engineering, K. N. Toosi University of Technology, 2013

A Dissertation submitted to the Faculty of the J.B.Speed School of Engineering of the University of Louisville in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Civil Engineering

Department of Civil and Environmental Engineering

University of Louisville

Louisville, Kentucky

May 2022

DATA LINKAGE FOR CRASH-INJURY OUTCOME ASSESSMENT

Aryan Hosseinzadeh

M.S., Transportation Planning, Tehran Polytechnique, Iran, 2016

B.S., Civil and Environmental Engineering, K. N. Toosi University of Technology, 2013

A Dissertation Submitted on

April 28, 2022

to the Following Dissertation Committee

Dr. Robert Kluger, Committee Chair

Dr. Zhihui Sun

Dr. Richard Li

Dr. Monica Gentili

ACKNOWLEDGEMENT

First and foremost, I wish to express my wholehearted gratitude to Dr. Robert Kluger, my committee chair and advisor, who not only patiently guided me through every step of my Ph.D. journey but also stood by me at moments of frustration and disappointment. I feel so grateful that as I will pursue my career in academia, I will always have a perfect example of a mentor, a scientist, and a wonderful person, to look up to.

I would like to thank my committee members: Dr. Zhihui Sun, Dr. Richard Li and Dr. Monica Gentili, for examining my work and providing me with insightful feedback. Special thanks go to Dr. Reginald Souleyrette from University of Kentucky for his generous help throughout the study. It was my great honor to collaborate with and be thought by such a high caliber scholar. Finally, this research would not have been possible without financial support from the School of Inter-disciplinary and Graduate Studies and the Department of Civil Engineering at the University of Louisville.

Getting a Ph.D. is a long and challenging journey. I cannot begin to express my gratitude to my family for all of the love, support, encouragement they have sent my way along this journey. To my parents, Nahid and Ali, thank you for being my champions. Your unconditional love has meant the world to me. I hope that I have made you proud. To my siblings, Anahita and Reza, for all the support you have always given me through any endeavor I have undertaken. To my niece, Sarina, you are my inspiration to achieve greatness.

ABSTRACT

DATA LINKAGE FOR CRASH-INJURY OUTCOME ASSESSMENT

Aryan Hosseinzadeh

April 28, 2022

Introduction:

Traffic crash reports lack detailed information about emergency medical service (EMS) responses, the injuries, and the associated treatments, limiting the ability of safety analysts to account for that information. Integrating data from other sources can enable a better understanding of the characteristics of serious crashes and further explain variance in injury outcomes. In this thesis, first, a heuristic approach is proposed and implemented to link crash data to EMS run data, patient care reports, and trauma registry data. Next, the method was adapted through larger datasets in a statewide linkage effort. The performance of the heuristic method was compared with the Bayesian probabilistic linkage method. Further, EMS times, along with other crash-related explanatory variables, were used to investigate influential factors on injury severity. The level of consistency in injury severity estimation among medical experts based on trauma registry data was investigated and factors that contribute to misclassification of injury severity in crash reports were identified.

Methods:

A heuristic framework was developed to match EMS run reports to crashes through time, location, and other indicators present in both datasets. A comparative bias analysis was implemented on several key variables. Bayesian record linkage was also performed, and the results were compared with the heuristic one. A random-effects ordered probit approach was implemented by employing crash-EMS runs linked data to study the impact of crash-related effective factors along with EMS times on injury severity. Three models of (1) crash-related variables, (2) crash-related and EMS times, and (3) crash-related, EMS times and interaction effects of EMS times and injury location on the body were developed. Furthermore, the discrepancy between police-reported injury severities and physicians' evaluations of corresponding trauma records was modeled using crash-related linked data. The trauma data were reviewed and classified by a panel of emergency physicians. Analysis of Variance was applied to model variation within the panel. An ordered probit model was used to model factors contributing to misclassification between police reports and emergency physicians.

Results:

72.2% of EMS run reports matched to a crash record, and 69.3% of trauma registry records matched with a crash record. Females, individuals between 11 to 20 years old, and individuals involved in single-vehicle or head-on crashes were more likely to be present in linked data sets. The heuristic linkage method performs better compared to Bayesian linkage, and the reasons behind the linkage rate gap were discussed. In EMS times impact on injury severity analysis, although the outcome could not find the impact of faster EMS times on injury severity in the general model, but when the interaction effects were considered, faster EMS response time was associated with decreasing the severity of entire-body injuries. According to the discrepancy analysis results, age, internal injury, and a proposed field - injury visibility-

were found to be contributing factors to injury severity discrepancy. Internal injury and injury visibility were among the trauma-related factors that were developed to explore their impact on injury severity discrepancy. Results show inconsistent physicians' injury severity evaluation based on injuries' detailed information.

Conclusions:

Linking data from other sources can significantly enhance the information available to address road safety issues, data quality issues, and more. Linking data can result in biases that should be investigated as they relate to the use-case for the data. Based on the EMS times association with injury severity outcome, although a significant relationship between EMS times and injury severity in all types of injuries was not found, EMS times based on injured body locations shed light on the relationship between EMS times and injury severity. In discrepancy analysis, findings indicate officers tended to underestimate injuries associated with a high gore factor, increasing age and the presence of an internal injury, specifically among trauma patients.

Practical Applications:

Linked crash-related datasets provide a valuable opportunity to evaluate the impact of prehospital care and emergency department care on crash outcomes. In general, policy steps could be taken to require cross-reporting and linkage of the data sets as the events occur to better monitor outcomes of injury crashes without requiring post-hoc linkage. This method can also realistically be integrated into a tool or software to undergo record linkage automatically. The findings of this study could act as a base for further investigation of EMS impact on injury severity, particularly with respect to effective use of EMS times in the evaluation of service quality. Further research should also be devoted to developing field tests

that support officer injury assessment and identifying the factors leading to underestimating injuries identified in this study. Also, results suggest that injury visibility is important and should be investigated further for reporting purposes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction and contributions

Inaccurate crash injury severity identification in crash reports may result in missed injuries in the field, incorrect estimation of parameters in models, and low-impact roadway safety investments. Identifying the factors that lead to misclassification is crucial to improving the data quality. Traffic crash reports lack detailed information about emergency medical service (EMS) responses, the injuries, and the associated treatments, limiting the ability of safety analysts to account for that information. Integrating data from other sources can enable a better understanding of the characteristics of serious crashes and further explain variance in injury outcomes. EMS runs, and trauma registry data are not an inherent part of traffic crash reports. By linking crash-related databases, a vast opportunity comes up to expand the knowledge regarding the variables affecting the crash injury outcome, including post-crash variables, such as EMS times. Moreover, detailed descriptions available in trauma registry records can be used to cross-check and verify the credibility of police-reported injury data.

Contributions in this dissertation are in two main domains:

1. Introducing a method to link crash-related datasets to use for safety analysis and evaluating characteristics of the resulting dataset. Crash-related datasets, including police-reported crash data, Emergency Medical Services – Computer Aid Dispatch (EMS CAD) data, and Patient Care Report (PCR) and Trauma Registry were included in the linkage. Further, the transferability of linkage expanded into a larger dataset and a larger geographical context.

2. Applying methods to analyze the linked crash-related dataset in transportation safety. The linked dataset has unleashed new potential in safety analysis by adding new

variables to explain variance in safety research, including EMS runs data and trauma registry data. Investigating the association of EMS times and injury severity, exploring factors affecting EMS times, and identifying factors contributing to the misclassification of injury severity in police crash reports are among the linked data applications that were investigated.

1.2 Structure of this dissertation

This dissertation follows with Chapter 2 includes a review of the existing literature on data linkage and further analysis. Chapters 3 to 5 of this dissertation are five academic papers and a technical report. The chapters are therefore self-contained, that is, each of them has its own introduction, method, results, discussion and conclusion and notations.

Chapter 2 is an overview of the existing literature, focusing on three main topics: (a) crash-related data linkage, (b) Association of injury severity and EMS times, and (c) Injury severity misclassification in motor vehicle crashes

Chapter 3 includes two sections. First, an academic paper proposes a heuristic algorithm to link crash data, EMS runs and trauma registry records in Jefferson County, Kentucky. In the second section of chapter 3, the heuristic algorithm expanded and adapted for the Kentucky statewide crash-related dataset. Integrating data from other sources can enable a better understanding of the characteristics of serious crashes and further explain variance in injury outcomes. Furthermore, the selectivity biases were investigated and based on a manual review of the records, and the reasons behind linkage failure in records were categorized.

Chapter 4 compares the heuristic algorithm developed in this study with a Bayesian probabilistic record linkage. The records were categorized based on the ones matched in both methods; the ones matched as the outcome of only one of the methods and the ones that resulted differently based on each of the two approaches. Different types of matches were investigated, and the reasons behind each of the groups were discussed.

Chapter 5 explores the applications of the linked data in transportation safety. First, the association between EMS times, along with other crash-related explanatory variables and

injury severity were investigated. Next, EMS response time was modeled and compared using four machine-learning approaches, as well as regression analysis. Furthermore, factors contributing to the misclassification of injury severity in police crash reports were identified. The discrepancy between police-reported injury severities and physicians' evaluations of corresponding trauma records was modeled. The trauma data were reviewed and classified by a panel of emergency physicians.

Chapter 6 provides a summary, highlights the contributions, and offers future directions. In this chapter, crash-related linkage research framework was discussed in a bigger picture and outreach and potential transferability of approaches in other linkage frameworks were emphasized. Figure 1.1 shows the steps that were taken to conduct this research.

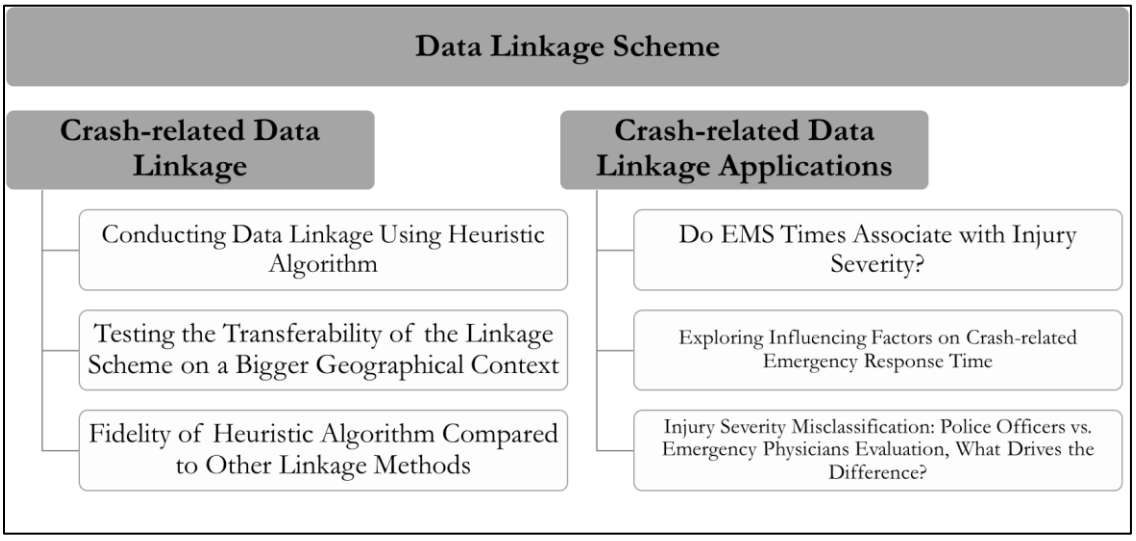


Figure 1.1. Research steps were taken in the dissertation

CHAPTER 2
BACKGROUND

2.1 Objective

In this section, an overview of previous studies related to data linkage was reviewed. The objective in the first part of the paper is to review crash-related data linkage, including the type of datasets, methods, linkage rate and geographical contexts. The objective of the second part is regarding a review of the literature about the applications of the linked data. The applications included three parts of "association of injury severity and EMS times," "determinants of EMS times," and "injury severity discrepancy."

2.2 Introduction

Crashes are one of the leading causes of preventable death in the United States, and they carry a severe burden on public health and wellness. Police-reported crash data is the primary source of information for transportation engineers to address safety systematically. However, additional data sets exist that can help explain factors associated with variance in crash outcomes and inform safety assessments. Emergency medical services (EMS) and hospitals both collect data about victims of traffic injuries. Both include specifics of the injury (Burch et al., 2014; Hosseinzadeh and Kluger, 2021a; Hosseinzadeh and Kluger, 2021b) through diagnoses and narratives. However, to date, they are rarely used to inform transportation engineering decision-making. Specifically, linking the crash records with trauma registry records and further analysis based on resulting data has been recognized as a gap in the literature (Tainter et al., 2020).

Several issues of interest can be investigated or addressed by linking crash data with other public health data sets. Some studies have expressed doubt about the accuracy of the crash reports, specifically, the injury severity field (Couto et al., 2016; McDonald et al., 2009; Watson et al., 2015). The KABCO scale is used by officers in the United States filing crash reports with the following designations: K – fatal; A – incapacitating injury; B – non-incapacitating injury; C – possible injury; and O – no injury. These definitions can vary slightly by state (for example, some list A as suspected serious injury). In recent years, crash reports have been subject to scrutiny regarding the quality of injury ratings. Crash injury severity is recorded using the officer's judgment based on limited information at the crash scene and can be incorrect when compared with medical professionals' assessments of a victim's condition (Benavente et al., 2006). Therefore, tracking crash injuries through emergency services and

hospitals can provide valuable knowledge about crashes and could be used to cross-check fields in the police reports for severe crashes. By linking crashes to a data set containing medical experts' diagnoses, the factors affecting injury severity discrepancy between police-reported crash data and trauma data could be investigated. Inaccuracies in severity reporting may also alter decision-making surrounding road safety issues (Das et al., 2021; Khoda Bakhshi & Ahmed, 2020). Furthermore, by relying solely on police-reported crashes, a portion of crashes might go unreported (Amoros et al., 2006; Boufous et al., 2008; Loo and Tsui, 2007; McDonald et al., 2009; Sciortino et al., 2005; Short and Caulfield, 2016; Tin et al., 2013a; Yannis et al., 2014), particularly for vulnerable road users such as bicyclists and pedestrians (Alsop and Langley, 2001; Amoros et al., 2006; Bakhshi and Ahmed, 2020; Loo and Tsui, 2007; Sciortino et al., 2005; Short and Caulfield, 2016; Tin et al., 2013b; Watson et al., 2015). While these issues are prevalent in crash data across the United States, police records are currently the most comprehensive source of information for monitoring road safety.

One of the potential factors that could have an impact on reducing severity is quick and efficient Emergency Medical Services (EMS). EMS characteristics are neglected in most studies, mainly due to data availability. Specifically, in the U.S., EMS characteristics are not an inherent part of datasets provided in the police-reported crash records that utilize database structure following Model Minimum Uniform Crash Criteria guidelines (NHTSA, 2017).

There are numerous challenges currently associated with linking crash data with EMS data and trauma records in the United States. Different agencies are responsible for collecting different information in EMS data and trauma records, open record data does not contain identifiable information such as name or driver license number, and privacy laws, such as the

Health Insurance Portability and Accountability Act (HIPAA), makes it so that health information and personal information are often inaccessible.

2.3 Crash-related data linkage

Record linkage is the process of linking data from different sources. There are three techniques used to link data: manual, deterministic, and probabilistic. Manual linkage is defined as "a process that requires human labor and involves visually comparing two (or more) data sets and determining whether each individual episode is the same across data sets" (Dean et al., 2001). Manual linkage is impractical with large volumes of records. Deterministic linkage "involves linking records based on an exact agreement of the selected match variables," such as personal identifiers (Karmel et al., 2010). The deterministic approach requires strong identifiers to be present in both data sets, which is often not the case, particularly in data sets that are open to the public, which have often been stripped of identifiers. Probabilistic linkage is defined as "linking records in two (or more) files and is based on the probabilities of agreement and disagreement between a range of match variables" (Karmel et al., 2010). Probabilistic linkage utilizes models to determine the likely matches.

A commonly used probabilistic approach is Bayesian record linkage (Conderino et al., 2017; McGlincy, 2004, 2006; Short and Caulfield, 2016; Watson et al., 2015; Winkler, 2002). Multiple existing software suites can guide users through the implementation of the Bayesian record linkage approach (Cook et al., 2015). Bayesian record linkage has also been used in the transportation safety context. A study in Dublin, Ireland, used Bayesian record linkage to link crash data with both hospital discharge records, and injury insurance claims based on age, gender, time, road user type, collision type, crash severity, and county. Their findings indicated a substantially lower match rate among bicyclist and motorcyclist injuries (Short and Caulfield,

2016). Conderino et al. (2017) used Bayesian record linkage, to link crash and in-patient hospitalization administrative records in New York City, NY, linking 52% of total hospital records to a crash by using date, time, gender, age, role, collision type, injury body location, and injury occurrence (Conderino et al., 2017). Milani et al. (2015) noted that the complexity of the Bayesian approach to probabilistic record linkage was one of the barriers to implementation in states across the U.S. (Milani et al., 2015).

In the United States, Crash Outcome Data Evaluation System (CODES) was a national effort led by the National Highway Traffic Safety Administration (NHTSA) to link hospital records with crash data (Cook et al., 2015). Each participating state was responsible for implementing linkage, and numerous studies utilized the linked data sets to investigate healthcare costs related to specific circumstances such as demographics (Shen and Neyens, 2015), aggressive driving (Chitturi et al., 2011), barrier and median-crossing crashes (Conner and Smith, 2014), seatbelt usage (Han et al., 2017), and motorcycle crashes (Olsen et al., 2014). CODES data sets have also been used to evaluate the quality of police reporting of injuries compared to injury severity ratings by medical professionals. Burdett et al. (2015) found significant differences between KABCO injury severity and Maximum Abbreviated Injury Scale (MAIS) in Wisconsin (Burdett et al., 2015). Burch et al. (2014) found consistency between distributions of injury reports in Maximum Abbreviated Injury Scale (MAIS) between Utah and Maryland crash data among injured persons involved in crashes, while KABCO injury severity varied (Burch et al., 2014). In the United States, the focus has been to link various hospital data sets with crash data, primarily through CODES (Cook et al., 2015), while only few studies were identified by the authors that linked EMS data with crash data.

Regarding studies across the world, a study in Portugal linked EMS, crash, and hospital data (Amorim et al., 2014) and used it to assess the quality of injury severity classification by the police using MAIS and length of hospital stay from the hospital data (Couto et al., 2016; Ferreira et al., 2017; Ferreira et al., 2015). The method was also used to assess the length of the prehospital impact on crash injury (Ferreira et al., 2019). A study in Queensland, Australia, linked patient admissions and crash data sets and found that motorcyclists, bicyclists, males, younger demographics, and injuries occurring in remote locations were more likely to go unlinked (Watson et al., 2015).

Errors and bias associated with data linkage is a relevant issue in data linkage exercises. Cryer et al. (2001) found significant differences in the distributions of variables including age, gender, and road user type between crash and hospital admissions data sets (Cryer et al., 2001). Justrap et al. (2014) found that certain attributes, including injury severity, speed, lane numbers, pedestrians, and females were more likely to result in a record being present in both trauma registry and crash data (Justrap et al., 2014). Tarko and Azam (2011) found selectivity bias in a linked crash and hospital data set to predict low injury levels among pedestrian-involved crashes. They found gender, age, crash type, and roadway geometry at the crash location were associated with the presence of a record in the linked data set (Tarko and Azam, 2011). Moore (1998) linked crash-hospital data in Alaska. Significant differences were not found between the age and gender of linked and unlinked records; however, significant differences were observed based on geographical location and crash type (Moore, 1998). Across the studies on selectivity bias in linking crash data to public health data, specific characteristics were consistent in most of them, including gender and injury severity or proxy for injury severity such as speed and crash type.

Table 2.1 summarizes the crash-related data linkage in previous studies, the method for data linkage used, the data sets, and their match rates. The match rate among studies in the literature varies from 29.8% to 74%. Most of the literature employed police-reported crash data and either EMS dispatch data or hospital data. Utilizing four crash-related data sets provides an opportunity to track and monitor crash injuries in each phase of the emergency. Due to the lack of personal identifiers and the complexity of the Bayesian approach raised in the literature (Milani et al., 2015), this study proposed an adaptive iterative heuristic approach to link crash data and public health-related datasets. Various sources of hospital-related datasets such as trauma registry, hospital admissions, hospital discharge, and in-patient hospital records were labeled as hospital data in Table 2.1

Table 2.1. Summary of crash-related data linkage implementations

Study	Objective	Method	Data sets	Linkage rate	Geographical context
(Moore, 1998)	Comparison of young and adult crashes	MINICODES software (Probabilistic method)	- Police-reported crash data - Hospital data	69% of MVC-related hospital data	Alaska, U.S.
(Stutts and Hunter, 1999)	Pedestrian and bicyclist crash analysis	Deterministic	- Police-reported crash data - Hospital data	California: 43%*, 45%** New York: 42%*, 56%** North Carolina: 66%*, 67%** *of Bicycle MVC-related hospital data **of Pedestrian MVC-related hospital data	California, U.S. New York, U.S. North Carolina, U.S.
(Cryer et al., 2001)	Investigating if hospital admission data linked to police MVC reports results in less biased information for the injury prevention policymaker and planner than police MVC reports alone.	Manual method	- Police-reported crash data - Hospital data	50% of MVC-related hospital admissions were found in the linked data set	England
(Alsop and Langley, 2001)	Exploring under-reporting of motor vehicle traffic crash	Automatch software package	- Police-reported crash data - Hospital data	63% of the total MVC-related hospital data	New Zealand
(Langley et al., 2003)	Exploring match rate of cyclist and the factors associated with the cyclist match rate	Automatch software package	- Police-reported crash data - Public road data	22% of cyclist crashes on public roads linked to a crash report	New Zealand
(Sciortino et al., 2005)	pedestrian injury surveillance	Matching thresholds	- Police-reported crash data - Hospital data	59% of the pedestrian MVC-related hospital data	California, U.S.
(Benavente et al., 2006)	Analysis of Injury Specifics and Crash Compatibility Issues	Probabilistic method	- Police-reported crash data - Hospital data	46% of MVC-related hospital admitted patients	Massachusetts, U.S.
(Boufous and Williamson, 2006)	Investigating factors affecting work-related traffic crashes	Probabilistic method	- Police-reported crash data - workers compensation data	46% of MVC-related work compensation claims	Australia
(Amoros et al., 2006)	Exploring under-reporting of road crash casualties	Semi-automated record-linkage procedure	- Police-reported crash data - Hospital data	37% of the total MVC-related hospital data	France
(Gonzalez et al., 2006)	Exploring factors affecting mortality in rural areas	Probabilistic algorithm	- Police-reported crash data	73% of the total MVC-related patient care reports	United States

			- Patient Care Reports - Hospital data		
(Lujic et al., 2008)	How comparable are road traffic crash cases in hospital admissions data and police records?	Linkage Wiz software	- Police-reported crash data - Hospital data	45% of the total MVC-related hospital data	Australia
(Tarko and Azam, 2011)	Investigating linked data selection bias in pedestrian crashes	Probabilistic method	- Police-reported crash data - Hospital data	51% of the MVC crashes matched with hospital records	Indiana, U.S.
(Wilson et al., 2012)	Validity of using linked hospital and police traffic crash records to analyse motorcycle injury crash characteristics	Automatch software	- Police-reported data - Hospital data	46% of the hospital data, 60% of serious injuries and 41% of moderate	New Zealand
(Kudryavtsev et al., 2013)	Evaluating reliability of police and healthcare data	Manual	- Police-reported crash data - Hospital data	162 matched fatality cases among 217 police records (74%) and 237 healthcare data (68.3%)	Russia
(Tin Tin et al., 2013a)	Completeness and accuracy of cyclist crash outcome Data	deterministic	- Police-reported crash data - Hospital data - Insurance data - Mortality record	13% of hospital reported crashes and 64% of hospital reported crashes were matched with police records, 39% of police reported crashes and 43% of police reported crashes were matched with hospital records	New Zealand
(Mitchell et al, 2015)	Comparison of novice and full-licensed driver common crash types	Choice maker software (Probabilistic method)	- Police-reported crash data - Hospital data	54% of MVC-related hospital admitted patients	Australia
(Watson et al., 2015)	Estimating under-reporting of road crash injuries	Linkage Wiz software (Combination of both deterministic and probabilistic approaches)	- Police-reported crash data - Hospital data - EMS data - Injury surveillance unit data	54% of MVC-related hospital admitted patients 29% of MVC-related EMS dispatch data 36% of MVC-related injury surveillance unit:	Australia
(Paixao et al., 2015)	Exploring motor vehicle crash death in high-risk population subgroup	Link Plus (Probabilistic approach	- Police-reported crash data - Mortality information system	1,072 resulted in initial match but manual review showed 311 of them are true matches	Brazil

(Short and Caulfield, 2016)	Linking police data with hospital and injury claims data	Probabilistic approach (Bayesian)	- Police-reported crash data - Hospital data - Injury claims -	61% of the total MVC-related hospital data	Ireland
(Janstrup et al., 2016)	Understanding traffic crash under-reporting	Deterministic approach	- Police-reported crash data - Hospital data	23% of the total MVC-related hospital data 34% of the MVC crashes matched with hospital records	Denmark
(Conderino et al., 2017)	Linking traffic crash and hospitalization	LinkSolv 9.0 (probabilistic approach)	- Police-reported crash data - Hospital data	52% of the total MVC-related hospital record	New York, U.S.
(Kamaluddin et al., 2018)	Matching of police and hospital road crash casualty records to investigate underreporting	Deterministic and probabilistic using Microsoft SQL	- Police-reported crash data - Hospital data	4% of MVC-related hospital records matched with police-reported crash data	Malaysia
(Tainter et al., 2020)	Data linkage approach to investigate potential reductions in motor vehicle crash severity	Iterative approach	- Police-reported crash data - EMS data	58% of the total MVC-related EMS data	Massachusetts, U.S.
(Ceklic et al., 2021)	Investigating MVC characteristics that are predictive of high acuity patients	Linkage Tool (v2. 1.5, Emory University, U.S.)	- Police-reported crash data - EMS data -	62% of MVC-related EMS record matched with police-reported crash data	Australia

2.4 Association of injury severity and EMS times

EMS play a vital role in linking individuals with trauma injuries to emergency care systems. EMS runs include phases, defined in Figure 2.1 by the National EMS Information System (NEMSIS). In terms of prehospital time, there is not a consensus among researchers about the impact on injury outcome (Harmsen et al., 2015, Lu and Davidson 2017, Ferreira et al., 2019, Katayama et al., 2019, Medrano et al., 2019). While some recent studies have focused on how reducing the prehospital time impacts fatality (Lee et al., 2018, Medrano et al., 2019, Nasser et al., 2020), other studies cast doubt about the universal effectiveness of reducing prehospital time (Newgard et al., 2010, Dharap et al., 2017, Möller et al., 2018). A shorter prehospital time can provide the injured individual with more advanced hospital care as quickly as possible. However, in some cases, on-scene care is shown to be more critical (Doggett et al., 2018a). Moreover, some severe injuries require transport to more advanced, potentially farther emergency departments.

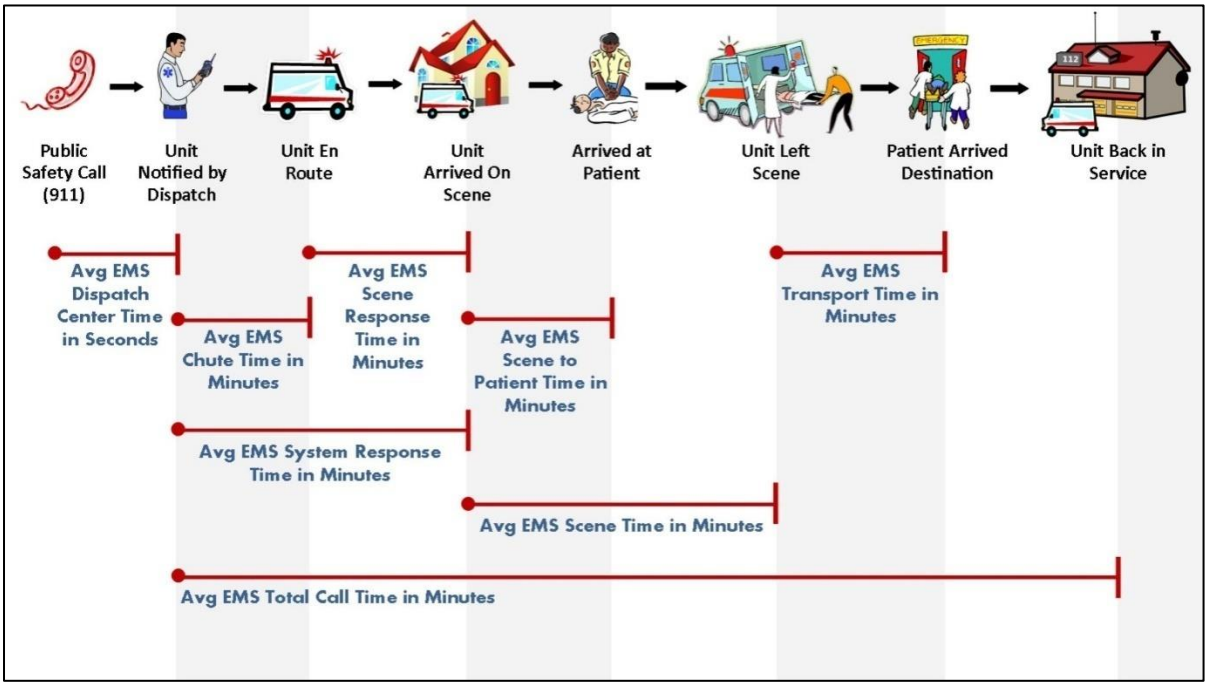


Figure 2.1. EMS times timeline (NEMSIS, 2013)

Some studies have explored the impact of response time on injury severity. Zeng et al. (2019) considered response time among explanatory variables and concluded that every minute increase in EMS response time increased the probability of medium and severe crash injuries by 0.36% and 0.11%, respectively (Zeng et al., 2019). Sanchez-Mangas et al. (2010) explored the leading factors for death in a crash and found EMS response time influential. According to their results, they estimated that traffic accident fatalities could be reduced by 30% by decreasing the average response time from 25 minutes to 15 minutes in Spain (Sánchez-Mangas et al., 2010). Lee et al. (2018) studied the impact of response time as well as two other prehospital times on injury severity. These intervals included crash-reporting time (the interval between occurrence and notification to EMS), response time, and transport time. Fatality Analysis Reporting System (FARS) data were utilized in this study which limited the EMS runs to the ones only including at least a fatal injury. Based on their model, they concluded increasing all three prehospital factors significantly increased the severity of the crash (Lee et al., 2018).

Feero et al. (1995) investigated the impact of out-of-hospital EMS time on survival and found shorter EMS time is significantly associated with unexpected survivors. This study was conducted among 848 injuries, of which 52% of them were related to motor vehicle crashes (Feero et al., 1995). Chen et al. (1995) found a higher preventable death rate among crashes in urban settings compared with crashes in rural areas with higher out-of-hospital time. Although the authors expected higher preventable death in rural areas due to longer EMS times, the outcome shows a 37.1% preventable death rate in rural settings compared to 48% in urban areas (Chen et al., 1995). Lovely et al. (2018) also did not find a significant relationship

between in-hospital mortality and either scene time or transport time. This study was conducted among about 4,000 injuries in Pennsylvania, United States (Lovely et al., 2018).

2.5 Crash-related determinants of emergency response time

Shorter prehospital time can quickly provide the injured individual with more advanced hospital care. However, in some cases, prehospital care administered by first responders such as paramedics or emergency medical technicians is shown to be more critical (Doggett et al., 2018a). Due to uncertainty around the impact of prehospital time on injury outcome, researchers focused on decreasing the response time, or the time specifically between an injury occurring and EMS arrival to the scene. The World Health Organization (WHO) has set reducing EMS response time as a goal, as it is expected to save lives (World Health Organization, 2009). A study in Costa Rica reported almost half of the fatalities are on the scene and could be partly due to insufficient EMS (Picado-Aguilar & Aguero-Valverde, 2020). Gonzalez et al. (2009) found a significantly higher EMS response time for fatal crashes when compared to non-fatal crashes in rural areas of Alabama (Gonzalez et al., 2009). Sanchez-Mangas et al. (2010) estimated that traffic crash fatalities could be reduced by 30% by decreasing the average response time from 25 minutes to 15 minutes in Spain (Sánchez-Mangas et al., 2010).

Since there are many ongoing efforts to reduce response time, some studies propose a practical threshold for response time and investigate the impacts. It has been generally recommended that response time be less than 8 minutes for at least 90% of calls (Stiell et al., 1999). In a county-level analysis in 2,268 counties across the US using National Emergency Medical Services Information System (NEMSIS) data from 2013 to 2015, longer response

times were significantly associated with higher rates of MVC fatalities. While the median county response time was 9 minutes, response times over 12 minutes have a 46% higher fatality rate ratio than those less than 7 minutes. In a study in Denver, Colorado, an eight-minute threshold did not result in a significantly lower fatality rate (Pons & Markovchick, 2002). A study in Calgary, Canada, also found the eight-minute threshold to be insignificant; however, there was a statistically significant decline in fatality for rising response time by one-minute increments (Couperthwaite, 2015). Ma et al. (2019), studied the response time in the United States and found two critical values: 5.5 minutes as the fastest decline in chance of survival and 17 minutes as the most critical cutoff for saving lives (Ma et al., 2019).

Due to the importance of response time in crash injury outcomes, some studies explored the factors affecting the EMS response time. He et al. (2019) used a spatial analysis approach to examine the impact of case-specific and service-specific variables on response time. Case-specific variables included caller's complaint (severe or minor), response mode (light/siren on or not), time of the day (day or night), time of the week (weekday or weekend), location (public or private area), visibility, wind speed and weather indicators. Service-specific variables included highway density, highway connectivity, speed, level of proficiency, and the number of ambulance and EMS demand. According to the results, response mode, mean visibility, EMS demand, highway connectivity, and level of proficiency were found significant (He et al., 2019). A five-year analysis in Michigan indicated that urban classification, day of the week, and month of the year were influential. In this regard, crashes in rural areas, on weekends, and during December, January and February showed higher EMS response time (Kumar et al., 2017). In a study in Malaysia, travel distance, age of patients, type of treatment, and peak hours were found as significant factors on EMS response time (Chin et al., 2017). A

study in Singapore explored the factors affecting short, intermediate, and long EMS response times and found weather, traffic, and location as significant impedances of swift response (Lam et al., 2015). Zhan et al. (2020) investigated the impact of call volume, precipitation, and temperature on response time. Based on their results, every additional EMS call, one °C increase in temperature, and one mm increase in daily precipitation could increase response time up to 8.79, 2.44, and 9.01 seconds, respectively (Zhan et al., 2020).

While a limited number of studies were investigated the determinants of EMS response time, other similar related definitions were used in safety literature, such as traffic incident duration (Cong et al., 2018; Hojati et al., 2013; Laman et al., 2018; Li et al., 2018), the incident response time (Hou et al., 2013) and clearance time (Ding et al., 2015; Tang et al., 2020). For example, Hojati et al. (2013) investigated the determinants of traffic incident duration in Southeast Queensland, Australia. They found variables such as distance from the central business district, being a major event, diversion/towing/medical requirement, and PM peak as significant factors (Hojati et al., 2013). Hou et al. (2013) developed a probability model to mathematically formulate incident response process based on incident response truck activities based on freeway incident data in Washington, United States. Debris, shoulder/median involved, total closure, injury involved, heavy trucks involved, work zone involved, average annual daily traffic, and weekends were identified associated with more prolonged incident response truck activities (Hou et al., 2013). Ding et al. (2015) used an endogenous switching model and found total closure, injury involved, work zone-involved and heavy truck-involved as influencing factors on clearance time (Ding et al., 2015).

2.6 Injury Severity Misclassification in Motor Vehicle Crashes

In recent years, crash reports have been subject to scrutiny regarding the quality of injury ratings. Crash injury severity is recorded using the officer's judgment, which is based on limited information at the crash scene and can be incorrect when compared with medical professionals' assessments of a victim's condition (Benavente et al., 2006). Brubacher et al. (2019) found that only half of the hospitalized crash injuries in their study were classified as a serious injury on the report (Brubacher et al., 2019). A study in Queensland, Australia, found a rate of discordance 45% to 70% between police-reported crash data and other trauma-related data sources (Watson et al., 2015). Overestimation, or reporting an injury that was more serious than the true injury, was observed in one-third of recorded data in different studies (Dove et al., 1986; Popkin et al., 1991). However, Dove et al. (1986) also reported underestimation in another third of their data (Dove et al., 1986). Morris et al. (2003) found both overestimation and underestimation cases after comparing police data with Abbreviated Injury Scale (AIS) in the UK. However, the number of overestimations was significantly higher (Morris et al., 2003). The state of practice for injury severity scoring in motor vehicle crash reports, used by police and transportation engineers, is to use the KABCO scale outlined in the Highway Safety Manual (AASHTO, 2010).

Few studies assessed the determinant of injury severity misclassification. Tsui et al. (2009) utilized a linked crash–hospital data set in Hong Kong and evaluated the agreement between police-reported injury severity and the Injury Severity Scale (ISS). The results show the police data greatly overestimated the injury severity. Age and position of the victims in the vehicle were significant in specifying the level of misclassification (Tsui et al., 2009). Further, a study in New Zealand revealed 15% of reported minor injuries in police data were, in fact, life-threatening.

Moreover, they found that females, single-vehicle crashes, and victims aged 65 and above were more likely to lead police officers to overestimate the severity of the crash injury (McDonald et al., 2009). Ferreira et al. (2015) found a significant tendency to overestimate the injury severity and identified victims above 65 years old, females, single-vehicle crashes and crashes in suburban areas were more susceptible to misclassified (Ferreira et al., 2015). A similar trend was found across eight other countries in Europe (Couto et al., 2016).

Taking advantage of CODES datasets, Burdett et al. (2015) evaluated the quality of police reporting of injuries compared to injury severity ratings by medical professionals in Wisconsin. They compared the injury severity of KABCO scale with the maximum abbreviated injury scale (MAIS) and found two-thirds of victims' injury severities were overestimated while only 2.9% were underestimated. They furthered their study by exploring the under/overestimation in nine body regions among victims. For instance, while only 7.2% of crashes included head injuries, 16.8% of underestimations were related to head injuries. Their outcome shows overestimation and underestimation were statistically significant in almost all body regions (Burdett *et al.,* 2015). Farmer (2003) conducted a comparison of police-reported data and the National Automotive Sampling System/Crashworthiness Data System (NASS/CDS), which included injury severity of medical records. The study outcome shows 49% of reported incapacitating injuries were not more than minor injuries. Overestimation was more frequent among females and young to middle-aged adults (Farmer, 2003).

CHAPTER 3

CRASH-RELATED DATA LINKAGE

3.1 Objectives

In this chapter, an approach was developed to link EMS computer-aided dispatch (CAD), EMS patient care reports (PCR) and a hospital trauma registry with police-reported crashes in Jefferson County, Kentucky, an urban county surrounding the city of Louisville, KY and expanding the results for the whole state of Kentucky. The seven main objectives in this research include: (1) proposing an adaptive stepwise algorithm to link four crash-related data sets, (2) defining types of matched and unmatched records, (3) comparing the match rate results with the previous data linkage frameworks in the literature, (4) identifying factors that affect records linkage and bias, (5) visualizing the results and drawing some inferences from the matched data, (6) tracking the injuries from crash to EMS and trauma registry and highlighting the potential discrepancies, and (7) exploring the transferability of the already developed method for other datasets. This study suggests an approach to linking transportation safety data sets for future analysis to investigate factors associated with variance in crash frequency and severity, evaluate EMS response times, and study health outcomes as they relate to crash circumstances

3.2 Data Linkage for Crash Outcome Assessment in Jefferson County, Kentucky: Linking Police-reported Crashes, Emergency Response Data, and Trauma Registry Records[1]

3.2.1 Data Description

This study uses four data sets in Jefferson County, KY. Crash records were collected by the Kentucky State Police through local law enforcement (Kentucky State Police, 2018), EMS CAD data was generated by dispatch software used by Louisville Metro Government's Department of Emergency Services, PCR data is reported by the EMS unit responding to an emergency, and trauma registry records are compiled by physicians at the University of Louisville Hospital (ULH). The study period was from July 2018 to March 2019. The data sets include all individuals involved in crashes in Kentucky over the study period. Figure 3.1 shows the visual framework of the data linkage used in this study.

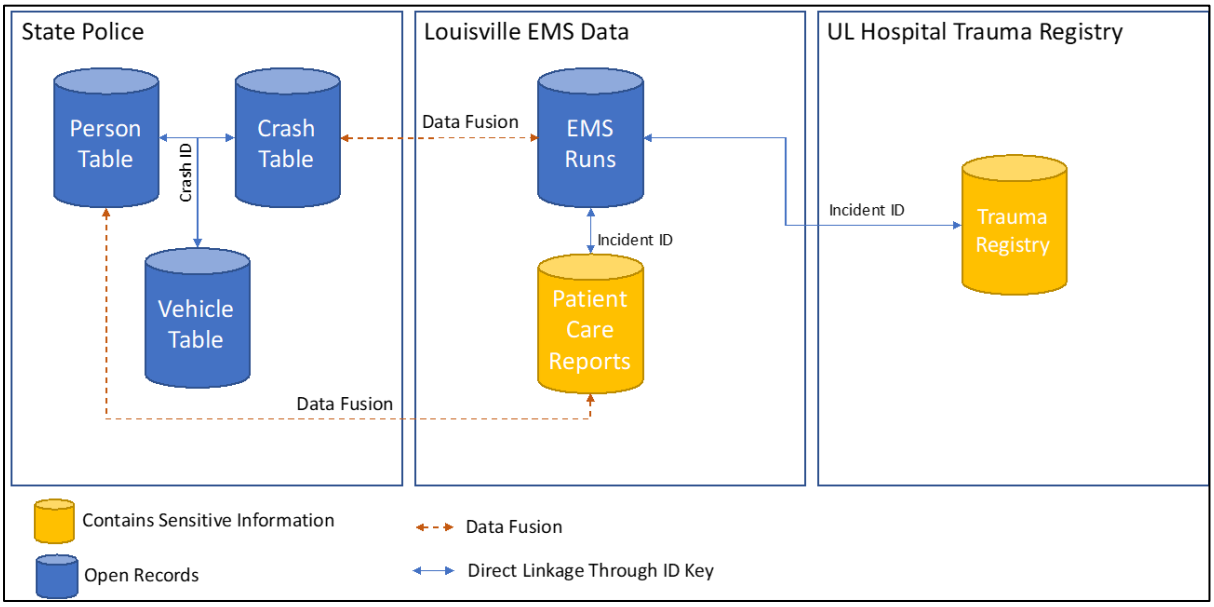


Figure 3.1. Visual framework of data linkage

[1] Sections from "Hosseinzadeh, A., Karimpour, A., Kluger, R., & Orthober, R. (2022). Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records. Journal of Safety Research *(in press)."* Included in this chapter.

Table 3.1 provides a summary of the data sets used, and the following sections describe them in further detail. Key fields used in the linkage methodology in EMS, PCR, and trauma registry were over 90% complete and over 70% complete in crash data. It should be noted that the incompleteness in crash data is mostly related to the events which were not severe and did not warrant immediate care; hence the detailed information regarding those events may not have been recorded. For more information regarding the fields available in the data sets and their completeness, please see the entity-relationship diagram in Figure 3.2.

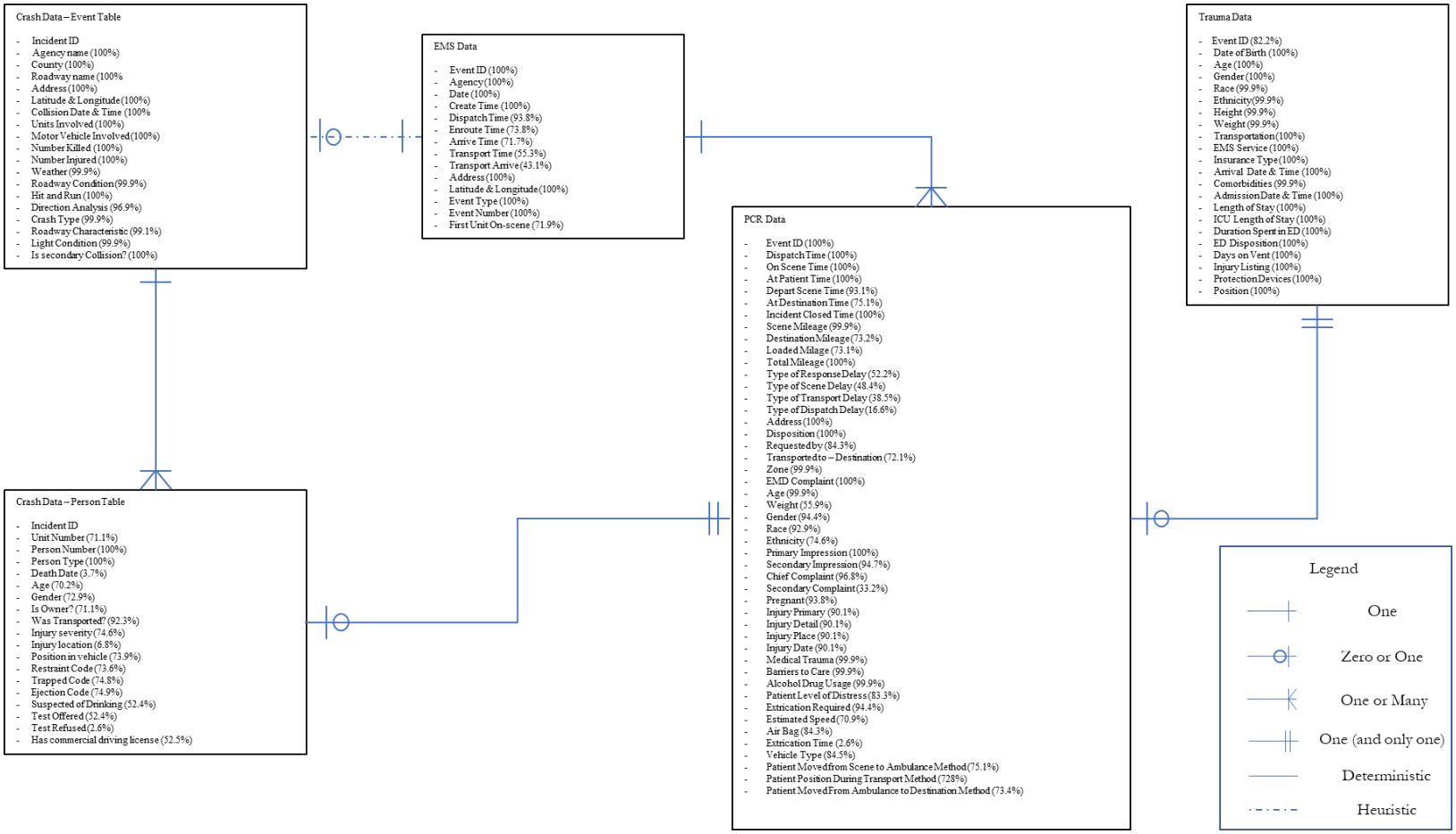


Figure 3.2. Entity Relationship diagram of study databases – Crash, EMS, PCR and Trauma - fields (completeness %)

Table 3.1. Summary of Jefferson County data sets used

Data Source	Number of records	Linkage variables	Data completeness
EMS	5,473 records	Lat/long	99.9%
		Event type	100%
		Date	100%
		Create time	100%
		Scene time	-
		Transport time	-
PCR	4,242 records	Incident date	100%
		Dispatch time	100%
		Age	99.9%
		Gender	94.5%
Crash (Event Table)	21,358 records	Lat/long	100%
		Crash time	100%
		Number of injuries	100%
		Number of killed	100%
		Street/intersection	95.3%
Crash (Person Table)	80,786 records	Age	70.2%
		Gender	72.9%
		Was transported?	92.3%
Trauma Registry	163 records	Arrival Date/time	100%
		Age	100%
		Gender	100%

The 9-month period contained 21,358 crash records with the database structure following Model Minimum Uniform Crash Criteria guidelines, including a crash table, person table, and vehicle table with records from each table linked by a key field (NHTSA, 2017).

CAD data is collected by software used by emergency dispatchers who field 911 calls and direct first responders to the scene. CAD systems record information from emergency services, including police, fire, and EMS. A total of 5,473 EMS run reports were recorded for Motor Vehicle Crashes (MVC) during the study period. The data includes run time features (create, dispatch, en-route, arrive, transport, and transport arrived), approximate location (block-level address), event type, and run priority. The locations were made available through an open record request but were reduced to the block level to avoid HIPAA violations.

Patient Care Report data is gathered by the first responders as they respond to incidents. A total of 4,242 PCR records across 2,883 EMS run reports, labeled as a MVC during the study period, were used. Unlike EMS CAD data, which only contains run time features, the PCR provides information regarding the patient's condition. The data includes run time features, approximate address, patient dispositions, injury impression, patient complaint at the scene, and patient socio-demographic information. PCR events contain a run ID, which enables a direct linkage with EMS CAD data. PCR data collected follows national standards for EMS care reporting outlined in the National EMS Information System (NEMSIS) (Legler et al., 2017).

All level 1 Trauma Centers in the United States are required to maintain a registry of trauma cases for performance evaluation. Trauma registry data was obtained from the Emergency Department (ED) at ULH over the study period. The data set contains 194 records where a patient was admitted to the ED with an MVC injury. 163 of those patients arrived at the ED via EMS and included a corresponding valid EMS run ID linked back to the CAD system at Louisville Metro's Department of Emergency Services. The variables included are patient characteristics, including age, gender, race, ethnicity, height and weight, arrival and admission date/time, and injury severity indicators, including injury severity score (ISS), length of stay in the hospital, and length of stay in the intensive care unit. ISS ranges from 1 - 75 and is based on the worst injuries in six different parts of the body: head and neck, face, chest, abdomen, extremity and external (Baker et al., 1974, Greenspan et al., 1985).

ULH is the only Level 1 Trauma Center in Jefferson County. The majority of the most severe injuries involving motor vehicles should end up at ULH, however, lower severity injuries may be taken to other hospitals in the region depending on proximity to the hospital

and patient preference. Additionally, patients will arrive from crashes in nearby counties in many cases.

3.2.2 Methodology

In this section, the heuristic framework to link data is proposed. In the first step, various thresholds of time and distance differences in crash and EMS runs were tested. Further, the initial matches based on different thresholds went through an adaptive iterative framework to reduce the number of duplicates and find the unique associated records. In the next step, random manual checks were conducted to investigate the fidelity of the proposed algorithm and to clarify some suspicious cases. In the final step, some variables of interest were visualized to investigate the linkage bias and examine linkage credibility.

Linkage framework

In this section, the possible match outcomes are defined and then the methodology to arrive at those outcomes has been described. The match rate of crash-related databases was defined in Equation 3.1. This match rate includes true matches and, presumably some false matches.

$$Match\ rate = \frac{True\ matches + False\ matches}{All\ records\ in\ a\ dataset} \qquad (3.1)$$

Most literature reviewed simply provided a match rate in their results (Alsop and Langley, 2001; Conderino et al., 2017; Short and Caulfield, 2016). However, not all unmatched records are the same. In this study, possible match outcomes were defined to improve understanding. The proposed matching approach has a finite set of pre-defined outcomes encompassing all possible cases, presented in Table 3.2.

Table 3.2. Classification of match types

Match Outcome	Description
1 Crash – 0 EMS run report	In this case, the crash was not linked with an EMS run report. The crash most likely did not require EMS to be sent to the scene.
1 Crash – 1 EMS run report	In this case, only one crash feasibly matched the EMS run report after the approach was implemented.
1 Crash – 2+ EMS run reports	In this case, two or more independent EMS run reports were sent to locations near a crash the method was unable to establish which EMS run was intended for the crash.
0 Crash – 1 EMS run report	In this case, the EMS run report was not successfully linked to a crash, despite being tagged as an MVC in the CAD data. These cases are unreported crashes, erroneously labeled in CAD, or maybe entries for complicated scenarios.
2+ Crashes – 1 EMS run report	In this case, two or more crashes occurred near each other at a similar time, and the approach was unable to distinguish for which crash the EMS run report was called.

To meet objective (1) of proposing an adaptive stepwise algorithm to link the study data sets, a heuristic algorithm was developed. Figure 3.3 summarizes the record linkage process in this study. The proposed approach implements a series of checks and filters to match records in a stepwise manner. For this section, the subscript *C* corresponds to a field from the crash data, *E* corresponds to the EMS CAD data

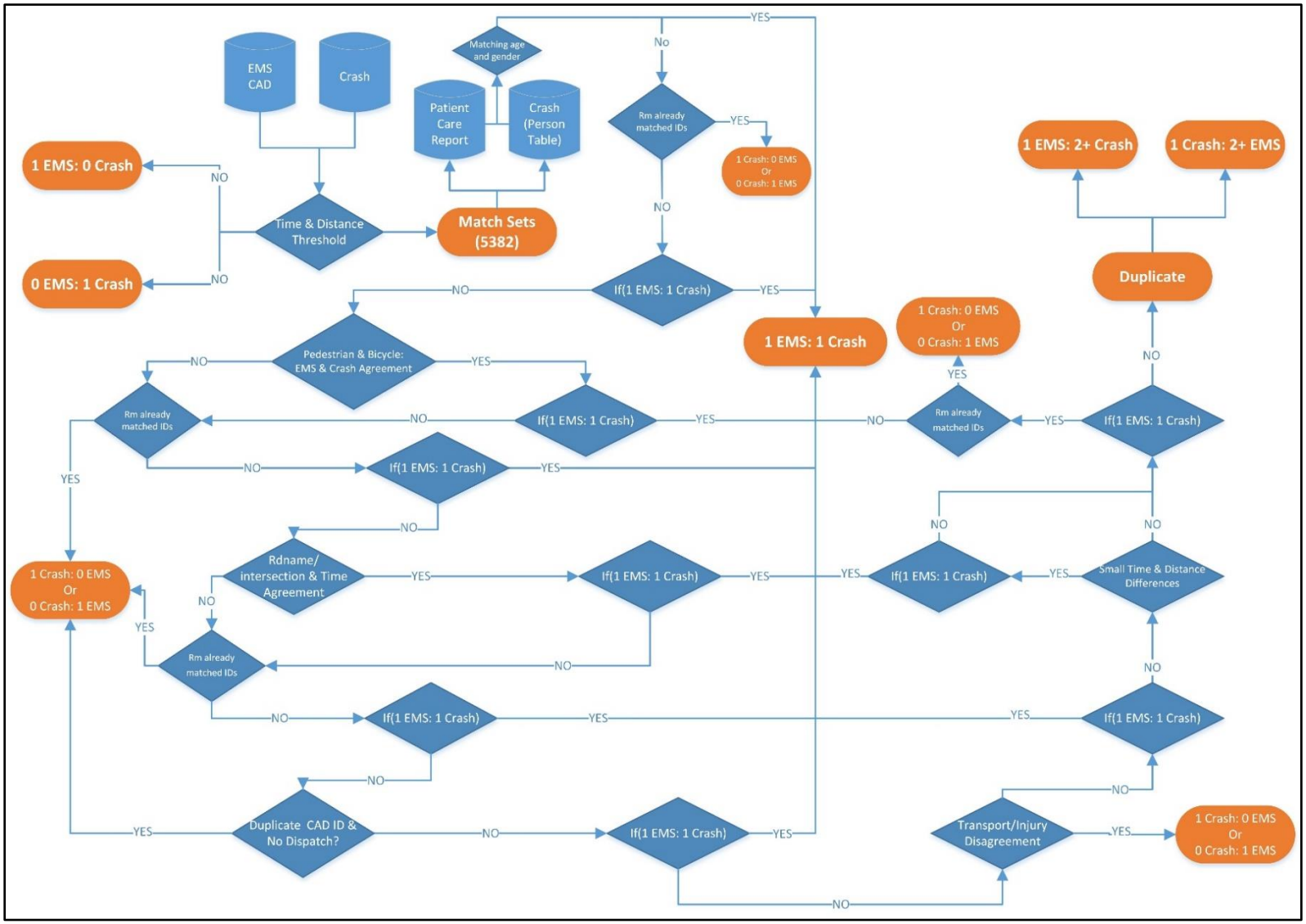


Figure 3.3. Heuristic algorithm to link crash data and EMS data

The initial check utilizes a time-distance boundary to determine a pool of possible matches between EMS run reports and crash events. Date-time was extracted from the EMS CAD data as the time the 911 call was received, while the time of crash filed in the police report was used for the crash data. Equation 3.2 defines two distance thresholds *D*, as a function of time, D_t, and Euclidian distance, D_d (using position, *x,* and *y*) and allocates the EMS run report (*j*) to a crash event set (*i*).

$$\begin{cases} |t_{Ci} - t_{Ej}| \le D_t \; and \; \sqrt{(x_{Ci} - x_{Ej})^2 + (y_{Ci} - y_{Ej})^2} \le D_d & j \in i \\ |t_{Ci} - t_{Ej}| > D_t \; or \; \sqrt{(x_{Ci} - x_{Ej})^2 + (y_{Ci} - y_{Ej})^2} > D_d & j \notin i \end{cases} \quad (3.2)$$

The result of implementing Eq. 3.2 is *i* crash "event sets" (1 for each crash), with each set containing elements of EMS run IDs potentially matched to the crash. The majority of simple cases will be matched as 1 Crash-1 EMS matches or 1 Crash-0 EMS matches, using Equation 3.2 assuming the distance threshold is set reasonably, as the probability of multiple crashes occurring at the same time within a short distance of each other and both requiring EMS is low. After implementing the time/distance threshold in equation 3.2, a series of additional checks are conducted for cases where any the following was true:

1. EMS run report *j* was not matched to a crash event set *i*.
2. EMS run report *j* was uniquely matched to a crash event set *i*.
3. Multiple EMS run reports {$j_1, j_2, \ldots$} were matched to crash event set *i*.
4. Crash event set *i* was not matched to an EMS run report *j*.
5. EMS run report *j* was assigned to multiple crash event sets. {$i_1, i_2, \ldots$}.

a. Multiple EMS run reports $\{j_1, j_2, \ldots\}$ were each matched to multiple crash event sets $\{i_1, i_2, \ldots\}$.

In case 1, an EMS run report was not associated with any crashes resulting in a 0 Crash-1 EMS Match. In case 2, a unique EMS run reports matched with a unique crash. In case 3, two or more EMS run reports were made to locations near a crash. In case 4, a crash was not successfully linked to an EMS run report, despite being tagged as an MVC in the CAD data. In case 5, two or more crashes occurred within the proximity of an EMS run report. A special scenario of case 5 involved two or more EMS run reports to the proximity of two or more crashes, so two crash event sets were assigned had the same possible EMS matches.

After the time distance threshold was implemented, each event set underwent several checks to find inconsistencies. For example, pedestrian/bicycle crash was a strong indicator of a likely match for EMS run reports that were labeled and matched to a crash with a pedestrian or bicyclist as the likelihood of two pedestrian/bicyclist involved crashes in the same short time period is even lower than that of two crashes occurring within a short timeframe. Validation of matches was conducted through a set of manually implemented diagnostic procedures. The first manual check was conducted by randomly sampling matches and examining all the fields. The locations were plotted on a map, and checks were made regarding the details of both the EMS run report and the crash event. Particular attention was paid to cases other than 1 Crash-1 EMS matches to determine why the algorithm was unable to identify a match. Figure 3.4 shows an example of a manual check conducted. Crash A and the EMS run report share the same road name (*Road/Intersection Name check)*. Furthermore, there are four injuries in crash A and the EMS run report transported a crash victim to the ED (*Transport Field/Injury Field check)*. Louisville Metro Government's Department of

Emergency Services conducted additional validation by auditing a set of specifically identified cases that were unable to be explained within the existing data and manual validation process. The audit involved utilizing the CAD ID to read narratives written by police, fire, and EMS who responded to the event. Some interesting lessons from the audit are shared in the results section.

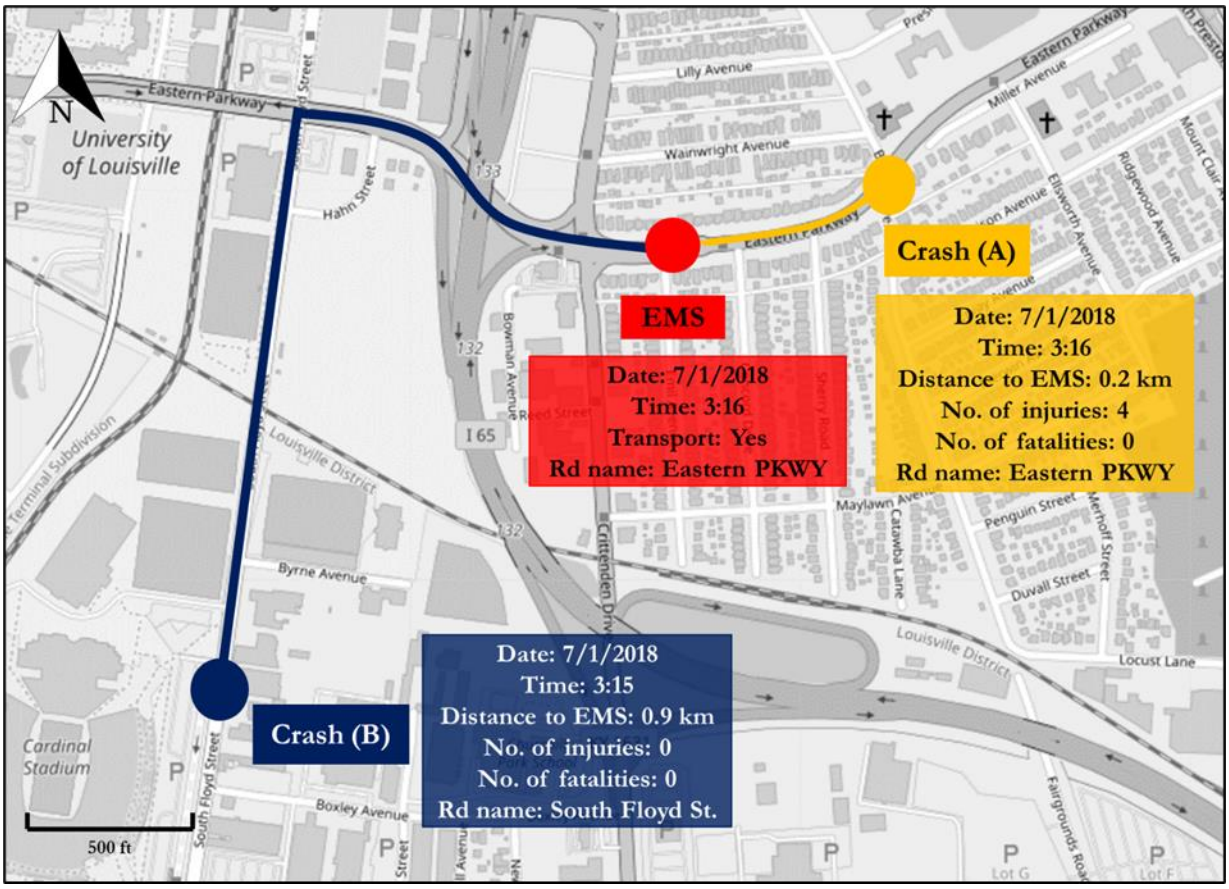


Figure 3.4. Visual Representation of Initial match candidates - Crashes and EMS locations

Once the EMS run reports were matched to crashes, the trauma registry data could be matched to the specific occupants of the vehicles from the person table in the crash data. The EMS run reports were already linked to the Trauma Registry via the EMS run ID, but the individual person in the crash person table still can be matched to a trauma record. After identifying the

crash-EMS match from the heuristic algorithm described, the demographic information between crash occupants in the person table and trauma registry records linked to the EMS run reports were compared. Age and gender were the primary fields used to match the specific occupants to the trauma record. Most cases of matched records were not present in the trauma registry as it requires the patient to be transported and admitted to the emergency department at ULH.

Linkage credibility and bias investigation

Using the linked data to build a model is going to result in some standard error associated with both modeling process and linkage process. Two possible issues can arise from linkage process: random error and bias. Random error is an error associated with incorrect linkage due to random chance, while bias is induced because certain data types may be more or less likely to be linked for systemic reasons within either the linkage process or the data itself.

Without ground truth data, simulation can be used to investigate the impact of random error on the modeling process. By simulating specifically erroneous data points, one can investigate how sensitive the results are to errors rates in linkage. If the results are highly sensitive to simulated errors in data, the model results may be unreliable if the user has doubt about the fidelity of the linkage process. This approach could be implemented on any modeling or data analytics exercise using the linked data.

For bias, quantitative bias analysis can be used to examine biases that may be in the data set. Quantitative bias analysis is a general approach to understand the extent that these errors produce bias on the results (Doige and Harron, 2019; Harron et al., 2020; Harron et al., 2014; Janstrup et al., 2016; Tarko and Azam, 2011). The first step is comparing linkage rates between different study variables to assess how much and in which direction the results might

have been influenced by bias. Visualizing variables of interest can help to gain deeper understanding regarding the bias imposed by data linkage. Also, the results should be in line with the expectation, for example, in our study, there were lower matches among individuals with no injury. This can be explained because emergency services are called less frequently for low-severity events. When looking to use a linked data set for modeling, it is important to investigate variables of interest for bias and better understand the population contained in the data set.

3.2.3 Results

R Programming Language (R Core Team., 2019) was used for data wrangling and matching. As stated in section 3 (Data Description), matching was conducted in the first stage using Eq. 2. Different thresholds for distance (D_d) and time (D_t) were tested. By implementing the distance and time thresholds, the (1 km, 60 min) and (1 km, 120 min) thresholds resulted in similar numbers of 1 Crash-1 EMS matches. However, the lower time threshold removed many more of the 1 Crash-2+ EMS and 2+ Crash-1 EMS sets. The threshold has an inherent trade-off between false positive and true negative pairs. As the threshold becomes stricter, the probability of matching events incorrectly lowers but it can also eliminate the true matches. How the remaining steps perform in reducing the size of 1 Crash-2+ EMS and 2+ Crash-1 EMS cases will dictate how restrictive the threshold should be. After applying the remaining steps proposed in the algorithm, Table 3.3 presents different match type numbers based on D_d , and D_t . The 1 km and 60 minutes threshold were chosen and used for further analysis.

	$D_d \leq$ 1 km $D_t \leq$ 60 min	$D_d \leq$ 0.5 km $D_t \leq$ 60 min	$D_d \leq$ 1 km $D_t \leq$ 30 min	$D_d \leq$ 0.5 km $D_t \leq$ 30 min	$D_d \leq$ 1 km $D_t \leq$ 120 min

1 Crash – 0 EMS	**17144**	17466	17432	17710	16878
0 Crash – 1 EMS	**1351**	1607	1576	1831	1194
1 Crash – 1 EMS	**3955**	3780	3787	3587	4019
1 Crash – 2+ EMS	**107**	44	68	34	176
2+ Crash – 1 EMS	**258**	93	118	44	498

Table 3.3. Record matching after the algorithm process

The full narrative was investigated manually by Louisville Metro Government's Department of Emergency Services for specific cases to determine the possible reasons behind the match type. The cases with an EMS run, but no crash, were of particular interest since these cases may be able to provide an indicator of underreporting of injury crashes. Some of the interesting cases discovered include 1) an officer was on the scene but did not file the report for an unknown reason 2) an arrest was made at the scene of the crash after a police chase and only an arrest report was filed. The first case is a clear case of underreporting, while the second would be an example of a more complicated scenario.

Of the 163 trauma records with arrival to the ED via EMS, 113 of them were matched to a crash and validated through the person table records of that crash. The unmatched records likely include a mix of invalid EMS Run IDs in the trauma data set, unreported crashes, unsuccessful matches, and most likely, crashes that happened outside of Jefferson County

where the patient was taken to the University of Louisville hospital due to the severity of the injury and the quality of the hospital. 13 Trauma records where a motor vehicle crash was the cause of injury indicated an arrival by personal vehicle. These indicate either unreported injury crashes or crashes that were reported where the person declined EMS and chose to go to the ED later. These 13 records could not be matched to a specific crash through this methodology.

As the outcome of linkage process, three linked data sets were generated: 1) crash-EMS CAD with 3955 records, 2) Crash-EMS CAD-PCR with 3002 records, and 3) Crash-EMS CAD-PCR-trauma with 113 records. To meet objective (3), the results of this study were compared with the similar studies and the heuristic algorithm outperformed in terms of match rate. This study was able to link 72.2% (3,955/5,473) of the MVC-related EMS data to a crash and 69.3% (113/163) of the MVC-related trauma registry data to a crash, respectively. 18.5% (3,955/21,358) of crashes were linked to an EMS record. The complete structure of the linked database is presented in the entity-relationship diagram in figure 3.2 . The current framework's most important advantage is utilizing an adaptive approach iteratively evaluates the pair's status at each stage.

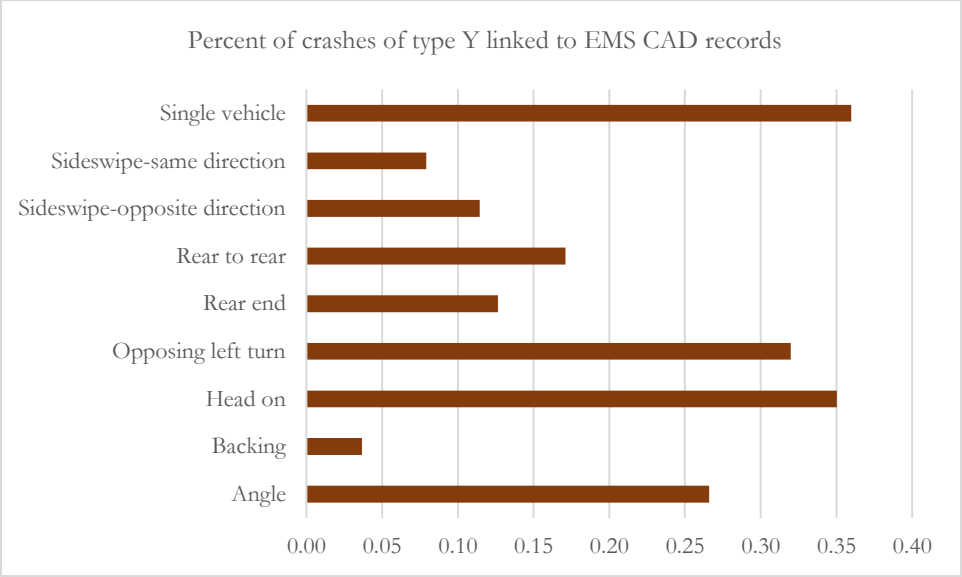
3.2.4 Discussion

Linked data bias

Before diving into the applications of the linked data, this section investigates the biases associated with the linked data set to fulfil objective (4). Figure 3.5 shows the percentage of crashes in the crash-EMS CAD linked data set broken down by (a) crash type, (b) number of injuries, and (c) injury severity. Distribution of variables' break down based on their categories in linked data could be different from either crash data or PCR data. Three factors could be the leading causes: first, some characteristics inherently have a higher chance of getting

reported in the linked data. For example, more severe crashes have a higher chance of requiring EMS and subsequently ending up in hospital. Second, some could be due to false matches. However, by applying rigorous stepwise adaptive algorithm and random manual checks, we are confident that this error is negligible. Third, under-reporting crashes or EMS runs. For rest of the section, some variables of interest were visualized to diagnose the biases in the linked data set.

Figure 3.5(b) demonstrates that crashes with injuries are more likely to be present in the data set. Also, according to Figure 3.5(a) it can inferred that the distribution of crash types in the linked data is different than distribution of crash types in crash data. This could be true for any other variable of interest. While it is not surprising, it is important to consider when developing research questions and applications that use this data set.



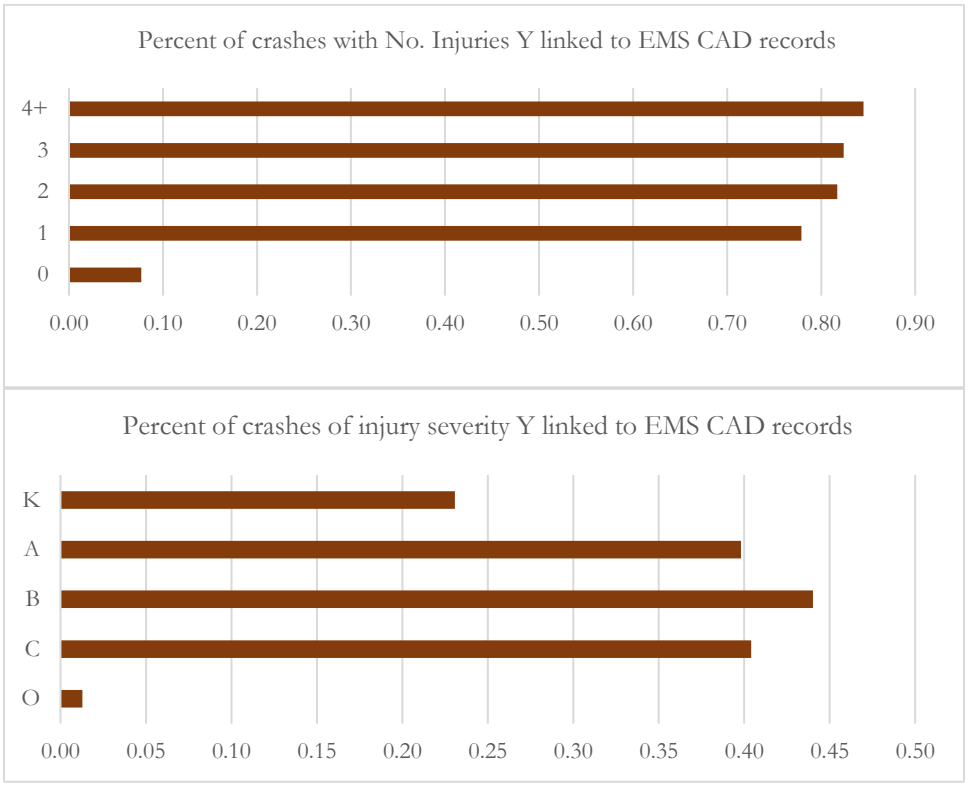


Figure 3.5. Percent of crashes in the crash-EMS CAD linked data set broken down by (a) crash type, (b) number of injuries, and (c) injury severity.

The age distribution was also investigated for bias. Figure 3.6(a) depicts the number of records in each age group of the crash-EMS CAD-PCR linked records divided by the number of records in the same age group in PCR. Since records in PCR were already transported to the hospital, significant differences among specific age groups would represent underreporting bias. A discernable difference cannot be captured in Figure 3.6(a). Figure 3.6(b) shows the number of each age group records in the crash-EMS CAD linked records divided by the number of records in the same age group in police-reported crash data. However, based on Figure 3.6(b), it can be speculated that younger individuals 11 to 20 years old and elderly individuals are more likely to be present in the crash-EMS CAD linked data because those individuals are either more likely to go to a hospital for care or are more frequently involved in serious crashes.

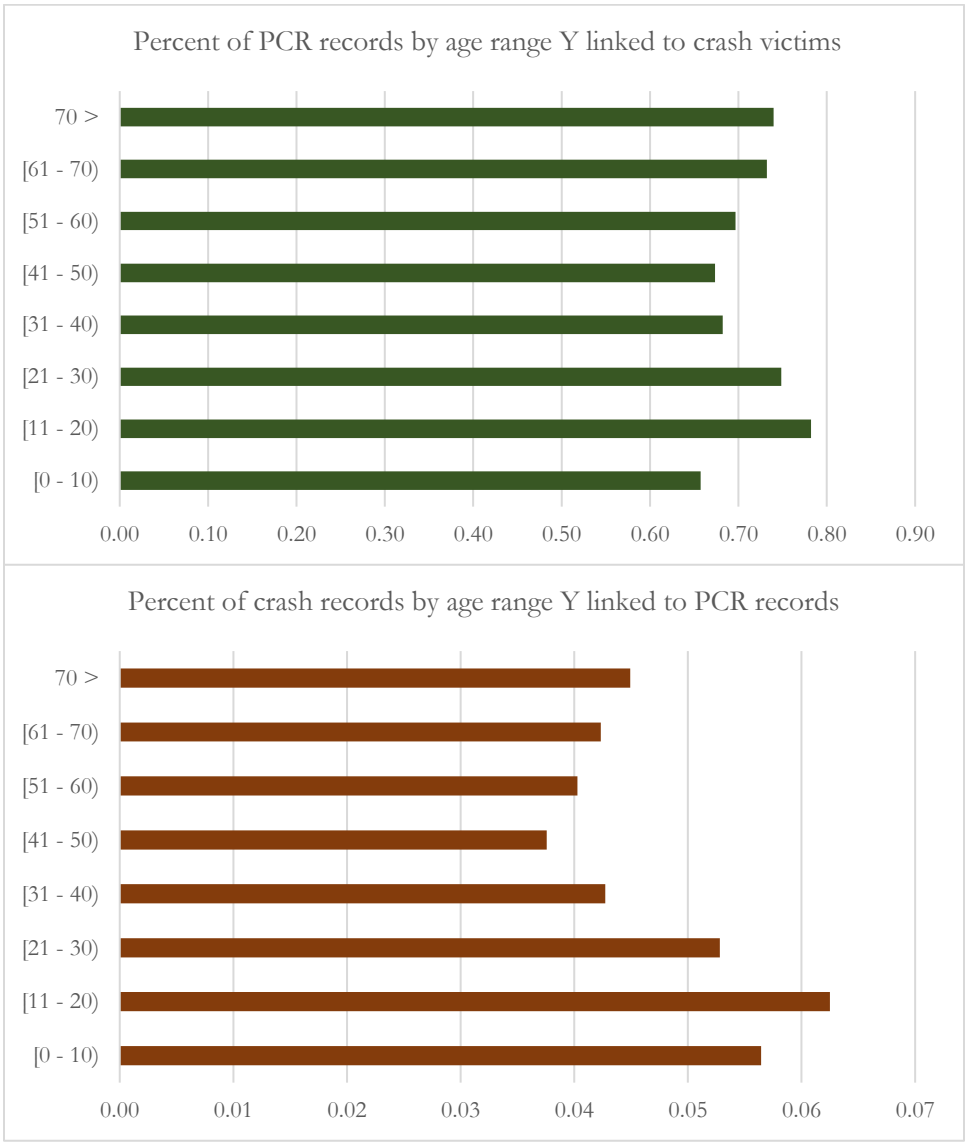


Figure 3.6. (a) Percent of PCR records by age range Y linked to crash victims (b) Percent of crash records by age range Y linked to PCR records

Investigation of gender shows that females are more likely to be captured in the linked data (Figure 3.7), despite the number of males in police-reported crash data (n=35,397) being higher than females (n=31,528). The numbers of females in PCR (2,099 female vs. 1,907 male) and crash-EMS CAD-PCR linked data (1,624 female vs. 1,377 male) were higher.

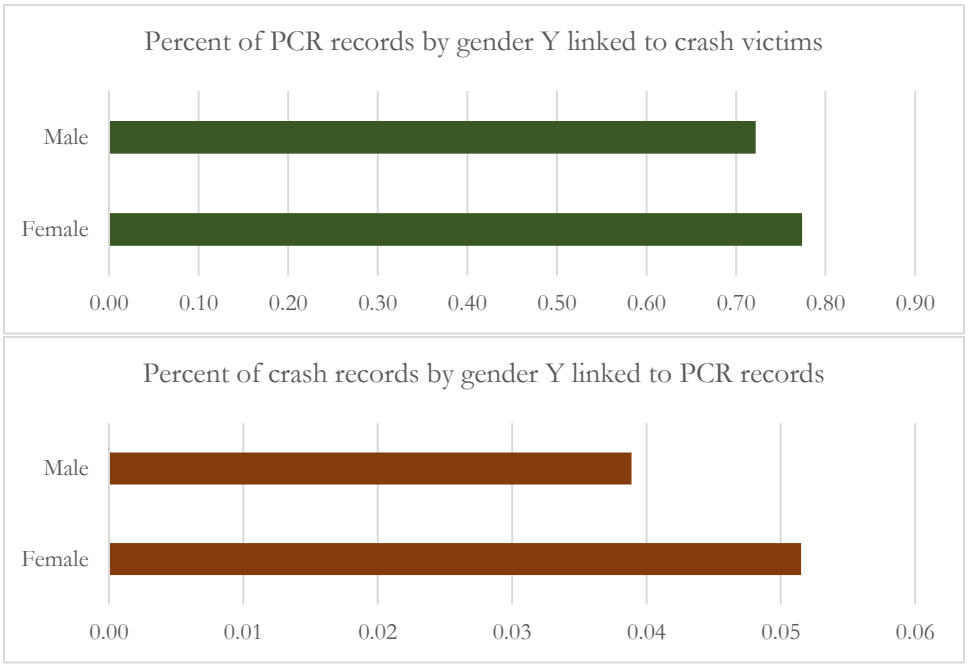


Figure 3.7. (a) Percent of PCR records by gender Y linked to crash victims (b) Percent of crash records by gender Y linked to PCR records

In terms of event type, bicycle, pedestrian, and motorcycle crashes are linked at a lower rate than general MVC crashes (Figure 3.8). One of the reasons could be higher under-reporting of motorcycle, bicycle and pedestrian crashes, which were found in other studies as well (Doggett et al., 2018b; Short and Caulfield, 2016).

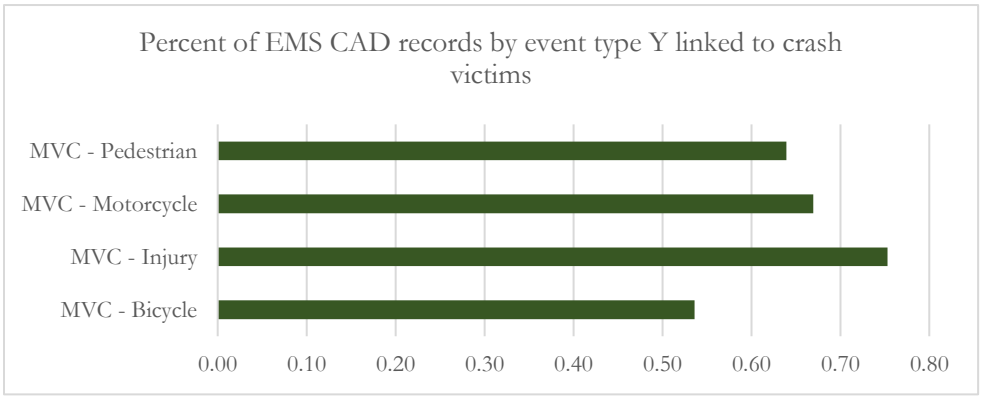


Figure 3.8. Percent of EMS CAD records by event type Y linked to crash victims

Linked data applications

Using the linked data sets, general safety monitoring, and data quality measures were obtained for four linked data sets to meet objective (5). In the next section the Crash-EMS CAD matches were used to compare EMS response time with crash data broken down by severity. In the section after, Crash-Trauma matched data was used to quantify the quality of injury that was reported in crash data by comparing ISS and emergency department disposition with KABCO injury severity.

Crash – EMS CAD

The time gap between when the CAD system received the 911 call and when the crash was reported by the officer has the potential to impact real-time applications of safety monitoring. If the time on the police report is assumed to be the exact time of the crash, an error will be induced when modeling the relationship between that crash time data from other sources at the time of the crash. In this effort, 11.9% of the matched records had police-reported times after the 911 call had been received in the CAD system, while 4.6% had a reported time more than 10 minutes after the 911 call was received. Assuming that the 911 call would not be made before the crash occurred, then these occurrences are possible errors in the police report.

Figure 3.9 shows a sample chart comparing the EMS response to features of the crash table. Response time is defined as the time between the 911 call and the EMS arrival at the scene. The importance of EMS response time is highlighted in the literature as a factor that impacts the survival rate (Amorim et al., 2019; Hu et al., 2017; Ma et al., 2019). EMS response time in Louisville was shown as a function of injury severity. Figure 3.9 contains 3,520 records. The parameter did not exhibit a clear pattern. This is not surprising as EMS typically makes

every effort to arrive on the scene as quickly as possible regardless of severity since the information dispatch receives from the callers is not always comprehensive.

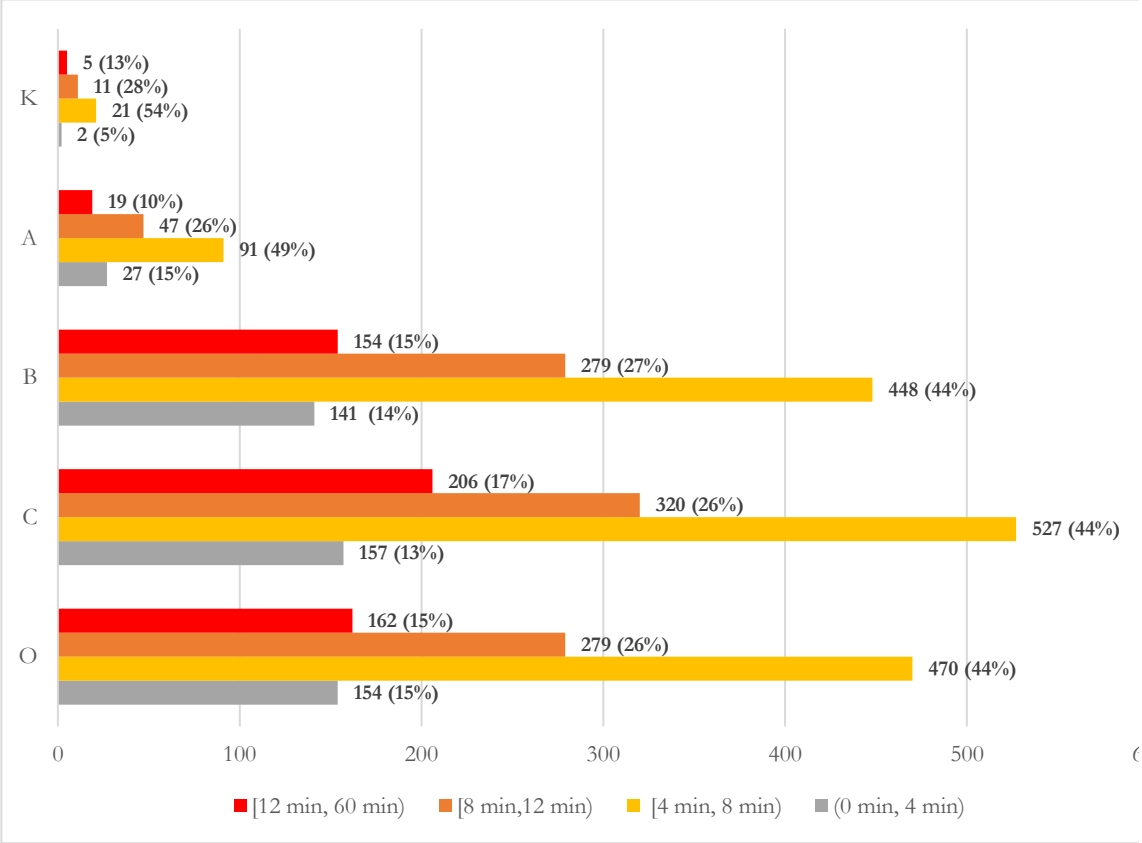
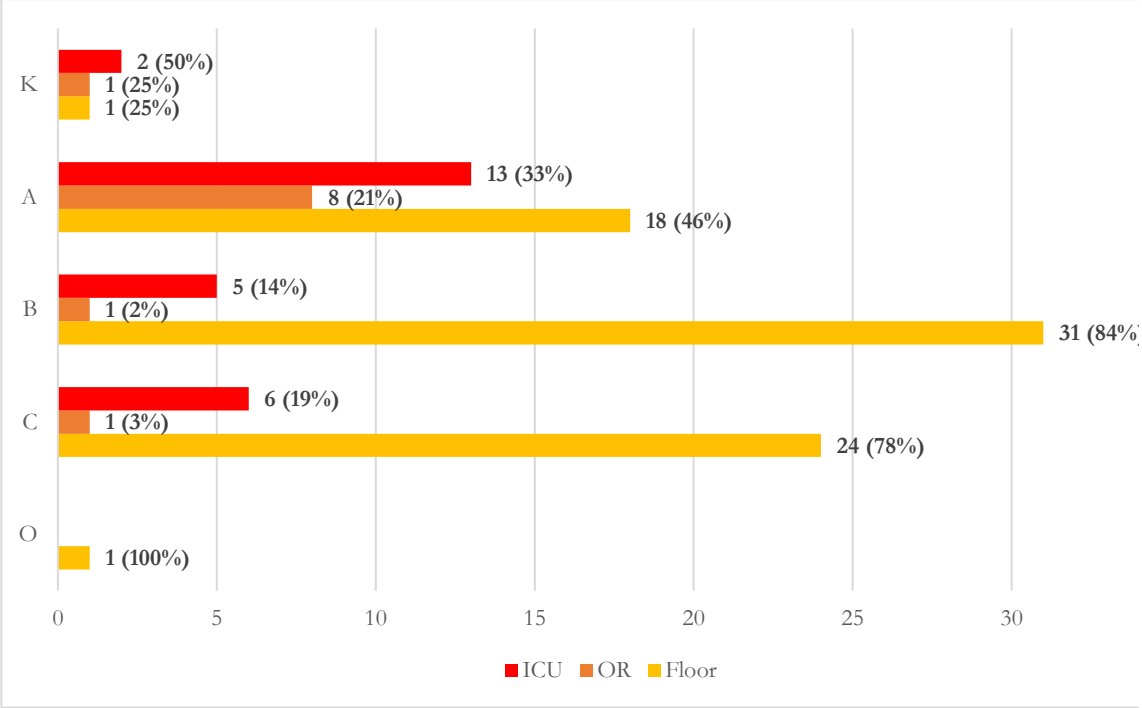


Figure 3.9. Distribution of response time (minutes) by KABCO injury severity (N = 3520)

Crash – EMS CAD – PCR - Trauma Registry

Figure 3.10 displays the relationship between the crash tables and the ULH trauma registry. Figure 3.10(a) displays the relationship between ED disposition and their police-reported injury severity. 13/113 records involved a B or C level injury ended up in either the Intensive Care Unit (ICU) or the Operating Room (OR). Figure 3.10(b), compares ISS to police-reported injury severity. ISS between 9 and 15 indicates a severe, non-life-threatening injury while an ISS of 16 or higher is life-threatening (Copes et al., 1988). Again, multiple C and one

O level crash were classified with an ISS of 16+ by a physician. These findings demonstrate that data linkage can help with tracking crash injuries to satisfy the objective (6). They also highlight the changing nature of injuries during the emergency response.



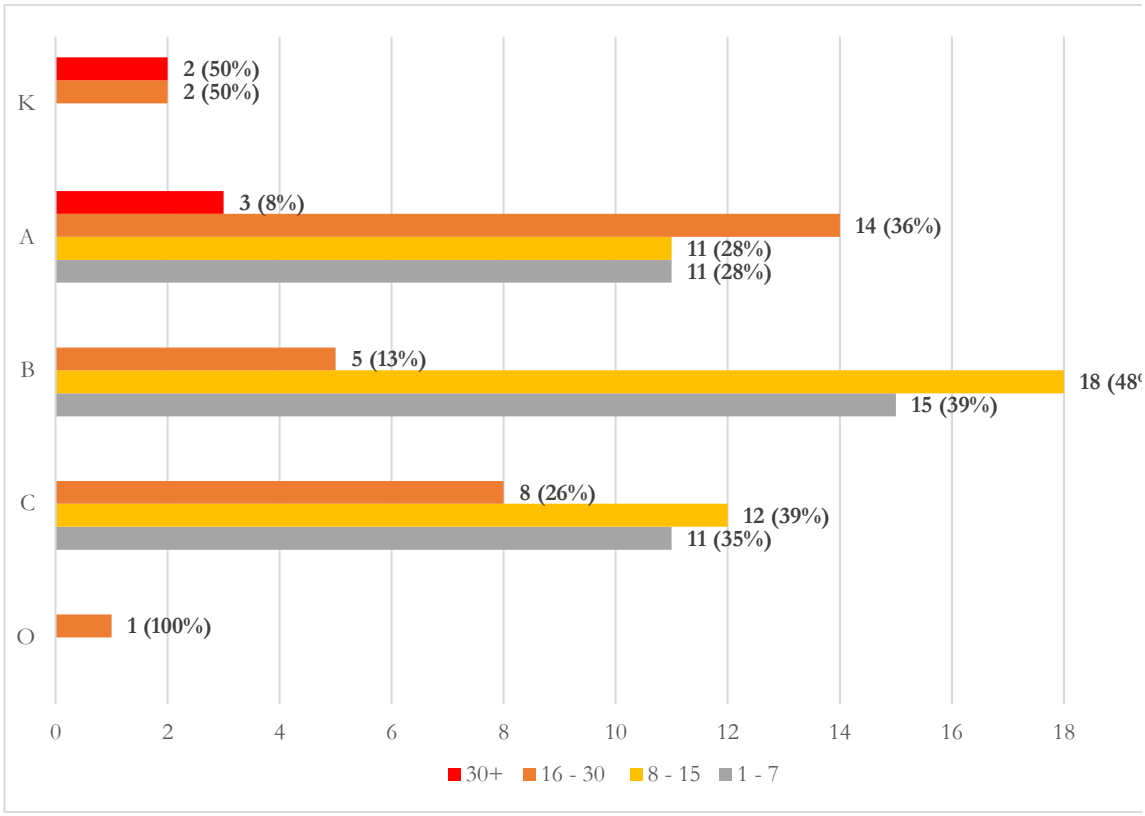


Figure 3.10(a) Distribution of ED disposition by KABCO injury severity (N = 113), **(b)** Distribution of ISS by KABCO injury severity (N = 113)

It should be noted that using the data for crash analysis purposes requires careful consideration of the study's objective. Some characteristics enhance the chance that the data go underreported and introduce bias to the analysis. For instance, pedestrian and bicycle crashes are more likely to go unreported, according to previous studies (Sciortino et al., 2005). This study's results should only apply to the cases where the outcome of the linked data is valuable. For instance, the trauma data cannot use to predict crash frequency since the data was reduced in size and in a biased way. However, it could be used to look at commonalities in crashes that end up in the trauma registry.

3.2.5 Conclusions

This study proposes a scalable heuristic algorithm for matching crashes to EMS and hospital trauma data; data sets that are not already inherently linked despite the clear benefits demonstrated both in this paper and other literature. The approach was implemented on a 9-month data sample from KY State Police, Louisville Metro Government, and the University of Louisville Hospital.

The six main objectives in this research included: (1) proposing an adaptive stepwise algorithm to link four crash-related data sets, (2) defining types of matched and unmatched records, (3) comparing the match rate results with the previous data linkage frameworks in the literature, (4) identifying factors that affect records linkage and bias, (5) visualizing the results and drawing some inferences from the matched data, (6) tracking the injuries from crash to EMS and trauma registry and highlighting the potential discrepancies. All six objectives were addressed in this paper. Based on the selected thresholds, results show 72.2% matches in EMS CAD data and 69.3% match rates in trauma registry records which are decent results comparing the studies in the literature. The sensitivity analysis result also suggests underreporting crash data. Further efforts have been conducted to provide some practical outcomes of the linked data.

The results of this study indicate that heuristic algorithms can achieve high linkage rates compared to previously achieved rates. A similar algorithm can be implemented beyond Jefferson County with some small adjustments to input parameters. It is anticipated that the distance and time threshold may need adjustment based on local crash frequency. While all the data sets used in this study followed nationally recognized standards, if users had additional fields with which to use in matching, those could be easily integrated within the heuristic proposed through a further check. When implementing the heuristic algorithm, it is critical

that users conduct a manual review of the results, and when using the data for subsequent analyses, the user should investigate biases associated with variables of interest.

Future studies could relax some of the assumptions made, such as restricting the EMS CAD data and trauma data to MVC-labelled, to evaluate the match rate further. Additionally, identifying a way to reliably match patients in the trauma registry that did not arrive by EMS to crashes would be valuable for quantifying underreporting. More investigation on bias, such as applying a statistical analysis approach to investigate which factors affect linkage rate and induce a bias is necessary. Moreover, a margin of error in the linkage is an inherent part of the linked data. Therefore, a sensitivity analysis needs to be conducted to quantify how the different error rate would affect the robustness of further inference. Finally, further research should be conducted to determine how to best use the resulting linked data to model and improve highway safety monitoring and data quality.

3.2.6 Practical Implications

Linked crash - EMS CAD – PCR – trauma registry data provides a valuable opportunity to evaluate the impact of prehospital care and emergency department care on crash outcomes. In general, policy steps could be taken to require cross-reporting and linkage of the data sets as the events occur to better monitor outcomes of injury crashes without requiring post-hoc linkage. This method can also realistically be integrated into a tool or software to undergo record linkage automatically.

3.3 Kentucky Statewide Crash-related Data Linkage[2]

3.3.1 Introduction

In this chapter, the process and outcome of a data linkage effort between the Kentucky State Crash Database, Kentucky Emergency Medical Services Information System, and the Kentucky State Trauma Registry were described. The result shows linked crash rate (linked crashes/total crashes) varies 0% to 23.9%, county-level injured persons match rate (linked individuals/total injured crash-involved individuals) ranges from 0% to 57.3% and county-level patient care reports match rate (linked individuals/total patient care reports) varies from 0% to 75%. A variable-level analysis was conducted to show which variables were more likely to be present in the linked data set compared to the individual data sets. The project team recommends investigation into additional data sets for inclusion in the linkage activities moving forward, updating query language for improved linkage rates, and investigation into low-linkage rate counties.

3.3.2 Data Sources and Management

This section will outline which and how datasets were obtained, and what fields were used in the data linkage approach. All datasets obtained were from 2018-2019.

Crash Data

Crash data consists of key information collected on police reports filed for crashes across the state. Crash data were obtained from the Kentucky State Police under a memorandum of understanding (MOU). The data are formatted following Minimum Model Uniform Crash

[2] Sections from "Kluger, R., Hosseinzadeh, A., Souleyrette, R. and Wang, T. Statewide Linkage of Crash, EMS, and Trauma Records. Kentucky Transportation Cabinet, 2022." included in this sub-chapter.

Criteria (MMUCC) standards (National Highway Traffic Safety Administration, 2017) with three tables (crash, vehicle, and person) linked by a unifying crash ID field. Both the crash and the person tables were used extensively in the data linkage.

Each crash record has a unique crash ID field and contains information about crash time, location, type, and more. In 2018, a total of 157,351 crash records were obtained. Table 3.4 outlines all fields present in the crash table. The specific fields used in the data linkage are in bold font.

Table 3.4. Fields available in crash table dataset

Master File #	Mile Post
Collision Date	Motorcyclist
Collision Time	Commercial Vehicle
Latitude Decimal Number	Young Driver
Longitude Decimal Number	Mature Driver
Weather Code	Pedestrian
First Aid Scene Indicator	Bicyclist
Time Notified	Distracted
Time Arrived	Aggressive
Time Roadway Opened	Impaired
Directional Analysis	Unrestrained
Time Last Left	Intersection
Year	Lane Departure
KABCO	Roadway Departure
KTC_RT	Median Crossover

For every individual involved in the crash, there is a record in the person table. Each person has a unique ID and is mapped to an individual crash through the crash ID. For 2018, a total of 458,546 crash–person records were obtained. Table 3.5 outlines all fields present in the person table. The specific fields used in the data linkage are in bold font.

Table 3.5. Fields available in a crash-person table dataset

Master File #	Injury Location Code
Unit Number	Position In/On Vehicle Code
Person Number	Restraint Use Code
Person Type Code	Trapped Code
Birth Date	Ejection From Vehicle Code
Death Date	Ejection Path Code
Age at Collision Time	Suspected Drinking Indicator
Gender Code	Year
Injury Severity	

Figure 3.11 shows the distribution of crashes in Kentucky. Note the larger clusters of crashes in Jefferson (Louisville), Fayette (Lexington), and northern Kentucky counties (Campbell, Kenton, and Boone).

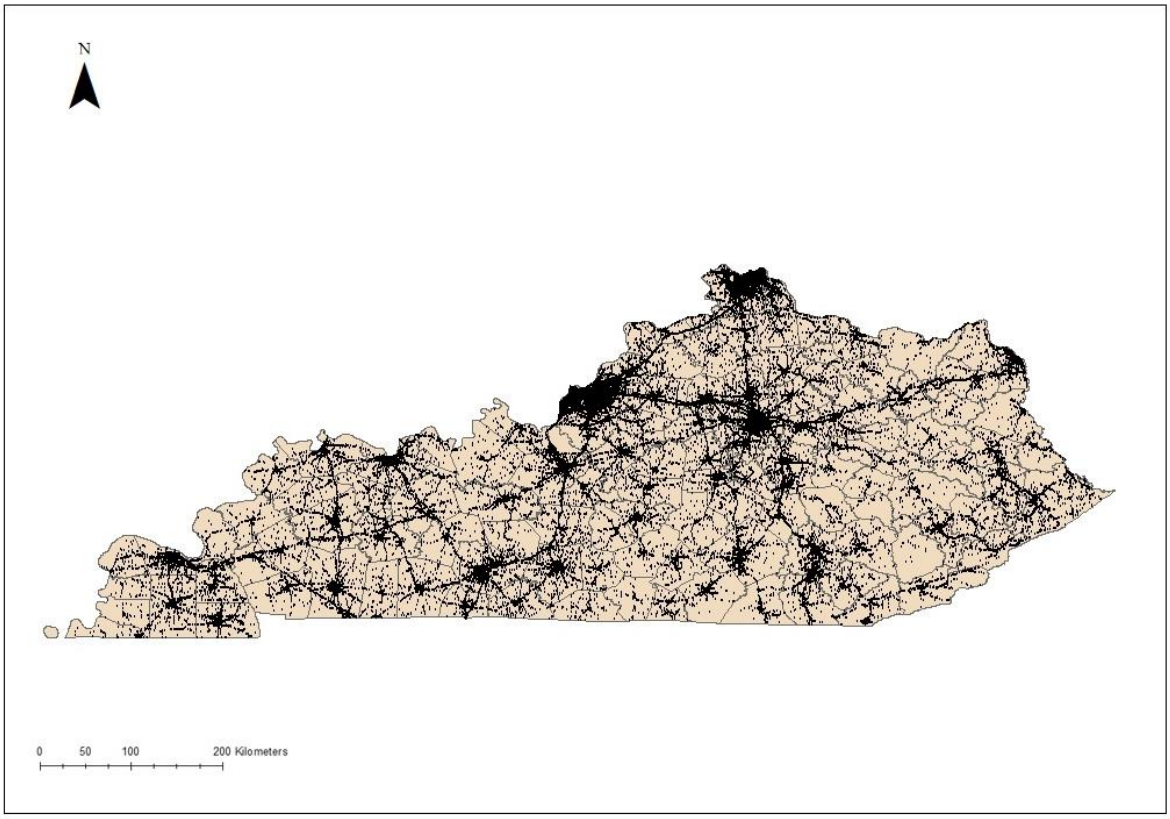


Figure 3.11. Distribution of crashes in Kentucky

EMS Data

EMS data contain a wide range of information about the EMS response to 911 calls. Each record represents a patient care report (PCR) filed by the team that responded to the emergency. KBEMS collects the data from EMS agencies across the state, standardizes it, and stores it in a state database called KEMSIS. The KEMSIS database follows National EMS Information System (NEMSIS) standard and contains 11 Tables:

- Table 1: EMS responded agency information
- Table 2: Patient medical examinations outcome
- Table 3: Injury automated collision notification

- Table 4: Patient medications given
- Table 5: Patient general body assessments
- Table 6: EMS response description
- Table 7: Scene information and status
- Table 8: EMS times
- Table 9: Vitals information
- Table 10 & 11: Patient examination information

In this study, EMS data were obtained through an open records request to KBEMS which required IRB protocols to be filed with the University of Louisville (U of L) and Kentucky Community and Technical College System (KCTCS), the parent organization of KBEMS. In the open records request, the following criteria were used to query the data from the KBEMS data repository:

1) Response Type (eResponse.05) matches 911 Response (Scene)
2) Complaint Reported by Dispatch (eDispatch.01) matches Traffic/Transportation Incident OR Scene Incident Location Type (eScene.09) contains any Street, Highway, Roadway.
3) Patient Care Report Narrative (eNarrative.01) contains one of the following keywords:

    Motor vehicle crash, Motor vehicle, accident, Motor vehicle incident, Car crash, Car accident, Car incident, Traffic crash, Traffic accident, Traffic incident, Transportation incident, Car wreck, Traffic collision, Motor vehicle collision, Fender bender, Automobile accident, Rollover, Hit-and-run, Traffic Incident, Transportation Incident, Truck Crash

For 2018-2019, a total of 57,083 records were requested. Under the HIPAA privacy rule requirements for de-identification, personally identifiable information was stripped from the dataset.

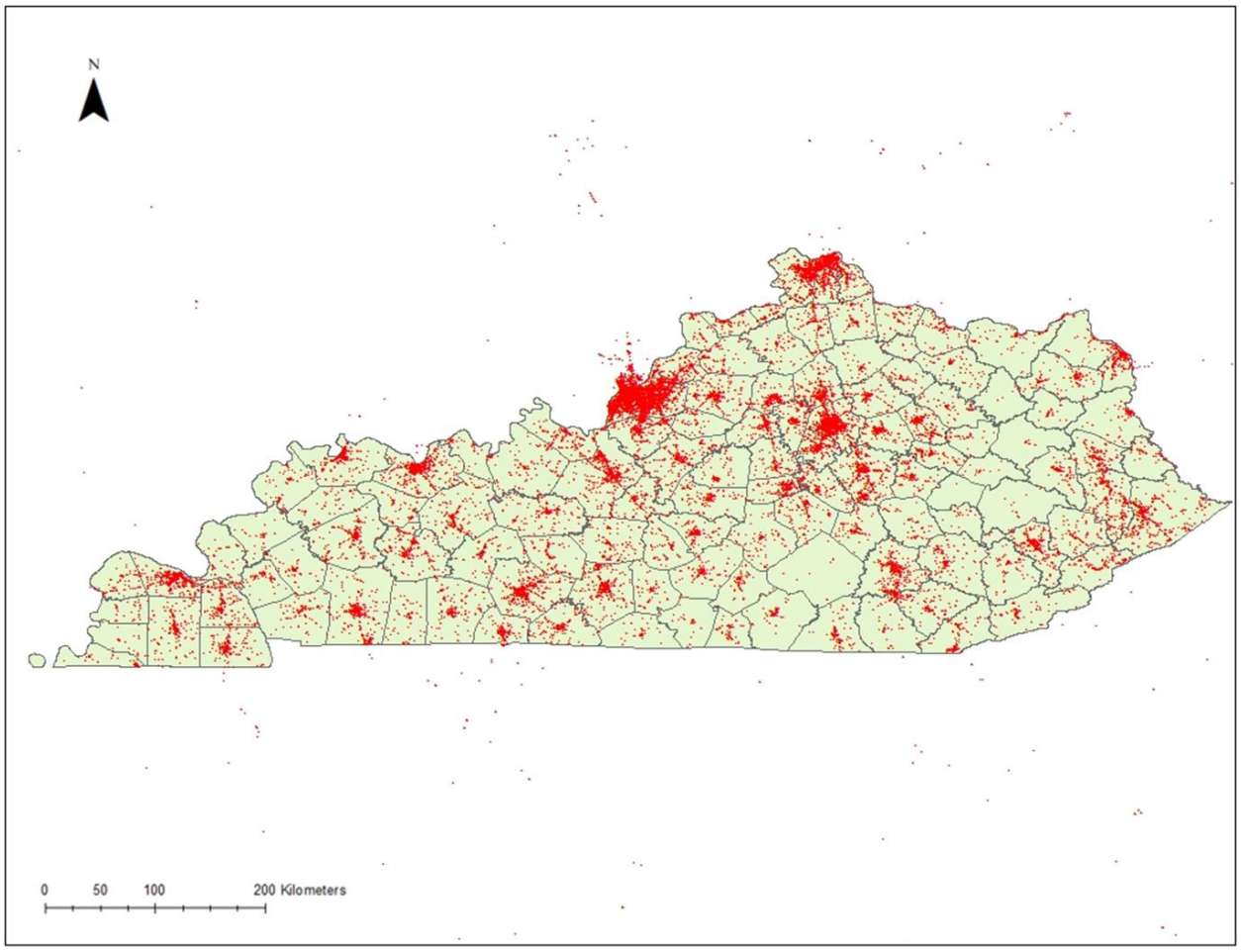


Figure 3.12. Distribution of EMS runs in Kentucky

Figure 3.12 shows the density of EMS runs at the county-level. Note the pronounced differences between counties. Jefferson County (26.72 per sq.mi) and Fayette County (13.07 per sq.mi) are the only counties with a density of EMS over 10. At the other extreme, 78 counties (out of 120) recorded less than 1 EMS run per square mile.

Trauma Registry Data

The State Trauma Registry is owned by the Cabinet for Health and Family Services (CHFS) and maintained by KIPRC. It contains data on emergency department admissions reported by trauma registries across the state.

The acquisition of Trauma Data required the signing of a data sharing agreement between U of L, UK and CHFS. Data is accessed through a secure virtual machine housed at KIPRC through a VPN. Table 3.6 outlines all fields present in the trauma data.

Table 3.6. Fields available in a trauma dataset

Date of Birth	Hospital Arrival Date & Time
Age	Temperature
Race	Alcohol Use
Gender	Drug Use
Incident Date & Time	Emergency Department Discharge Disposition
Injury Zip Code	Comorbid Condition
Airbag Deployment	Injury Diagnosis
EMS Notify Date and Time	Total ICU Level of Service
EMS Arrival Date and Time	Total Vent Days
EMS Left Date and Time	Hospital Discharge Date and Time
Transport Mode	AIS Severity
EMS Pulse Rate	Trauma Type
EMS Respiratory Rate	Cause Code
EMS Glasgow Coma Scale	Injury Detail
Inter Facility Transfer	Death in Emergency Department

Injury Severity Score	Trauma Type
Admit Service	Blood Alcohol Level
Injury Details	Position in the Vehicle
International Classification of Diseases, Tenth Revision (ICD-10)	ICD-10 Procedure

For 2018-2019, 12,803 trauma records are available in the dataset. Among them, 2979 records labeled as motor vehicle crashes, 267 pedestrian and 167 bikes. Also, there are 734 unlabeled records, 1217 records labeled as "other", 32 records labeled as unspecific, 12 not elsewhere classified and 7 not documented in the dataset that could possibly be related to motor vehicle crashes. However, due to the fact that the cause of the injury could be reported as "not-motor vehicle crashes" but still be related to motor vehicle crashes, the other causes of injury were not filtered out. A closer examination of the cases was conducted after linkage to filter out incorrect matches.

3.3.3 Method

Data Management and Preparation

MySQL was used in this project for data management, and datasets were stored in a relational database. R studio software was used for data management and statistical analysis (R Core Team, 2019). ArcGIS was used for mapping and spatial analysis. Moreover, although PCR data included latitude and longitude of the events, crash data used the addresses. The Google Maps platform (geocoding API[3]) was employed to provide latitude and longitude of crash locations. The addresses were prepared in a single field to be readable by the Google API. Of

[3] https://developers.google.com/maps/documentation/geocoding/overview

158,332 addresses (Jan 2018 to September 2019) representing all EMS runs, 150,662 were successfully geocoded (geocoding rate: 95.1%). The remaining 7760 records were returned as "NA" or the coordinate found was out of the study area and clearly wrong. For the rest of 7760, the google spreadsheet geocoding add-in tool (Awesome Table) was used. Using this tool successfully geocoded 6540 addressed in the study area (successful geocoding rate: 84.2%). With limiting the data to transportation-related EMS runs and 2018, the number of EMS runs entered to the linking process was 57,083.

Data Linkage

EMS runs and crash incidents are linked through location, time, age, and gender. Incidents reported within a three-kilometer distance and a 3-hour time window, for individuals with the same age and gender in the EMS PCR and crash reports database were considered to be matching pairs. Loops in R studio software were used to compare every two pairs in the crash and EMS data to find candidate matches. Figure 3.13 shows the algorithm used for this task.

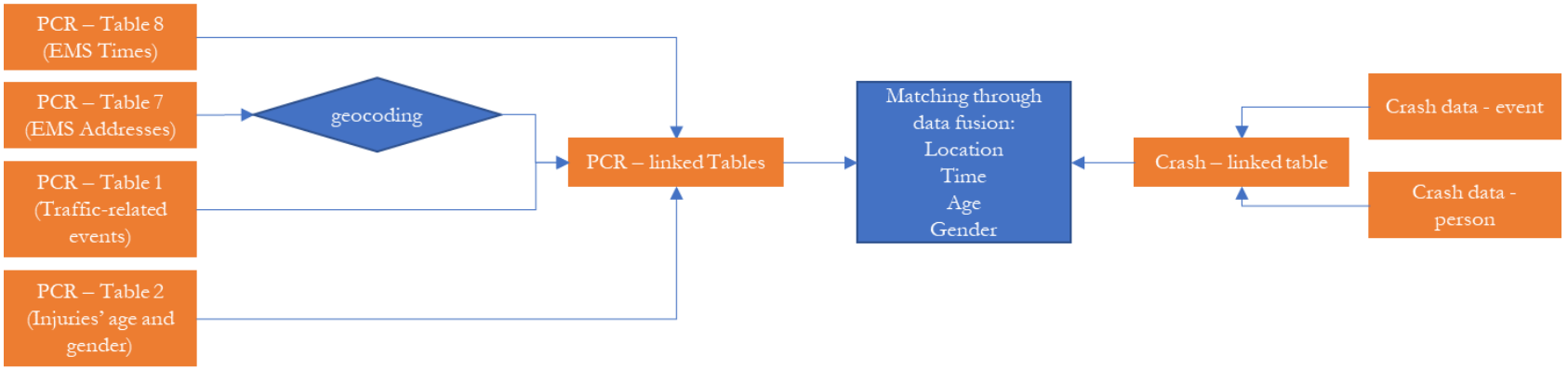


Figure 3.13. The algorithm applied to link PCR data and crash data

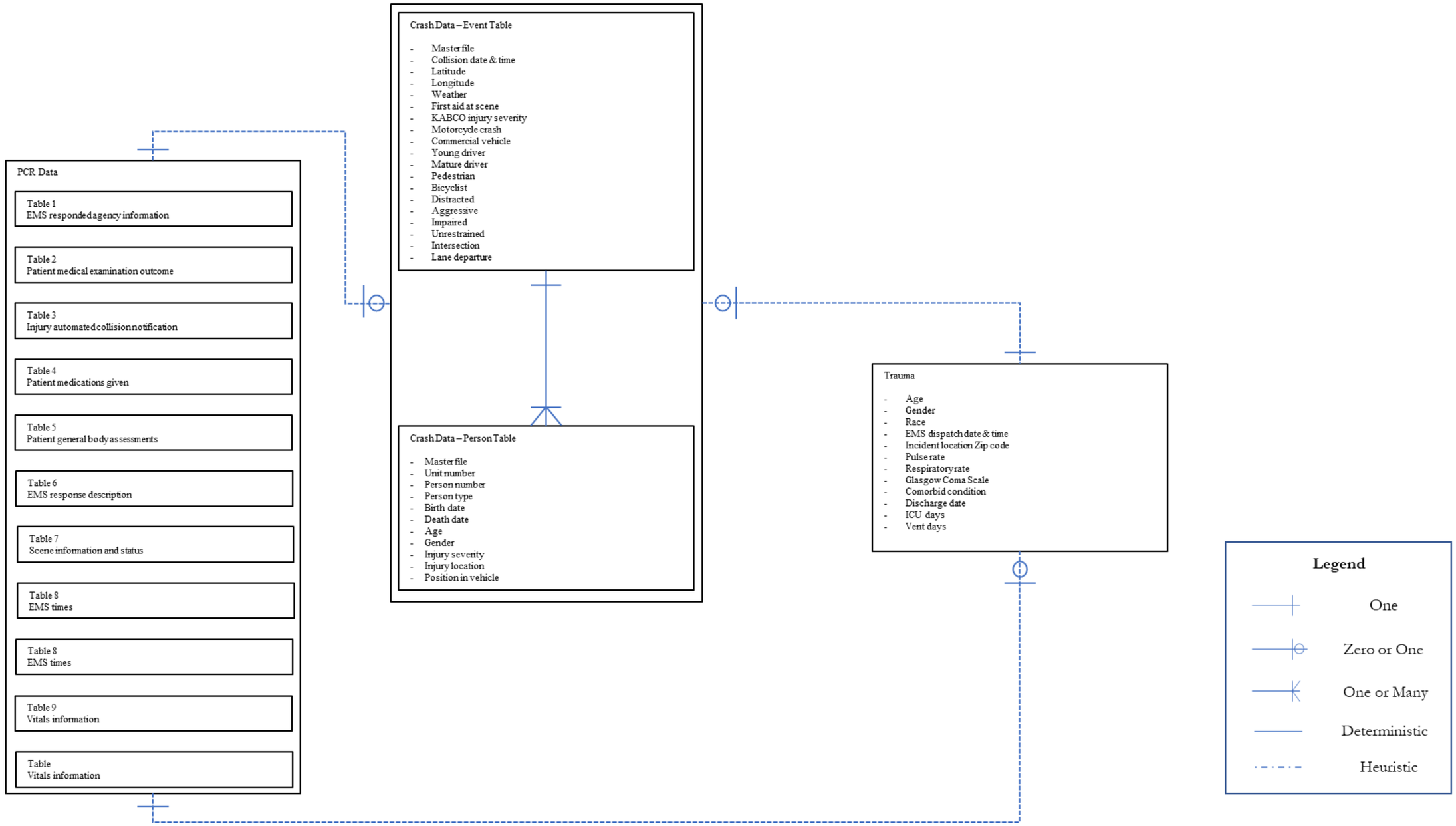


Figure 3.14. Entity relationship diagram of linked dataset

Figure 3.14 shows the entity-relationship diagram of datasets used in this project and relationships among them. A unique match is the favorable result (i.e., one crash-person linked with one EMS PCR). There were a few duplicate matches (i.e., one crash-person linked with two and more EMS runs, or one EMS run linked with two and more crash- persons), but these were not considered for further analysis in this project.

3.3.4 Police-reported Crash-EMS Linkage - State and County-Level Results

Key metrics tracked include the total number of records in each linked database and the rate at which a match was obtained for each database. These metrics were calculated for the entire state, as well as on a county-by-county basis. Table 3.7 shows the linkage rates of matched records on a state-level basis.

Table 3.7. Linkage percentage of crash-events/crash-person/EMS runs

Metric	Description	State-level Outcome	Map
% of linked crash records	# of linked crash IDs (matched with EMS runs) / # of all crash IDs	8.4%	Figure 3.15
% of linked crash-person records	# of linked crash-person IDs (matched with EMS runs)/ # of all crash-person IDs	5.5%	Figure 3.16
% of linked injured crash-person records	# of linked injured crash-person IDs (matched with EMS runs)/ # of all injured crash-person IDs	44.7%	Figure 3.17
% of linked EMS runs	# of linked EMS runs (match with crash- person table) / # all EMS runs	44.9%	Figure 3.18

Figure 3.15 shows the county-level crash data match rate (Linked Crashes/Total Crashes). Note that the match rate varies from 0 to 23.9% across counties. Most crashes do not require EMS, so the low percentage of total crashes linked is expected.

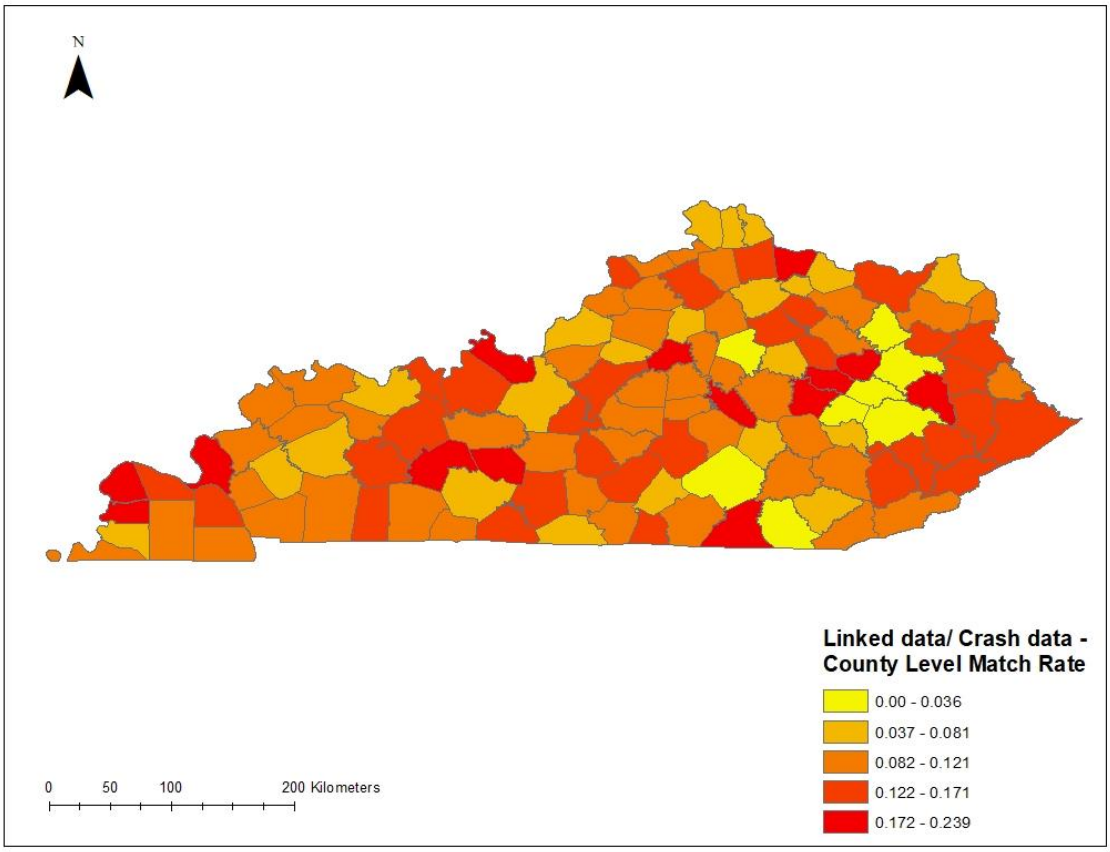


Figure 3.15. County-level crash data match rate

Figure 3.16 shows the county-level crash-person data match rate (Linked Crash-persons/Total Crash persons). Note that the match rate varies from 0 to 17.2% across counties.

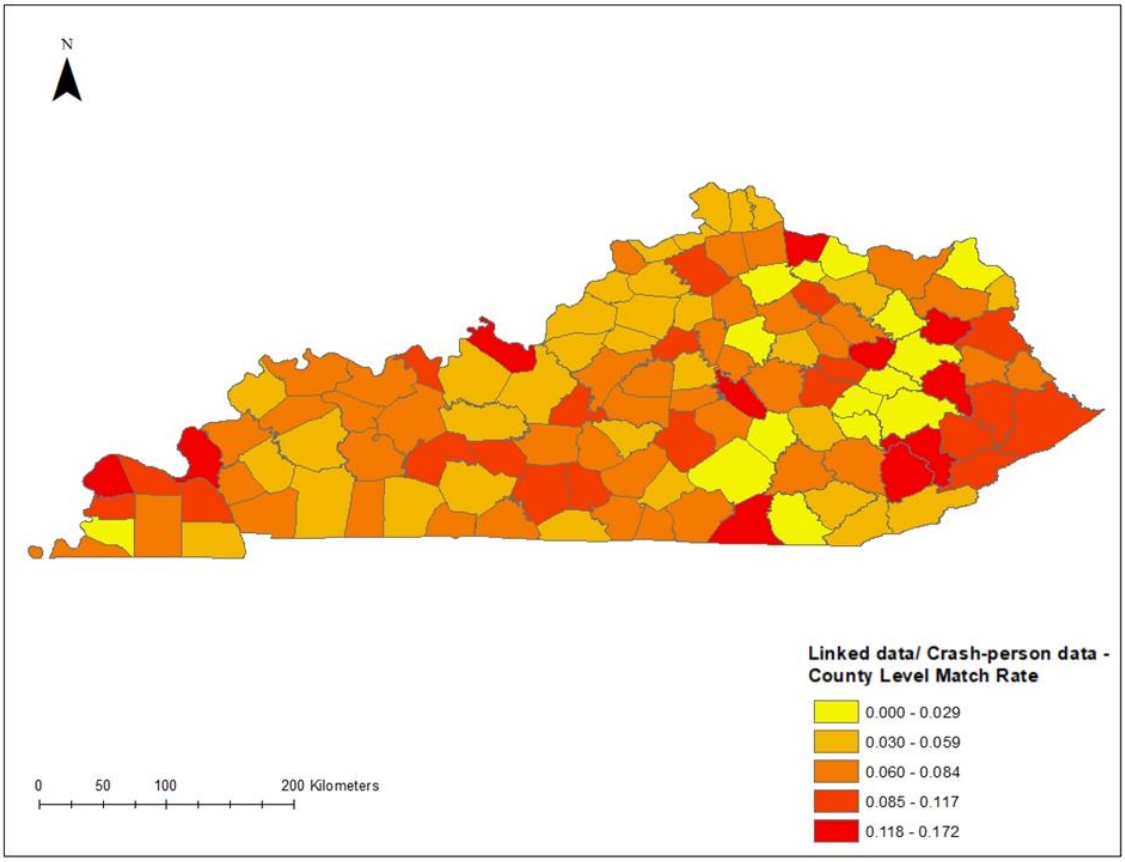


Figure 3.16. County-level crash-person data match rate

Figure 3.17 shows the county-level injured persons match rate (Linked Individuals/Total Injured Crash-involved Individuals). The match rate varies from 0 to 57.3% across counties.

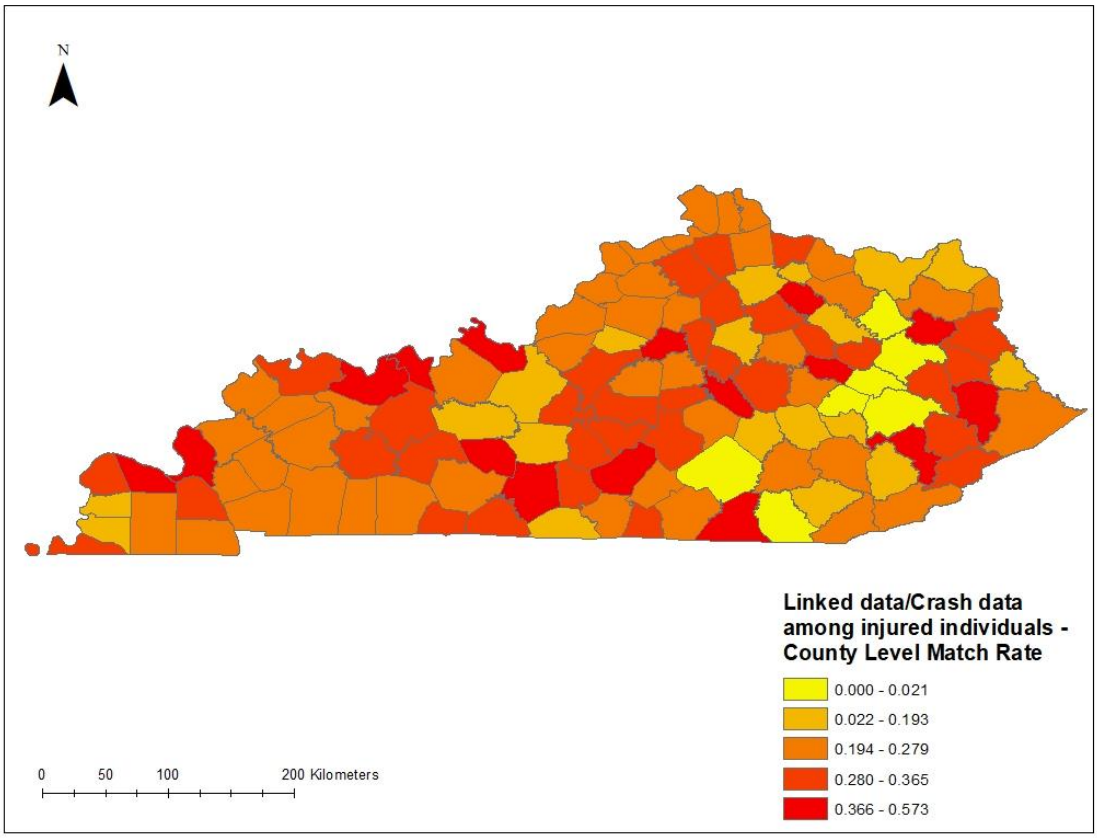


Figure 3.17. County-level injured persons match rates

Figure 3.18 shows the county-level PCR match rate (Linked Individuals/Total Patient Care Reports). The match rate varies from 0 to 75% across counties.

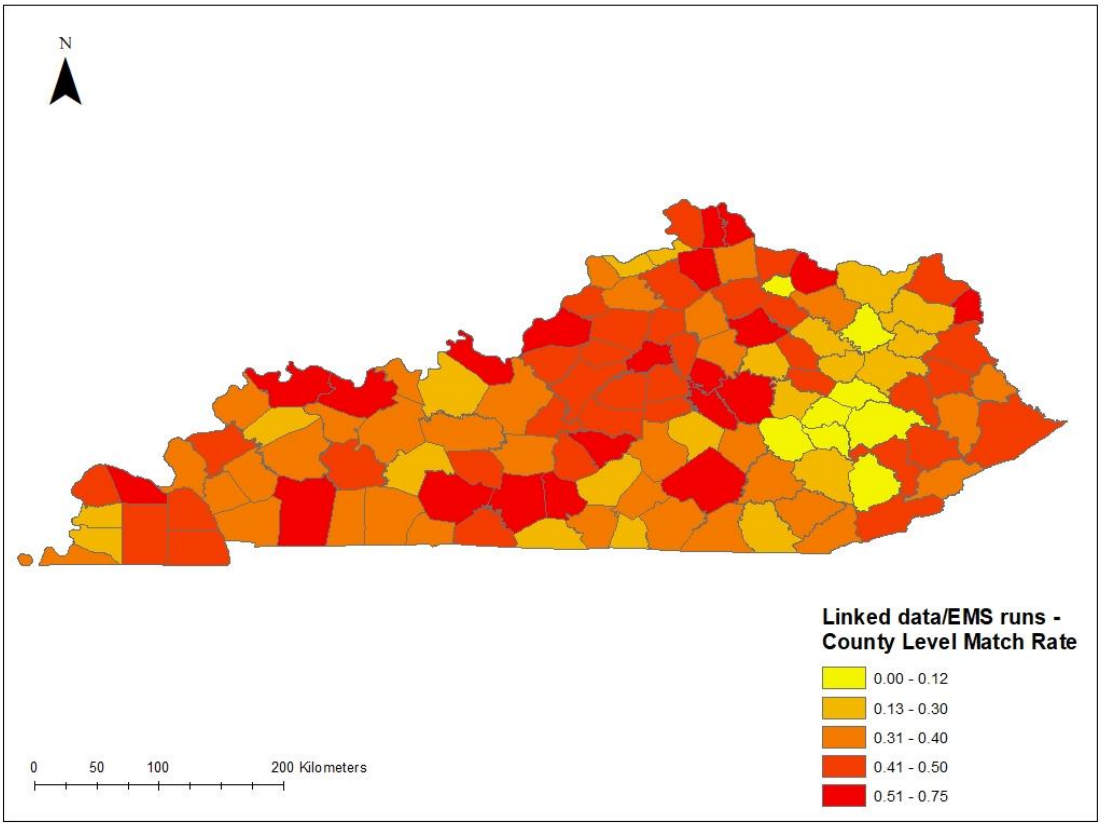


Figure 3.18. County-level PCR data match rate

Several observations can be made regarding the linkage success rate. While one would not expect every crash to match to an EMS patient care report, it should be expected that most EMS patient care reports should be assigned to a crash-involved individual, given how the EMS runs were queried.

Lower rates of crash linkages can be explained through several characteristics. First, and foremost, not all crashes require an EMS response. Of those that do require an EMS response, fatal crashes where this is not an opportunity to provide care also do not have patient care reports filed. Finally, it is possible that the query used excluded some cases. For example, if an EMS agency doesn't define a motor vehicle crash correctly, it might not end up in the EMS runs dataset based on the search parameters defined.

3.3.5 Variable-level Analysis of Match Rates

This section investigates differences between the linked datasets and the original datasets in terms of variable distributions. Table 3.8 displays characteristics of several variables among the linked data, crash data, and PCR data.

Table 3.8. Descriptive comparison of records in linked data, crash data and PCR data

	Linked dataset (n = 25,664)		Crash data (n = 157,351)		PCR data (n = 57,083)	
	Avg	sd	Avg	sd	Avg	sd
Age	38.23	20.17	37.91	19.69	40.36	21.01

	Linked dataset (n = 25,664)	Crash data (n = 157,351)	PCR data (n = 57,083)
Gender			
Male	47.99%	52.83%	54.16%
Female	52.01%	47.17%	45.84%
Injury severity			
O	37.87%	90.19%	-
C	33.05%	5.42%	-
B	22.61%	3.39%	-
A	5.29%	0.78%	-
K	1.15%	0.20%	-
Pedestrian			
Yes	2.59%	0.77%	1.48%
No	97.41%	99.23%	98.52%
Bicycle			
Yes	0.62%	0.18%	-
No	99.38%	99.82%	-
Intersection			
Yes	35.44%	25.93%	-
No	64.56%	74.07%	-
Suspect of Drinking			
Yes	4.66%	2.18%	-
No	95.34%	97.82%	-

Although the average age in the linked data and crash data are almost the same, injuries transferred to the hospital averaged approximately two years older. More males were involved in the crashes. However, more females were transferred to the hospital, and more females were available in the linked data. Moreover, more than 90 percent of the incidents in crash data are labeled as no-injury crashes. In comparison, this percentage for linked data is less than 40 percent. It's expected to have fewer no injury crashes in the linked data since the probability of request for an EMS would decrease for cases without injuries. The percentage of pedestrian and bicycle crashes is more than three times that of the linked data. More intersection crashes are also available in the linked data, probably because intersection crashes tend to be more severe than other crashes and involve more people (since there are usually multiple cars), leading to more opportunities for injury. Suspected of drinking cases were found to be more likely to be linked.

At the county level, there are different reasons for low match rates PCR data. First, these are the counties with very low numbers of crash/EMS runs, sometimes just because of the small size of the county. For example, the match rate in Roberson County is only 9 percent. However, one should consider that only 11 EMS runs met the query criteria in this county in 2018. In some counties, the match rate is suspiciously low. For example, for Lee and Wolfe counties, no traffic incident EMS runs were reported in 2018. Wolfe County had 448 crashes and 92 injuries during that time period. One recommendation from this finding is to reevaluate the query used and to investigate further how possible errors in reporting may have led to this issue.

Some counties with even relatively high numbers of EMS runs produced poor linkage results. For instance, in Leslie County, among 176 EMS runs, only nine were matched by PCR (5

percent). Pulaski (75 percent), McCracken (66 percent), and Meade (59 percent) counties have the highest PCR data match rates (Although in Pulaski, only 4 EMS runs were recorded in 2018).

Police reported Crash - EMS Runs -Trauma Linkage

The police-reported crash - EMS runs -trauma linkage was conducted between the linked dataset and trauma data. Date of birth, age, gender and race of the injured individuals in the linked data matched with the ones in the trauma data. Also, crash date and time in crash data matched with hospital admission time and a window of 12 hours have been used as the threshold. Incident date and time and EMS times reported in trauma data were also used; however, this field is not reported for most of the crashes thus were not helpful extensively. Incident zip codes in trauma data were the only location specific field to use for the linkage and matched with zip code reported in crash data.

After performing the initial linkage a few steps were conducted to validate the linked data. First, a couple of fields, such as position in the vehicle in both crash data and trauma data were compared. Second, the based on the location of the hospitals that the injured individuals were transported, the cases with high transported distance (more than 100 km) from scene to the hospital were gone under close attention to make sure these cases are true matches. The third step focused on injury details description. Text mining approach was used to make sure all the records, regardless of injury cause listed in another field, are actually related to motor vehicle crashes. The fourth step was a manual random check to ensure there is no systematic error in the matched dataset and figuring out the reasons for unmatched pairs. A detailed elaboration on the reasons of unmatched pairs were provided in the next section.

As a result of the matching process, the final linked crash, EMS runs, and trauma data is included 235 records. Table 3.9 shows the attributes of the linked dataset and the descriptive information of the fields.

Table 3.9. Descriptive statistics of some of variables in crash-EMS runs-trauma registry linked data

Attributes	Frequency	Percentage
Injury severity		
K	8	3.4%
A	93	39.6%
B	75	31.9%
C	51	21.7%
O	8	3.4%
Pedestrian	23	9.7%
Bicyclist	2	0.8%
Gender		
Male	138	58.7%
Female	97	42.3%
Transport Mode		
Ground ambulance	205	87.3%
Helicopter	26	11.1%
Private/public vehicle	3	1.2%
NA	1	0.4%
Admit Service		
Trauma	148	58.4%
Neurosurgery	9	3.5%
Orthopedics	30	11.8%
Medicine	12	4.7%
Others/NA	36	15.3%
Position in the car		
Driver	147	58.1%
Front Passenger	27	10.6%
Back Passenger	6	2.3%
Not specified/ NA	73	28.8%

Attribute	Average	S.D.	Min	Max
Age	43.1	21.8	1	96
Injury Severity Score	11.9	10.11	1	66

The matching process of the trauma data was conducted separately with police-reported crash dataset and EMS runs. 246 crash-trauma data and 286 EMS-trauma records were available in the these linked datasets.

3.3.6 Discussion, Recommendations and Conclusions

The objective of this sub-chapter included building and applying a framework to link crash data to EMS records and trauma records on a statewide, county-by-county basis in Kentucky. Data were obtained from Kentucky State Police (KSP), the Kentucky Board of EMS (KBEMS), and the Kentucky Injury Prevention Research Center (KIPRC). The results section outlined the linkage performance at the state and county levels.

There are some suspicious results in which further investigation into the data is needed. For example, although there were 191 individuals involved in crashes, including 25 injury individuals in Lee County in 2018, there were no EMS runs reported during the same period. Additional suspicious results such as Pulaski County (6527 crash-person records, 930 injured crash-person records and only 4 EMS runs in 2018) and Rowan County (2797 crash-person records, 341 injured crash-person records and 3 EMS runs in 2018). These warrant a deeper look into the queries made for EMS data, the methods implemented, and more.

Non-matched Records

The manual review provides an opportunity to ascertain the performance of the linkage algorithm. Overall, more than 100 records were reviewed manually to investigate the quality of the matching algorithm and further fine tune the parameters. Specifically, we reviewed non-matches and how inconsistencies lead to a lower match rate.

a) Data incompleteness

Some variables play a vital role in the linkage process as strong identifiers such as age, exact date of birth and gender. However, data incompleteness in these attributes causes the linkage serious problems. Data incompleteness in some of the important attributes is provided in Table 3.10.

Table 3.10. Incompleteness percentage in some of the important attributes

Attribute	No. of incomplete records	Total No. of Records	Incompleteness percentage
Age (EMS runs)	10,487	57,082	18.37%
Gender (EMS runs)	10,427	57,082	18.26%
Date of Birth (Crash data)	72,260	458,545	15.70%
Age (Crash data)	72,260	458,545	15.70%
Gender (Crash data)	55,343	458,545	12.10%

b) Incomplete or Inconsistent Formatting of Text Fields

Due to the formatting of addresses in the EMS data, geocoding was implemented to determine the latitude and longitude in EMS data. The addresses sometimes are incomplete or imprecise resulting in geocoding failures. For example, there is a pair of records in trauma data, and linked crash-EMS run data in which all the indicators matched except the location. After a careful deeper look at the attribute, it can be realized the issue is how precise the recorded address was in the EMS

data. The address was "KY-194, Pikeville, KY 41501" which could be the span 30 kilometers of a road. Formatting of addresses was also a notable issue.

c) Data Entry Error

Another case found was two pairs of matches in linked crash-EMS runs and trauma in which the birthday of the injured individual may have been recorded incorrectly. While all the other attributes matched and insinuated the pair records were related to a specific injured individual, the birthday in EMS data was "10/3/1986", while it was recorded as "10/3/1987" in trauma data. It is not possible to fully correct for data entry errors, though it is possible to implement checks and relax the parameters of the matching algorithm to catch the most common suspected errors. The most common entry errors must first be identified to account for this.

d) Transported with the helicopter or private/public vehicle

Some of the true matches that were not matched successfully through the linkage scheme are related to the fact that the injured individuals in cases were transported by a helicopter or private/public vehicles. So, these cases are not in EMS runs data then cannot find in the crash-EMS runs linked dataset previously matched. Therefore, it's not available in the crash-EMS runs-trauma linked data. 67.52% (8,645/12,803) of the records used ground ambulance for the EMS transport mode and the rest of 32.48% of the records were used other methods of transport. EMS data is a critical part of the linkage methodology, and the gap will lead to lower success rates in matching.

e) The transported from the referred facility

Some EMS runs included inter-facilities transfers (transfers between hospitals). In these cases, the time between the crash and hospital admission might be several days, even since the EMS run is still associated with a crash. In these cases, it's difficult to ensure the matches are accurate. Only 39.1% (5018/ 12803) of the records were transported straight from the scene to the hospital.

f) Recorded as motor vehicle crashes but it's not

Some cases in trauma records are recorded as MVC in trauma records but may not be classical cases included in other datasets. Digging into the injury detail description shows this phenomenon. For example, one record was recorded as "Ped vs. dump truck while working". This will count as an unsuccessful match of an MVC-related trauma record even though matching this type of case is not among the objectives of this analysis.

g) Reporting

Gaps in reporting varied among datasets. Follow up with data managers indicated that several agencies are failing to fully report data to their respective systems, particularly within KEMSIS and the Trauma Registry. For example, Rowan County reported three total EMS runs that were valid to be included within the linkage.

h) Categorization and Capture of Data

When querying the data sets from their original sources it is possible that the query did not capture how certain counties or agencies recorded information. A review of the consistency and quality of reported data may help to ensure each field is operating as intended.

3.3.7 Recommendations

Based on the project outcomes the team recommends the following steps be taken to further the findings of this project.

1. Additional quality checks into counties with low linkage rates relative to expected. Subsequent adjustments to the algorithm to improve linkage rates.
2. Modeling of expected linkage rates for key benchmarks based on county characteristics
3. Identifying new data sources for inclusion in this database to improve linkage rates or data coverage.
4. Continue data linkage efforts moving into 2022.

CHAPTER 4

FIDELITY OF HEURISTIC ALGORITHM COMPARED TO OTHER LINKAGE METHODS

4.1 Objectives

The objective of this chapter is to compare approaches to data linkage in traffic safety. This study used police-reported crash data, and emergency medical services run data in Louisville, Kentucky, from July 2018 to March 2019 and implemented a Bayesian record linkage with improved prior probability informativeness along with a stepwise adaptive heuristic algorithm. None of the previous studies were found by the authors to compare crash-related data linkage approaches. This study compared two common approaches, and consistency rate and discrepancy rate were reported. The results suggest (1) an approach to improve prior probability informativeness in the Bayesian record linkage of crash data (2) the superiority of the proposed heuristic algorithm compared to the Bayesian record linkage in terms of match rate (3) the consistency of more than 94% between the match pairs resulted from the two approaches. Moreover, the possible reasons behind these findings were discussed. Crash-related data could potentially provide a valuable opportunity to evaluate the impact of prehospital care and emergency department care on crash outcomes. Gaining in-depth knowledge regarding the linkage method can result in better quality linked safety dataset.

4.2 Introduction

Motor vehicle crashes (MVC) are one of the leading causes of death globally, and they impose a severe threat to public health. Police-reported crash data is the main source of information for safety analysis. However, other additional datasets can add further explanations to associated crash outcomes. Emergency Medical Services (EMS) data contains data about crash injuries and includes specifics of the injury (Burch et al., 2014), which often are not included in the safety analysis. The EMS records are neither inherently linked nor do they include an identifier to connect datasets. However, there is valuable information in EMS records related to traffic safety.

Historically, probabilistic record linkage has been the preferred method in research to link crash-related datasets. Specifically, Bayesian record linkage, is a powerful statistical approach to quantifying the probability that two records belong to the same event. Bayesian record linkage has been found to be difficult to implement in practice and can have limitations associated with the informativeness of prior probabilities (Milani et al., 2015). Meanwhile, deterministic approaches that do not quantify probabilities have also been implemented historically (Karmel et al., 2010). The objective of this research is to compare the results of deterministic and probabilistic record linkage for crash data.

In this research, Bayesian probabilistic record linkage was compared with a previously developed heuristic algorithm by the authors for linking datasets. Linkage was implemented to connect two datasets (EMS computer-aided dispatch (CAD) and EMS Patient Care Reports (PCR)) with police-reported crashes to improve road safety monitoring capabilities in Jefferson County, Kentucky, an urban county surrounding the city of Louisville. CAD systems collect EMS run data, and PCR data is information recorded by paramedics or emergency

medical technicians. CAD and PCR data can both bring forth information about the EMS response time to the crash and aspects of patient transport to the hospital, both of which can influence the injury outcome of the crash. In this research, first, the Bayesian record linkage approach was applied considering the available information to enhance the prior probability informativeness, and second, the results were compared with a stepwise adaptive heuristic algorithm previously developed by the authors (Hosseinzadeh et al., 2022) to investigate differences in the outcome across methods.

After a review of data linkage methods and challenges in the literature, the study data was described, the fundamentals of both the probabilistic Bayesian method and the heuristic algorithm were presented and an approach to assess prior and posterior probabilities were outlined. Next, the results of different probability thresholds were presented and followed with a discussion, conclusion and practical implications regarding the linkage findings.

4.3 Method

4.3.1 Initial Assessment of Potential Matches

The initial check utilizes a time-distance boundary to determine a pool of possible matches between EMS run reports and crash events. Date-time was extracted from the EMS CAD data as the time the 911 call was received, while the time of crash filed in the police report was used for the crash data. Equation 4.1 defines two distance thresholds D, as a function of time, D_t, and Euclidian distance, D_d (using position, x, and y) and allocates the EMS run report (j) to a crash event set (i).

$$\begin{cases} |t_{Ci} - t_{Ej}| \leq D_t \; and \; \sqrt{(x_{Ci} - x_{Ej})^2 + (y_{Ci} - y_{Ej})^2} \leq D_d & j \in i \\ |t_{Ci} - t_{Ej}| > D_t \; or \; \sqrt{(x_{Ci} - x_{Ej})^2 + (y_{Ci} - y_{Ej})^2} > D_d & j \notin i \end{cases} \quad (4.1)$$

The threshold used for the Bayesian approach was 3-day and 3-km and for the heuristic algorithm a boundary of 1-hour and 1-km was used. The Bayesian used a larger span to allow the probability to filter pairs based on the probability of match outcome. Meanwhile, a heuristic should be more restrictive when determining likely matches and lean on the idea that two crashes at the same location and the same time are inherently unlikely. The Bayesian approach resulted in 582,298 EMS and crash pairs, while for the heuristic method, 5,382 potential pairs.

4.3.2 Bayesian Record Linkage

Considering crash record, Ci, and EMS record, Ej, and the definition of conditional probability and implementing Bayes theorem, the following can be derived:

$$P(\bar{C}_i) = 1 - P(C_i) \quad (4.2)$$

$$P(C_i|E_j) = \frac{P(E_j|C_i)\,P(C_i)}{P(E_j)} \quad (4.3)$$

$$P(\bar{C}_i|E_j) = \frac{P(E_j|\bar{C}_i)\,P(\bar{C}_i)}{P(E_j)} \quad (4.4)$$

By dividing the two probability equations in equations 4.3 and 4.4, the following can result:

$$\frac{P(C_i|E_j)}{P(\bar{C}_i|E_j)} = \frac{P(C_i)}{P(\bar{C}_i)} \frac{P(E_j|C_i)}{P(E_j|\bar{C}_i)} \quad (4.5)$$

The components in equation 4.5 are classified based on Bayes theorem. $\frac{P(\mathrm{C_i}|E_j)}{P(\bar{C}_i|E_j)}$ is called posterior odds, $\frac{P(\mathrm{C_i})}{P(\bar{C}_i)}$ is called prior odds and $\frac{P(E_j|\mathrm{C_i})}{P(E_j|\bar{C}_i)}$ is called the likelihood ratio. Assuming $E_1|\mathrm{C_i},\ \ldots E_j|\mathrm{C_i}$ are independent and implementing Bayes theorem:

$$\frac{P(\mathrm{C_i}|E_1,E_2\ldots E_j)}{P(\bar{C}_i|E_1,E_2\ldots E_j)} = \frac{P(\mathrm{C_i})}{P(\bar{C}_i)}\,\frac{P(E_1|\mathrm{C_i})}{P(E_1|\bar{C}_i)}\ldots\frac{P(E_j|\mathrm{C_i})}{P(E_j|\bar{C}_i)} \qquad (4.6)$$

Equation 4.6 is the general form of equation 4.5 when there is more than one likelihood ratio. $P(\mathrm{C_i})$ denotes the probability that a record in the crash data matches its associated EMS runs by random chance. $P(E_1|\mathrm{C_i})$ is the probability that EMS record E1 matches crash record Ci based on the information of first criteria prior information, and $P(E_1|\bar{C}_i)$ is the probability that EMS record E1 does not match crash record Ci based on the information of first criteria prior information (Clark, 2004).

Prior odds ratio assessment

In Bayesian record linkage, the prior odds are the probability of two records getting matched divided by the probability of those records not getting matched based on prior information. The next step is evaluating prior probability for each of the matching criterion. Figure 4.1 shows the frequency of crashes by spatial distribution ($P(\mathrm{C_{Di}})$), Figure 4.2 demonstrates the frequency of crashes by age ($P(\mathrm{C_{Ai}})$), Figure 4.3 exhibits the frequency of crashes by time of the day ($P(\mathrm{C_{Ti}})$), and Figure 4.4 displays the frequency of crashes by gender ($P(\mathrm{C_{Gi}})$). The information gained through the matching criteria distributions was applied to generate prior probabilities.

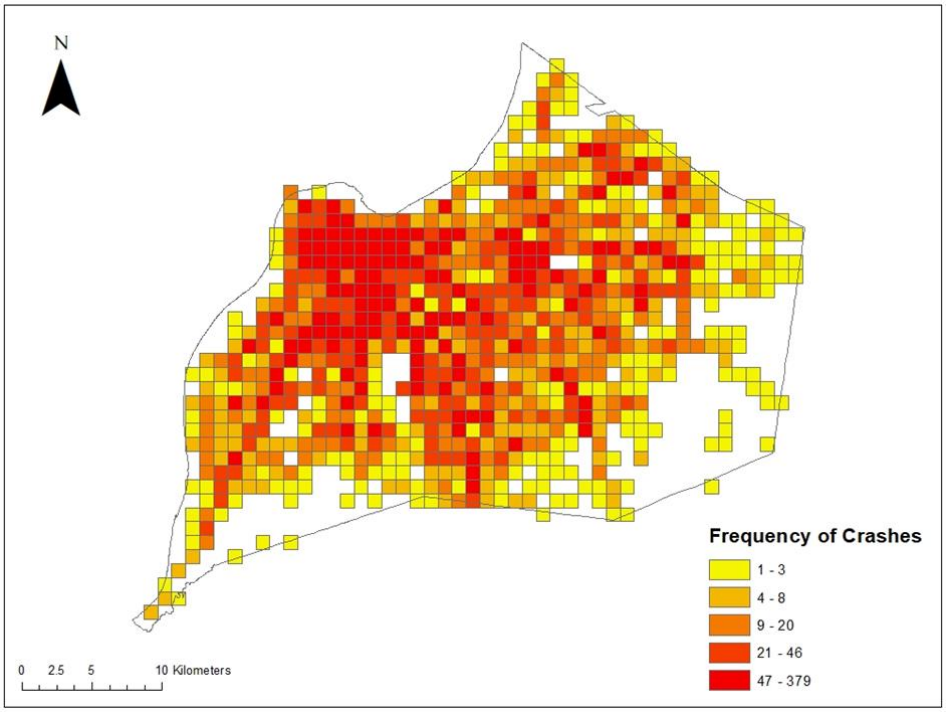


Figure 4.1. The frequency of crashes in sq-km unit cells in Jefferson County, Kentucky (July 2018 – March 2019)

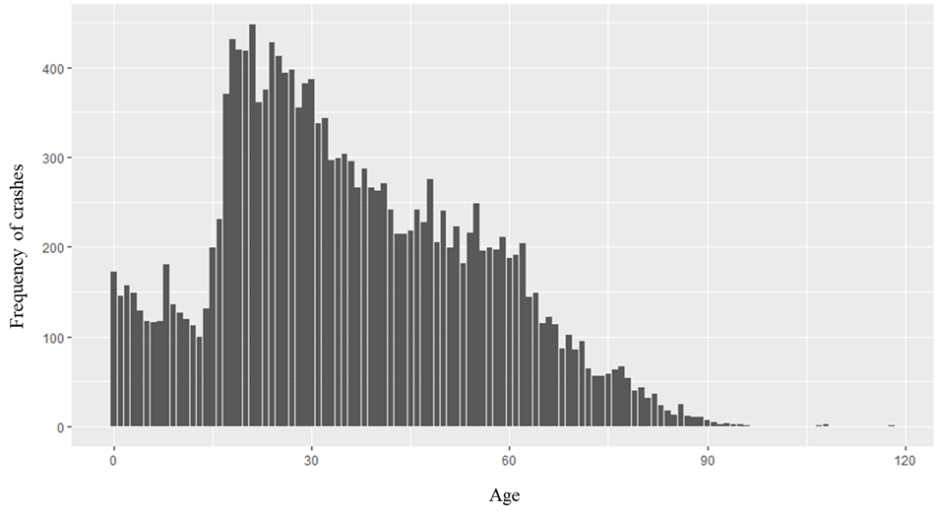


Figure 4.2. The frequency of crash-person records by age in Jefferson County, Kentucky (July 2018 – March 2019)

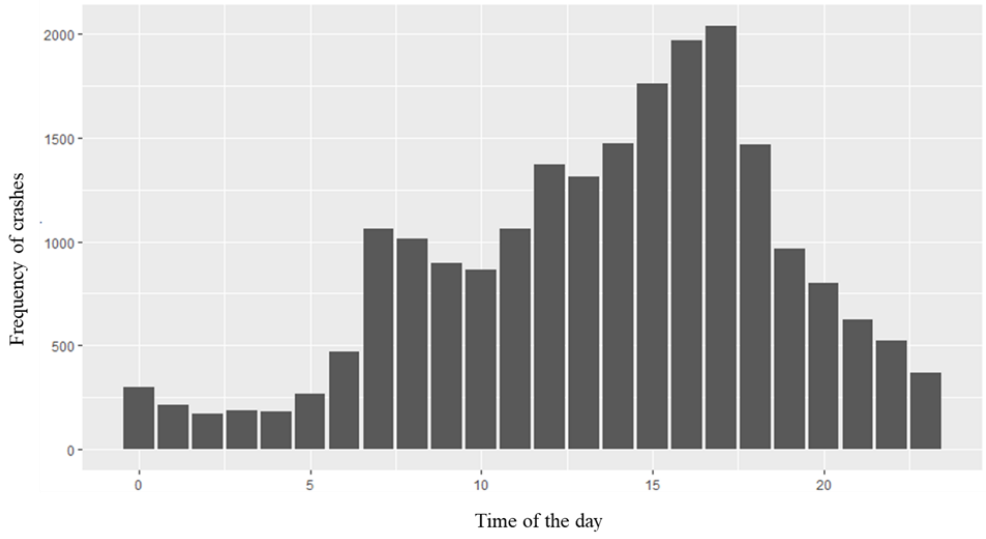


Figure 4.3. The frequency of crashes by time of the day in Jefferson County, Kentucky (July 2018 – March 2019)

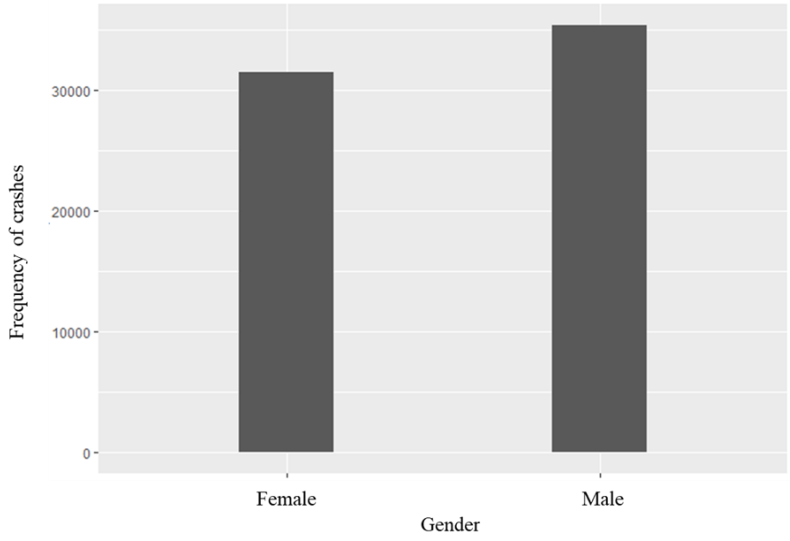


Figure 4.4. The frequency of crashes by gender in Jefferson County, Kentucky (July 2018 – March 2019)

Likelihood odds ratio assessment

In Bayesian record linkage, the likelihood probability assessment is the probability of two records matching based on the features. In this chapter, first, a large set of all possible matches was created using the Euclidean distance in equation 4.1. Reducing pairs of EMS runs and

Crash events that happened very far apart before implementing the Bayesian approach, reduced the computational burden. Afterward, features specific to potential EMS and Crash matches were used in the record linkage approach, including distance, time, gender, and age were considered as matching criteria.

Furthermore, likelihood probabilities based on each matching criterion were found. Two decay functions were used to assess the probability of matching for time and location. Based on these two decay functions, the probability of matching decreases as the distance and time between the two pairs increase. Since records were eliminated as possible matches if the time or distance boundary from the crash event was 3 km or 3 hours using equation 4.1, the decay function ranges from 1 to 0. Also, the likelihood was assumed for gender and age, as shown in Table 4.1. If the age exactly matches, $P_G(E_j|\mathrm{C_i})$ considered as 0.99. If there is a one-year difference, there is still a chance but not as high as being exactly match, so it was assumed as 0.8. In cases of more than one-year difference, the probability was assigned as 0.01. It means if the pairs are real matches, the other indicators should be completely matched to make up for the inaccuracy in the age, and the final likelihood ends up being more than 90%. For the records that age and gender were not available, the likelihood probability was calculated only using distance and time.

Table 4.1. The likelihood ratio of matching criteria assessment

Matching Criteria	Likelihood Ratio
Distance	$P(D_{ij}\|C_iE_j) = 1 - (\frac{D_{ij}}{3})^2$
Time	$P(T_{ij}\|C_iE_j) = 1 - (\frac{T_{ij}}{3})^2$
Gender	$If\ ((G_i = G_j), P(G_{ij}\|C_iE_j)$=0.99),
	$If\ ((G_i \neq G_j), P(G_{ij}\|C_iE_j)$= 0.01)
Age	$If\ ((A_i = A_j), P(A_{ij}\|C_iE_j)$=0.99),
	If (abs $(A_i - A_j) = 1, P(A_{ij}\|C_iE_j)$= 0.8),

$$If\ ((A_i \neq A_j), P(A_{ij}|C_iE_j) = 0.01)$$

4.4 Results

Types of different possible match outcomes are discussed in table 3.2. Each record status is described with one of the linkage terms defined in the table 3.2.

The results of implementing two record linkage approaches are presented in Table 4.2. The first row represents pairs of records whose posterior probability was higher than 99% in the Bayesian approach. The second column shows all the matches gained through Bayesian, including duplicate matches. Duplicate matches happen when an EMS run matches with more than one crash or a crash matches with more than one EMS run. After filtering out the results based on the calculated match probability, results demonstrate crash-EMS unique matches (column 3). The number of records with unique matches after applying the heuristic algorithm is shown in the fourth column. For example, for 99% threshold, Bayesian reached to 903 unique crash- EMS events match and the heuristic algorithm reached to 3955. The fifth and sixth columns demonstrate the number of unique matches in the Bayesian approach only and heuristic only, respectively. For example, the Bayesian approach identified 196 unique matches that the heuristic did not if the Bayesian probability threshold was 99%. Common crash/EMS event here means the ones that a crash or an EMS event in one can be found in the other one. The following columns show how many common crashes/EMS events were identical and how many of them were different.

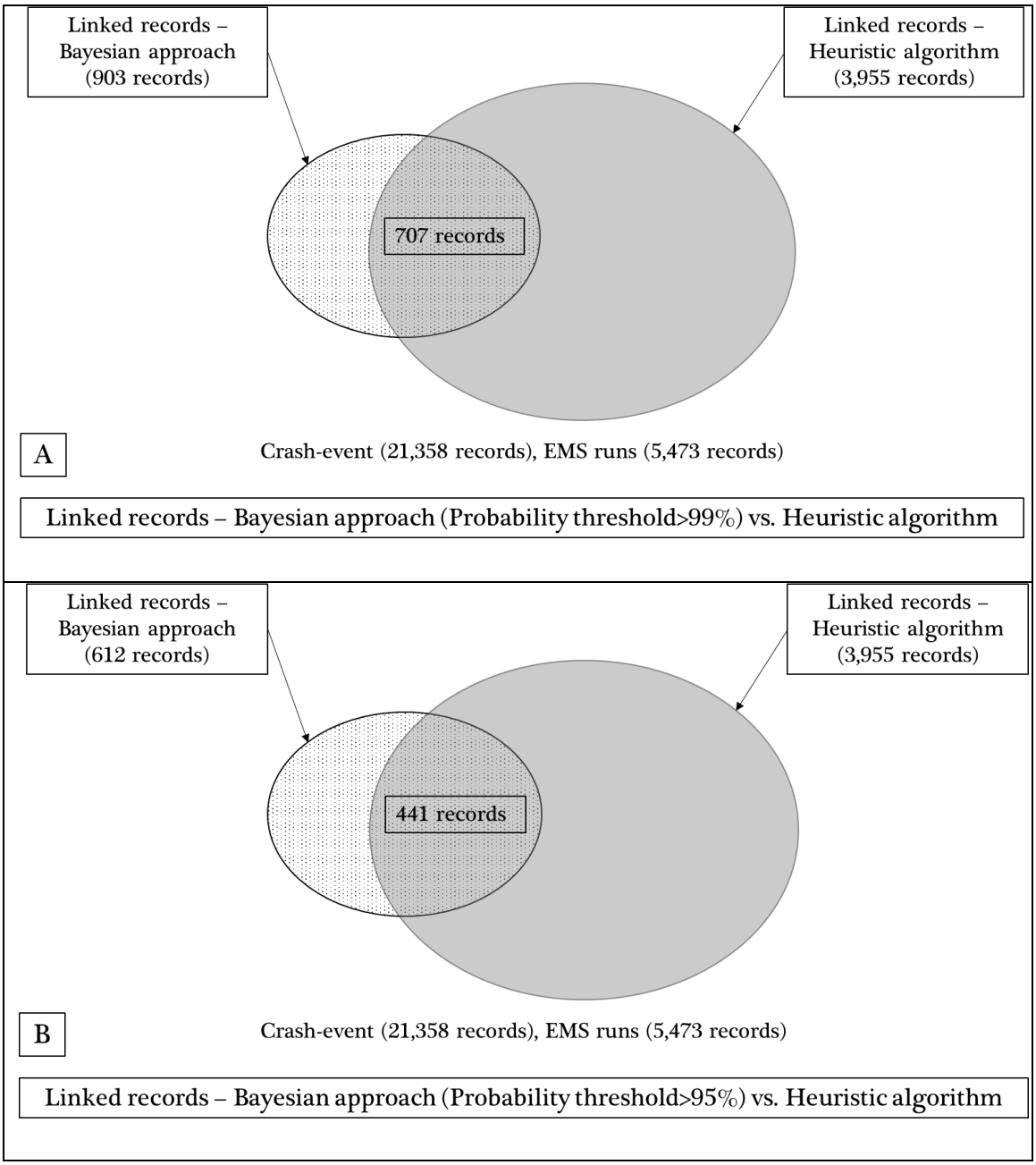
The number of common crashes/EMS events decreased as the cut-off threshold decreased from 99% to 90%. The same trend was observed in the number of the same matches between the two approaches. Results of the Bayesian approach in all three thresholds show substantial differences to the outcome of the heuristic algorithm. The 99% probability of a match reached

22.8% (903/3955) of the heuristic algorithm unique matches, and this rate for 95% and 90% thresholds were 15.4% and 17.2% of the unique matches in the heuristic algorithm. This outcome shows with loosening the threshold, the number of all Bayesian matches, including duplicates, would increase; however, it doesn't guarantee that the number of unique Bayesian matches also increases.

Among the unique matches from the Bayesian approach, 81.3% (735/903), 74.6% (457/612), and 57.8% (432/682) of 99% to 90% thresholds were found in the heuristic unique matches. 96.1% (707/735), 96.4% (441/457) and 94.4% (408/432) of the unique matches were available in both linked datasets for 99%, 95% and 90% thresholds, respectively, were identical. Figure 4.5 shows the visual representation of the two methods in different probability thresholds.

Table 4.2. Comparison of the Bayesian and heuristic data linkage results

1	2	3	4	5	6	7	8	9
Probability threshold (Bayesian)	All Bayesian matches-crash - person records (Including duplicates)	Bayesian unique matches	Heuristic unique matches	Bayesian only	Heuristic only	Bayesian and heuristic in common Crash ID or EMS ID	Same Crash -EMS match	Different Crash-EMS match
> 99%	2,755	903	3,955	196	3248	735	707	28
> 95%	7,522	612	3,955	171	3514	457	441	16
> 90%	12,287	682	3,955	274	3547	432	408	24


Linked records – Bayesian approach (903 records)
Linked records – Heuristic algorithm (3,955 records)
707 records
A
Crash-event (21,358 records), EMS runs (5,473 records)
Linked records – Bayesian approach (Probability threshold>99%) vs. Heuristic algorithm
Linked records – Bayesian approach (612 records)
Linked records – Heuristic algorithm (3,955 records)
441 records
B
Crash-event (21,358 records), EMS runs (5,473 records)
Linked records – Bayesian approach (Probability threshold>95%) vs. Heuristic algorithm

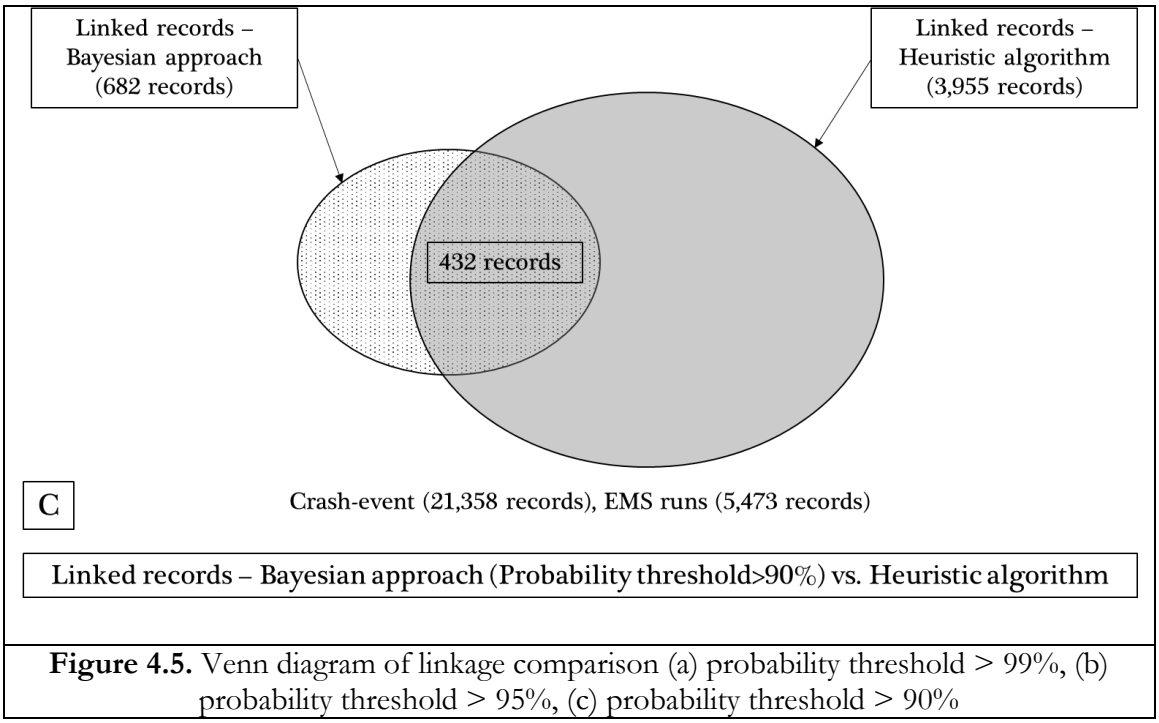


Figure 4.5. Venn diagram of linkage comparison (a) probability threshold > 99%, (b) probability threshold > 95%, (c) probability threshold > 90%

4.5 Discussion

Comparing two record linkage approaches shows the superiority of the heuristic stepwise adaptive algorithm compared to the Bayesian approach. The Bayesian approach was just able to reach up to 22.8% of the number of match records of the heuristic algorithm. However, there is no way to determine which one of the pairs available in heuristic algorithm matches and Bayesian matches are true matches.

The performance of the heuristic and Bayesian approaches highly depends on the linkage features available in both datasets. The main factor that drives the wide gap between two linkage approaches is the adaptive nature of the heuristic algorithm. The heuristic algorithm reconsiders the duplicates to find the true match between them. It moves the already unique match pairs to the unique match records in every step, freeing the chance to find a unique match for the potential match pairs that used to be among duplicates in the previous

step. Both datasets inevitably might include some false matches. However, the possibility is lower for the ones found as unique matches as a result of both approaches. The discrepancy between the two approaches was 3.9%, 3.6% and 5.6% for various thresholds suggesting a high overlap between the two approaches. However, comparing two approaches also allows taking different matches under close attention to distinguish some potential false matches. While there are no external independent data sources to verify the validity of matches, comparing would provide a valuable opportunity to double-check the matched pairs. One of the upsides of the Bayesian linkage record is the fact that each linkage record is provided with the probability of match leaving a margin of error, which is not able to be assessed for the pairs in the heuristic algorithm.

4.6 Conclusion

The data linkage gives significant insight into injury trends in several safety emphasis areas and provides a variety of underlying factors which would not be available without linking datasets. This research examined and compared two record linkage methods, which could help explain crash safety assessment. Results show the superiority of the proposed heuristic record linkage compared to the Bayesian approach, as the Bayesian approach reached only up to 22.8% of unique match pairs. Also, the results shed light on some match discrepancies for further investigation.

There are several limitations in this study that are necessary to point out. First, none of the match outcomes can be surely verified unless a unique identifier is present. To counteract this, manual review was conducted extensively, however there is no way to perfectly quantify the match rate. Second, the functions used to determine the likelihoods for time, distance, age and other factors can be optimized through further research. While these

assumptions are reasonable, further investigation might help to find the optimized decay functions. These likelihood ratio assumptions depend on the quality of the data of the study. For example, with a manual review of a sample of available Crash-EMS run linked data, it's possible to evaluate a more accurate likelihood ratio. Third, age and gender information in crash data was 70.2% and 72.9% complete, preventing inclusion in the analysis and reducing the match rate. However, the completeness for more severe injuries was higher.

4.7 Practical Implications

Policy steps could be taken to require cross-reporting and linkage of the datasets as the events occur to better monitor outcomes of injury crashes without requiring post hoc linkage. Incorporating two linkage methods (i.e., heuristic algorithm and Bayesian probabilistic linkage) to get a deeper insight into consistent and inconsistent records could possibly strengthen the linked data quality. This study layout the initial steps moving forward toward this goal.

CHAPTER 5

APPLICATIONS OF THE LINKED DATA

5.1 Objective

This chapter showcases what can be explored by using the linked crash-related datasets. After conducting the linkage and making sure the linked dataset adequately represents the datasets involved in the linkage process, applications of the resulted linked dataset can be investigated. Crash linked dataset added a couple of variables that have an impact on the crash outcome but were not available in police-reported and therefore were not included in the traditional safety analysis. These variables included the aftermath of the crash from on-scene to the hospital, such as EMS runs and trauma registry information. The following sub-chapters represent three examples of linked data usage. First, the association of injury severity and EMS times were explored. Furthermore, factors affecting EMS times were investigated. Last but not least, the injury misclassification of police officers and emergency physicians and the reasons behind the discrepancy were explored.

5.2 Do EMS Times Associate with Injury Severity?[4]

5.2.1 Introduction

In this section, two EMS times, response time and on-scene time, along with other crash-related explanatory variables, have been modeled to investigate influential factors on injury severity. It is worth noting that, among EMS times, transport time was not included in the model since the police-reported injury severity evaluations are often done on-scene without knowledge of transport and therefore transport time cannot impact the police-reported injury severity. To dig more into EMS times impact on crash outcome, the interaction effects of EMS times and injury location on the body were investigated in a separate model. Three sets of explanatory variables were considered in each model:

(1) crash-related variables

(2) crash-related variables + EMS times

(3) crash-related variables + EMS times + interaction effects of EMS times and injury location on the body

A limited number of studies are conducted to account for the role of EMS times on injury severity. These studies in the U.S. are scarce due to the fact that EMS data is not an inherent part of crash data. This study accounts for EMS times along with crash-related factors in estimating injury severity. Moreover, new variables are introduced, including patient level of distress and injury location on the body. Utilizing a linked dataset in this study explores the

[4] Sections from "Hosseinzadeh, A., & Kluger, R. (2021). Do EMS times associate with injury severity? Accident Analysis & Prevention, 153, 106053. Included in this chapter.

relationship between EMS times and injury severity among all injured individuals who were transferred to the hospital via EMS. The study also assesses the importance of EMS performance on injury severity in Motor Vehicle Crashes (MVC).

5.2.2 Data preparation

Police-reported crash data and EMS runs linkage in Jefferson County, Kentucky were used in this research. Table 5.1 shows the datasets and fields that were used for linkage. Furthermore, records which include missing KABCO injury severity, EMS runs that did not return to an emergency department, because either it was not serious or there was a death on the scene were, and no injury individuals (O level in KABCO scale) kept out of the further analysis. The 'O' crashes were excluded to make sure the data point actually required an emergency response and was not called out of precautionary measures. Figure 5.1 shows the linkage framework.

Table 5.1. Summary of datasets used for linkage purpose

Data Source	Number of records	Field name
Crash (Event Table)	21,358 records	Lat/Long Crash time Crash type Number of injuries Number of fatalities
Crash (Person Table)	80,786 records	Age Gender
EMS	5,473 records	Lat/Long Event Type Date Create Time Scene Time Transport time
PCR	4,242 records	Incident date Dispatch time Age Gender

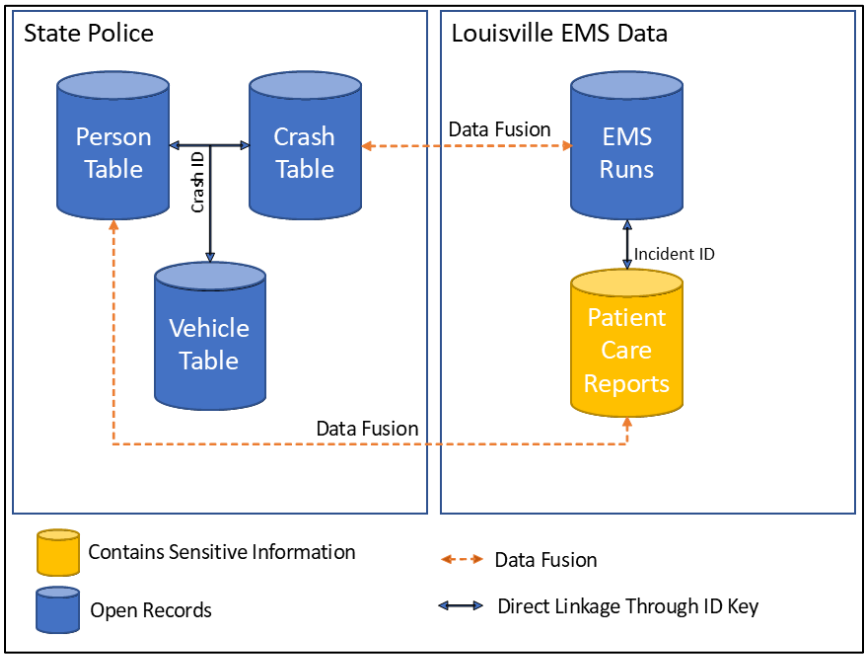


Figure 5.1. Visual framework of data linkage

The final linked data contained 1,572 unique MVC events and 2,192 unique people-crashes in Jefferson County, Kentucky, between July 2018 and March 2019. Attributes in the study are introduced in Table 5.2.

Table 5.2. The dependent and independent variables utilized in the model

Variable	Unit	Description	Levels/Interval	Frequency	Percentage
Dependent variable					
Injury severity		Injury severity based on KABCO scale	1.fatal and incapacitating (K& A)	139	6.3%
			2. non-incapacitating (B)	843	38.5%
			3. possible (C)	1210	55.2%
Independent variables					
Crash					
Age		Injured age	1. under 18	294	13.4
			2. 18-65	1722	78.6
			3.over 65	176	8.0
Gender		Injured gender	1.male	1023	46.7
			2.female	1169	53.3
Crash type		Crash type	1.angle	852	38.9
			2.head on	119	5.4
			3.opposing left turn	152	6.9
			4.rear end	490	22.4
			5.sidewipe	189	8.6

		6.single vehicle	390	17.8
Injury location code	Location of the injury on the body	1.head/face/ neck	735	33.5%
		2. chest/back/ abdomen/pelvis	462	21.1
		3.arms/hands/legs/feet	422	19.3%
		4.multiple-entire body	573	26.1%
Position in vehicle	Injured position in	1.front seat - left side	1394	63.6%
		2.front seat – right side	385	17.6%
		3.second seat – left side	110	5.0
		4.second seat – middle	51	2.3
		5.second seat – right side	114	5.2
		6.third seat	56	2.6
		7.none	82	3.7
Trapped	If the victim was trapped in the vehicle	1. not trapped	2078	94.8%
		2.trapped	114	5.2%
Ejection	If the victim was ejected from the vehicle	1.not ejected	2145	97.9%
		2. ejected	47	2.1%
Suspect of drinking	DUI test in a case that the injured was driver	1.no	1749	79.8%
		2.yes	443	20.2%
License restriction	License restriction in a case that the injured was driver	1.no	1660	75.7%
		2.yes	532	24.3%
Patient level of distress	Patient level of distress	1.none	490	22.4%
		2.mild	1146	52.3%
		3.moderate	373	17.0%
		4.severe	183	8.3%
Airbag	Airbag deployment status	1.air bag(s) deployed	997	45.5%
		2.no airbag(s) deployed	905	41.3%
		3.no airbag present	290	13.2%
Hwy	Highway	1.no	1891	86.3%
		2.yes	301	13.7%
Weather	Weather status	1.clear	1381	63.0%
		2.cloudy	432	19.7%
		3.rain/snow/fog	379	17.3%
Hit and run	Hit and Run	1.no	2047	93.4%
		2.yes	145	6.6%
Roadway character	Roadway characteristics	1.curve	237	10.8%
		2.straight	1955	89.2%
Light condition	Light condition	1.dark	184	8.4%
		2.dark – highway lighted	530	24.2%
		3.daylight	1324	60.4%
		4.dawn	68	3.1%
		5.dusk	86	3.9%
Time of the day	Crash time interval	1.early morning	336	15.3%
		2.morning peak	218	9.9%
		3.mid-day	636	29.0%
		4.evening peak	636	29.0%
		5. night	366	16.7%
Week time	Weekday/weekend	1.weekday	1539	70.2%
		2.weekend	653	29.8%
EMS				
Response time	EMS Response time in seconds	Base: RT < 240	323	14.7%
		240 < RT < 360	466	21.3%

		360 < RT < 480	475	21.7%
		480 < RT < 600	343	15.6%
		600 < RT < 900	414	18.9%
		RT > 900	171	7.8%
On-scene time	EMS on-scene time in seconds	Base: OT < 900	405	18.5%
		900 < OT < 1200	508	23.2%
		1200 < OT < 1500	486	22.2%
		1500 < OT < 2100	503	22.9%
		OT > 2100	290	13.2%

5.2.3 Method

Identifying factors that affect injury severity through various modeling frameworks has been covered well in the literature. A typical approach in these studies is to use a statistical modeling approach, with crash severity as a dependent variable and characteristics of the crash, driver, roadway, weather, etc. as independent variables (Mannering and Bhat 2014). A wide range of modeling approaches, including parametric and non-parametric, have been used in crash severity studies. In this study, a random effects ordered probit approach is utilized to model injury severity and study the impact of EMS response time. Assuming individuals involved in the same crash are expected to have comparable unobserved variables affecting their injury severity, a random-effect component was utilized in the model. Ignoring the similarity in each cluster (i.e., crash), known as intra class correlation (ICC), violates observations' independence assumption and leads into incorrect outcomes. Equation 5.1 shows the modeling formulation:

$$y_{ij}^{*} = X_{ij}\beta + v_{ij} + u_i \qquad (5.1)$$

Where X_{ij} is a (1×k) vector of observed explanatory variables of the *i*th individual in *j*th crash; β is a (k × 1) vector of coefficients for the explanatory variables, v_{ij} is the random-effects for individuals involved in the same crash *j* and u_{i} is individual-level random-effects. y_{ij}^{*} is a latent

variable. The observed variable with three injury severity levels of: (1) fatal/ incapacitating injury, (2) non-incapacitating injury and (3) possible injury can be written as equation 5.2.

$$y_{ij} = \begin{cases} 1 & if \quad y_{ij}^{*} \leq \mu_1 \\ 2 & if \quad \mu_1 < y_{ij}^{*} \leq \mu_2 \\ 3 & if \quad \mu_2 < y_{ij}^{*} \end{cases} \qquad (5.2)$$

Where the thresholds μ_1 and μ_2 are parameters to be jointly estimated with the vector of parameters β. ICC also is defined as portion of between cluster (i.e., crash) variance to total variance, as shown in equation 5.3.

$$\text{ICC} = \frac{Var\ (v_{ij})}{Var\ (y_{ij}^{*})} \qquad (5.3)$$

In this study, three models of (1) crash-related variables, (2) crash-related and EMS times and (3) crash-related, EMS times and interaction effects were estimated to identify the impact of factors on injury severity in MVC with a focus on EMS times. "Mixor" package (Archer et al., 2018) in R studio software was used in this study (R core team, 2019).

5.2.4 Results

Table 5.3 presents the outcome of each model. The first column shows model 1, consisting of only crash related variables. The second column shows the model that includes crash related variables in addition to EMS times. The third column includes the model with interaction effects. Significant variables are in bold font. According to Table 5.3, in all three models, age of the injured individual, trapped/ejected injuries of the crash in a motor vehicle, airbag status, weather, manner of collision and patients' levels of distress were found to be significantly associated with injury severity in a crash. Older occupants, trapped/ejected individuals, vehicles with the deployed airbag or without airbag available, cloudy/ foggy weather, single-

vehicle crashes, individuals with a position in the vehicle after the second-row seats, individuals with higher distress level were more likely to be more severe. Moreover, individuals with injury location of chest/back/abdomen/pelvis and arms/hands/legs/feet found to have less severe injuries than individuals with head/face/neck injuries.

In the second model, the EMS times were added to the model. Among EMS times, in the second model, none of the EMS times were found significant. In the third model, with interaction effects, injured individuals with entire body injuries and faster response time were associated with less severe injuries compared to the based level (response time < 4 minutes and injury location of head/face/ neck). Moreover, entire body injuries with low on-scene time were associated with more severe injuries compared to the base level. In arms/hands/legs/feet injuries, both very low and very high on-scene time were found to be significantly associated with more severe injuries. Light condition, time of the day and weekdays/weekends and the crash occurring on the highway are among the variables which were not found significant.

The intra class correlations were found as 0.575, 0.564 and 0.549 in first, second and third models, respectively, which imply the modeling approach was chosen properly. In terms of model fit, considering EMS times in model two (AIC = -1678) and EMS times + interactions in model three (AIC = -1684) improved the fit marginally compared to the model with only crash-related factors (AIC = -1696).

Table 5.3. Random effects ordered probit models

		Model (1) crash-related variables			Model (2) crash-related variables + EMS times			Model (3) crash-related variables + EMS times + interaction effects of EMS times and injury location on the body		
Variable	Levels	Coef	Std Err	P > z	Coef	Std Err	P > z	Coef	Std Err	P > z
Intercept		**3.37**	**0.35**	**0.001**	**3.37**	**0.39**	**0.001**	**3.29**	**0.46**	**0.001**
Age	Base: 18 < age < 65									
	Under 18	**0.49**	**0.16**	**0.003**	**0.48**	**0.08**	**0.004**	**0.53**	**0.17**	**0.002**
	Upper 65	-0.14	0.15	0.367	-0.13	0.15	0.397	-0.13	0.16	0.422
Gender	Base: Male									
	Female	-0.01	0.09	0.932	-0.01	0.09	0.940	0.01	0.09	0.979
Airbag	Base: No airbag deployed									
	Airbag deployed	**-0.58**	**0.11**	**0.001**	**-0.57**	**0.11**	**0.001**	**-0.60**	**0.12**	**0.001**
	No airbag present	**-1.05**	**0.16**	**0.001**	**-1.05**	**0.17**	**0.001**	**-1.02**	**0.09**	**0.001**
Highway	No									
	Yes	0.21	0.15	0.156	0.19	0.15	0.200	0.21	0.15	0.168
Weather	Clear									
	Cloudy	**-0.31**	**0.13**	**0.026**	**-0.31**	**0.13**	**0.024**	**-0.31**	**0.14**	**0.027**
	Fog	**-0.70**	**0.71**	**0.051**	**-0.76**	**0.72**	**0.066**	**-0.86**	**0.79**	**0.041**
	Raining	0.12	0.14	0.383	0.12	0.14	0.379	0.11	0.15	0.439
	Snowing	1.11	0.89	0.211	1.09	0.87	0.212	1.24	0.87	0.153
Crash type	Base: Angle									
	Head on	-0.18	0.25	0.465	-0.21	0.26	0.419	-0.22	0.27	0.414
	Opposing left turn	-0.09	0.22	0.660	-0.11	0.22	0.623	-0.15	0.23	0.491
	Rear end	0.05	0.14	0.721	0.04	0.14	0.781	0.04	0.15	0.753
	Sideswipe	-0.12	0.19	0.516	-0.16	0.19	0.413	-0.15	0.20	0.454
	Single vehicle	**-0.30**	**0.15**	**0.051**	**-0.30**	**0.15**	**0.053**	**-0.31**	**0.16**	**0.054**
Roadway characteristics	Base: Straight									
	Curve	0.14	0.16	0.371	0.11	0.16	0.473	0.10	0.17	0.525
Hit and Run	Base: False									
	True	-0.22	0.20	0.269	-0.21	0.20	0.296	-0.22	0.20	0.277
Light Condition	Base: Dark									
	Dark-highway lighted on	0.10	0.20	0.624	0.10	0.20	0.619	0.10	0.21	0.640
	Dawn	0.25	0.33	0.454	0.22	0.33	0.497	0.32	0.34	0.354
	Daylight	0.32	0.23	0.159	0.34	0.23	0.142	0.33	0.24	0.172
	Dusk	0.01	0.34	0.99	0.03	0.34	0.928	0.08	0.35	0.810
Position in Vehicle	Base: Front seat – left side									
	Front seat – right side	-0.01	0.14	0.94	-0.01	0.14	0.961	-0.03	0.15	0.822
	Second seat – left side	0.39	0.26	0.128	0.40	0.26	0.118	0.43	0.28	0.120
	Second seat – middle side	-0.20	0.27	0.467	-0.23	0.28	0.399	-0.29	0.28	0.307
	Second seat – right side	0.26	0.23	0.249	0.27	0.23	0.242	0.25	0.24	0.300
	Third seat	**0.70**	**0.35**	**0.044**	**0.70**	**0.35**	**0.044**	0.58	0.37	0.124
	After third seat	**0.57**	**0.28**	**0.044**	**0.17**	**0.28**	**0.035**	**0.55**	**0.29**	**0.063**
Trapped	Base: No									

	Yes	**-0.99**	**0.17**	**0.001**	**-0.99**	**0.17**	**0.001**	**-1.01**	**0.18**	**0.001**
Ejected	Base: No									
	Yes	**-1.09**	**0.35**	**0.001**	**-1.09**	**0.35**	**0.001**	**-1.17**	**0.36**	**0.001**
Driving under influence	Base: No									
	Yes	-0.20	0.14	0.162	-0.20	0.14	0.166	-0.18	0.15	0.220
License restriction	Base: No									
	Yes	-0.04	0.12	0.719	-0.04	0.12	0.714	-0.03	0.13	0.820
Time of day	Base: Mid-day									
	Early morning	-0.04	0.21	0.833	-0.02	0.21	0.905	-0.06	0.21	0.774
	Morning peak	-0.04	0.19	0.415	-0.15	0.19	0.420	-0.16	0.20	0.416
	Evening peak	-0.01	0.15	0.962	-0.01	0.15	0.943	-0.01	0.15	0.907
	Night	-0.15	0.22	0.482	-0.13	0.22	0.538	-0.15	0.22	0.496
Weekend	Base: No									
	Yes	-0.15	0.11	0.178	-0.15	0.11	0.199	-0.12	0.12	0.322
Patient level of distress	Base: None									
	Mild	-0.15	0.12	0.211	-0.15	0.12	0.213	-0.13	0.13	0.292
	Moderate	**-0.71**	**0.15**	**0.001**	**-0.69**	**0.15**	**0.001**	**-0.71**	**0.16**	**0.001**
	Severe	**-1.04**	**0.18**	**0.001**	**-1.04**	**0.19**	**0.001**	**-1.05**	**0.20**	**0.001**
Injury Location	Base: Head/face/neck									
	Chest/back/abdomen/pelvis	**0.68**	**0.12**	**0.001**	**0.69**	**0.12**	**0.040**	0.74	0.46	0.108
	Arms/hands/legs/feet	**0.29**	**0.13**	**0.026**	**0.30**	**0.13**	**0.023**	0.64	0.42	0.131
	Multiple-entire body	0.04	0.12	0.720	0.05	0.12	0.655	0.36	0.38	0.340
Response time (Seconds)	Base: RT < 240									
	240 < RT < 360				-0.03	0.17	0.837	-0.34	0.27	0.219
	360 < RT < 480				0.17	0.17	0.318	0.04	0.27	0.862
	480 < RT < 600				0.11	0.18	0.519	0.16	0.30	0.577
	600 < RT < 900				0.07	0.18	0.661	-0.08	0.29	0.771
	RT > 900				0.16	0.23	0.475	0.23	0.37	0.533
On-scene time (Seconds)	Base: OT < 900									
	900 < OT < 1200				-0.21	0.15	0.161	0.24	0.25	0.335
	1200 < OT < 1500				-0.12	0.15	0.436	0.18	0.26	0.487
	1500 < OT < 2100				-0.08	0.16	0.597	0.10	0.25	0.683
	OT > 2100				0.03	0.20	0.869	0.36	0.31	0.252
Response time	Injury location									
240 < RT < 360	Chest/back/abdomen/pelvis							0.23	0.44	0.599
360 < RT < 480	Chest/back/abdomen/pelvis							-0.27	0.44	0.542
480 < RT < 600	Chest/back/abdomen/pelvis							-0.27	0.45	0.542
600 < RT < 900	Chest/back/abdomen/pelvis							0.48	0.46	0.299
RT > 900	Chest/back/abdomen/pelvis							-0.23	0.56	0.677
240 < RT < 360	Arms/hands/legs/feet							0.25	0.42	0.544
360 < RT < 480	Arms/hands/legs/feet							0.53	0.41	0.202
480 < RT < 600	Arms/hands/legs/feet							0.04	0.51	0.923
600 < RT < 900	Arms/hands/legs/feet							0.49	0.43	0.263
RT > 900	Arms/hands/legs/feet							-0.19	0.63	0.756
240 < RT < 360	**Multiple-entire body**							**0.70**	**0.38**	**0.069**

360 < RT < 480	Multiple-entire body							0.34	0.41	0.397
480 < RT < 600	Multiple-entire body							-0.02	0.43	0.946
600 < RT < 900	Multiple-entire body							-0.20	0.42	0.637
RT > 900	Multiple-entire body							0.08	0.57	0.883
On scene time	Injury location									
900 < OT < 1200	Chest/back/abdomen/pelvis							-0.32	0.41	0.437
1200 < OT < 1500	Chest/back/abdomen/pelvis							0.03	0.41	0.933
1500 < OT < 2100	Chest/back/abdomen/pelvis							0.07	0.38	0.840
OT > 2100	Chest/back/abdomen/pelvis							0.11	0.50	0.822
900 < OT < 1200	**Arms/hands/legs/feet**							**-0.86**	**0.41**	**0.038**
1200 < OT < 1500	Arms/hands/legs/feet							-0.63	0.41	0.125
1500 < OT < 2100	Arms/hands/legs/feet							-0.60	0.41	0.145
OT > 2100	**Arms/hands/legs/feet**							**-0.86**	**0.51**	**0.073**
900 < OT < 1200	**Multiple-entire body**							**-0.46**	**0.24**	**0.063**
1200 < OT < 1500	**Multiple-entire body**							**-0.68**	**0.37**	**0.069**
1500 < OT < 2100	Multiple-entire body							-0.30	0.37	0.417
OT > 2100	Multiple-entire body							-0.62	0.47	0.193
μ_1		0.35	0.06	0.001	0.33	0.06	0.001	0.31	0.06	0.001
μ_2		1.80	0.06	0.001	1.79	0.06	0.001	1.80	0.06	0.001
ICC		0.575	0.007	0.001	0.564	0.005	0.001	0.549	0.005	0.001
LL		-1633			-1630			-1615		
AIC		-1678			-1684			-1696		

5.2.5 Discussion

Model Interpretation

While the association between EMS times and injury severity is investigated in this study, it is important to ensure the outcome is not under the influence of simultaneity. Simultaneity in this study would occur when EMS times impact injury severity while injury severity simultaneously influences EMS times. To alleviate the impact of simultaneity first, the no-injury crashes had to be excluded from the analysis. No-injury crashes seem to impact the EMS times since the lack of urgency likely results in the first responders taking their time. In those cases, it is possible one of the individuals still requests transfer to the hospital. In the injury cases, it seems the first responders behave in a relatively consistent manner following policy. Among 2480 injured individuals' data, 288 O injuries were excluded; therefore, the data reduced to 2192 injuries with (1) fatal and incapacitating (K&A) injuries (2) non-incapacitating (B) injuries, and (3) possible (C) injuries. By doing this we believe we have eliminated at least the worst cases of injury dictating EMS times. However, to ensure reverse causality was not extensively impacting our model, further diagnostics were completed. Considering equation 1 as the modeling formulation, in the presence of simultaneity effects, equation 5.4 should also be true.

$$X_{ij} = \beta y_{ij}^{*} + v_{ij} + u_i \quad (5.4)$$

Based on the reverse causation formula, if X_{ij} found not to be correlated with the error term, it could be claimed that the reserve causation is not a valid argument (Katz, 2006).

We explored this argument's validity by comparing the distribution of correct and wrong classifications in different EMS times intervals (Table 5.4). The Chi-square analysis result in

Table 5.4 indicates that the proportion of the wrong prediction to correct predictions by different response time intervals was not significant. Therefore, response time is not apparently correlated with the residual error. For on-scene time, the simultaneous effect was also not captured (Table 5.5) in a Chi-Square analysis.

Table 5.4. The prediction distribution based on response time

	RT < 240	240 < RT < 360	360 < RT < 480	480 < RT < 600	600 < RT < 900	RT > 900
Correct prediction	270	395	394	288	362	146
Wrong prediction	53	71	81	55	52	25
(Wrong/ correct) pct.	19.6%	17.9%	20.5%	19.1%	14.3%	17.1%
Chi-square: 0.547						

Table 5.5. The prediction distribution based on on-scene time

	OT < 900	900 < OT < 1200	1200 < OT < 1500	1500 < OT < 2100	OT > 2100
Correct prediction	314	413	295	402	431
Wrong prediction	56	80	50	74	77
(Wrong/ correct) pct.	17.8%	25.4%	15.9%	23.5%	24.5%
Chi-square: 0.970					

Model Outcome

The results showed, considering the age of those injured, younger individuals saw less severe injuries. Moreover, cloudy and foggy weathers were associated with more severe crashes in Jefferson County, Kentucky. In terms of weather impacts on injury severity, there is not a consensus among researchers. The results were interpreted in different ways; resulting in losing control of the vehicle and ending up more severe crashes (Eluru et al., 2008, Yu and Abdel-Aty 2014, Haleem et al., 2015) or making the drivers more cautious and resulting in a less severe crash (Naik et al., 2016).

Ejected and trapped individuals sustained more severe injuries. Individuals' position after the second-row seats in the vehicles with more than two-row seats involved in the crashes

saw less severe injuries than those in the front two rows. Moreover, single-vehicle crashes (e.g., pedestrian-/cycle- involved, rollover crashes) were found to be related to more severe injuries compared to the base case of angle crashes. The higher severity of single-vehicle crashes has been found in other recent literature (Hosseinzadeh et al., 2021b).

Results show individuals with higher distress levels were associated with more severe injuries. It seems individuals have the right perception about their level of injuries, or officers may be basing injury designation on distress level. Furthermore, crashes in which either the airbag deployed, or an airbag was not available were associated with a more severe injury outcome. The results are expected since the airbag status shows either it was that serious enough for the airbag to deploy, or it was not available in the first place.

Injury locations of chest/back/abdomen/pelvis and arms/hands/legs/feet were associated with less severe injuries than the base level of head/face/neck in the first and second models. However, entire body injuries were not found significantly different than head/face/neck injuries. The results indicate either chest/back/abdomen/pelvis and arms/hands/legs/feet led to less severe injuries than head/face/neck or chest/back/abdomen/pelvis and arms/hands/legs/feet injuries led officers to evaluate these injuries less severe than head/face/neck injuries.

While response time was not significant in the second model, the third model found that it was important interacting with multiple body injuries associated with less severe injuries. The second model results are in line with Lovely et al. (2018), who found no significant relationship between increasing response time and more severe injuries in general injuries (Lovely et al., 2018). This research, in the third model, found a significant relationship between reducing response time and decreasing the severity of injuries. This finding highlights the

importance of fast response in cases with entire body injuries. It is worth noting that the urban and suburban study areas in Jefferson County led to relatively fast EMS response with low variability. The authors believe that a rural setting with sparser EMS coverage and hospital density may see different results, with response time and scene time having a larger impact.

The outcome indicated on-scene time was not significant, according to the second model results. However, the third model results showed individuals with multiple parts of body injuries, and low scene time was associated with more severe injuries compared to the base case (On-scene time < 15 minutes and head/face/neck injuries). This finding highlights the importance of injury assessment on-scene by Emergency medical technicians (EMT) and taking adequate precautionary acts to stabilize individuals' status on-scene with entire body injuries. Moreover, arms/hands/legs/feet injuries with either very low or high on-scene time were related to more severe injuries.

5.2.6 Conclusion

Although the importance of optimized and efficient EMS in saving lives is undisputable, there is not a consensus on how EMS times impact injured individuals of the crashes. This study took into account EMS times along with other crash related variables to explore the impact on injury severity. Based on the outcome, although the authors did not find a significant relationship between EMS times and injury severity in all types of injuries, EMS times based on injured body locations shed light on the relationship between EMS times and injury severity. The outcome showed faster response time was associated with less severe injuries in cases with an entire body injury. Accounting for on-scene time, the results indicated that either very low on-scene times or very high on-scene times were related to more severe injuries in entire body parts and arms/hands/legs/feet injuries. Adding EMS times and interaction

effects of EMS times and injury location on the body to the model, improved the model quality marginally.

This study also has some limitations. First, the outcome showed some EMS-related factors were correlated with crash injury severity; However, the relationship does not imply causation. For instance, although outcome indicated the higher response time in cases with multiple body location injuries were associated with higher injury severity, it does not mean higher response time cause severe injuries. Response time may have contributed to the injury severity, injury severity may have led to the response time or there could be a latent factor correlated with both EMS response time and injury severity. Second, fatal and incapacitating injuries were merged into one category due to low numbers of fatalities in the data. There is a possibility that faster response time increases the chance of survival that the current research was not able to capture. In the current dataset, among fatal injuries, three individuals with average EMS response time of approximately 11 minutes died at the scene, while the same measure for 13 individuals who died at the hospital was about eight minutes. Further, 126 individuals with incapacitating injuries who survived had also about eight minutes response times. However, the insufficient records impede statistical investigation of those impacts. This could be further researched employing larger sample size. Third, the area of research was in an urban and suburban area, which resulted in low variation in EMS response time. Further research could elaborate on more diverse geographical area. The findings of this study could act as a base for further investigation of EMS impact on injury severity, particularly with respect to effective use of EMS times in evaluation of service quality.

5.3. Exploring Influencing Factors on Crash-related Emergency Response Time[5]

5.3.1 Data preparation

This section utilized linked police-reported crash data and EMS data from Jefferson, County KY (Figure 5.1). Jefferson County is a largely urban and suburban county in western KY where Louisville is located. The final linked data contained 2,009 unique MVC events and 2,977 unique people-crashes in Jefferson County, Kentucky, between July 2018 and March 2019. Table 5.6 represents the variables utilized in the study.

Table 5.6. Variables utilized in the model

Variable	Description	Levels/Interval	Frequency/average	Pct/sd
		Dependent variable		
EMS Response time (second)	EMS Response time	[36 – 2689]	495.73	293.47
		Independent variables		
		Demographics		
Age	Injured age	[0 – 95]	35.44	19.33
Gender	Injured gender	1.male	1367	46%
		2.female	1610	54%
Race	Injured race	1.white	1391	46%
		2.african-american	1358	46%
		3.hispanic/latino	111	4%
		4.others	117	4%
Weight (lbs)	Injured weight	[10 -475]	174.47	42.75
Pregnant	Injured pregnancy status	1.no	2871	96%
		2.yes	106	4%
	Event time			
Time of the day	Crash time interval	1.early morning	421	14%
		2.morning peak	859	29%
		3.mid-day	855	29%
		4.evening peak	302	10%
		5. night	540	18%
Crash Hour	Crash time	[0 – 23]	14.19	5.55
Week time	Week time of the crash	1.weekday	2125	71%
		2.weekend	852	29%

[5] Sections from "Hosseinzadeh, A., Haghani, M., & Kluger, R. (2021a). Exploring Influencing Factors on Crash-related Emergency Response Time: A Machine Learning Approach (No. TRBAM-21-00614)." Included in this section.

Police/EMS time discrepancy (minute)	Time difference between events record in police-reported data and EMS CAD data	[0 – 59]	5.18	9.17
Police/EMS location discrepancy (meter)	Police/EMS location discrepancy between events record in police-reported data and EMS CAD data	[0 – 999]	141.48	240.11
		EMS		
EMS travel distance (mile)	Since mileage from EMS center to event location	[0.1,124]	8.34	15.55
Disposition	How the injuries are transported	1.evaluated/treated on scene	752	25%
		2.transported light/siren	408	14%
		3.transported no light/siren	1812	60%
		4. dead on scene	5	1%
Requested by	Who requested for ambulance	1.by stander	1274	43%
		2.family	185	6%
		3.fire department	111	4%
		4.law enforcement	305	10%
		5.patient	904	30%
		6.other	198	7%
Patient level of distress	Patient level of distress	1.none	858	29%
		2.mild	1503	50%
		3.moderate	402	14%
		4.severe	214	7%
Extrication required	If extraction is required	1.no	2828	95%
		2.yes	149	5%
Estimated speed (Mph)	Estimated speed of the vehicle at the crash	[0,130]	27.93	14.85
Airbag	Airbag deployment status	1.air bag(s) deployed	1248	42%
		2.no air bag(s) deployed	1334	45%
		3.no airbag present	395	13%
Vehicle type	Type of vehicle of the injured person	1.automobile	2490	84%
		2.tractor-trailer	34	1%
		3.motorcycle	137	5%
		4.moped	125	4%
		5.others	191	6%
Event type	Event type	1.mvc- pedestrian	104	3%
		2.mvc-bicycle	21	1%
		3.mvc-motorcycle	43	1%
		4.mvc-injury	2465	83%
		5.mvc-ejected/fire	44	1%
		6.mvc-rescue	221	7%
		7.mvc-rollover	79	3%
EMS priority	EMS priority	1.high priority	506	17%
		2.low priority	2471	83%
Agency	Agency which responded to the request	1.louisivlle metro EMS	2914	98%
		2.fire department	63	2%
Hwy	Was the crash location on highways?	1.no	2553	86%
		2.yes	424	14%

Motor vehicles involved	Number of motor vehicles involved	1.single vehicle	503	17%
		2.two vehicles involved	2010	67%
		3.more than two vehicles involved	464	16%
Number of fatalities	Number of fatalities of the crash	1.non-fatal crash	2953	99%
		2.fatal crash	24	1%
Number of injured	Number of injuries of the crash	1.non-injury crash	446	15%
		2. 1 injury	1247	42%
		3. 2 injuries	699	23%
		4. 3+ injuries	585	20%
Weather	Weather status	1.clear	1906	64%
		2.cloudy	570	19%
		3.raining	471	16%
		4.snowing	18	1%
		5.fog	12	1%
Roadway condition	Roadway condition	1.dry	2278	77%
		2.wet /snow/flood/ice	699	23%
Hit and run	Was the event because of hit and run?	1.no	2766	93%
		2.yes	211	7%
Roadway character	Roadway characteristics	1.curve and grade	74	2%
		2.curve and hillcrest	35	1%
		3.curve and level	233	8%
		4.straight and grade	122	4%
		5.straight and hillcrest	65	2%
		6.straight and level	2448	82%
Light condition	Light condition	1.dark	249	8%
		2.dark – highway lighted	721	24%
		3.daylight	1808	61%
		4.dawn	89	3%
		5.dusk	110	4%
Was transported	If the EMS transported the injuries to hospital	1.no	863	29%
		2.yes	2114	71%
Crash				
Crash type	Crash type	1.angle	1169	39%
		2.head on	154	5%
		3.opposing left turn	214	7%
		4.rear end	661	22%
		5.rear to rear	42	1%
		6.sidewipe-opposing direction	49	2%
		7.sidewipe-same direction	200	7%
		8.single vehicle	488	16%
Injury severity	Injury severity based on KABCO scale	1.fatal (K)	142	5%
		2.incapacitating (A)	871	29%
		3.non-incapacitaing (B)	1261	42%
		4.possible (C)	703	24%
Injury location	Location of the injury based on different areas of body	1.head/face	546	18%
		2.neck	345	12%
		3.chest	279	9%
		4.back	329	11%
		5.abdomen/pelvis	204	7%
		6.arms/hands	302	10%
		7.legs/feet	334	11%
		8.multiple-entire body	638	21%

Position in vehicle	Injured position in	1.front seat - left side	1930	65%
		2.front seat – right side	511	17%
		3.second seat – left side	150	5%
		4.second seat – middle	74	2%
		5.second seat – right side	182	6%
		6.third seat	52	2%
		7.none	78	3%
Trapped code	How the injured trapped	1. not trapped	2855	96%
		2.trapped	114	4%
Ejection code	How the injured ejected due to the crash	1.not ejected	2928	98%
		2. ejected	49	2%
Suspect of drinking	DUI test in a case that the injured was driver	1.no	2462	83%
		2.yes	515	17%
License restriction	License restriction if the injured individuals was driver	1.no	2334	78%
		2.yes	643	22%

5.3.2 Methodology

In this research, a parametric approach, as well as four non-parametric approaches, were implemented. The comparison was conducted to choose the best model and find the most influential factors on EMS response time based on the results of the best model. Four tree-based ensemble learning approaches in two different categories were utilized: bagging and boosting. The former primarily focuses on understanding the variance, while boosting minimizes errors in prediction. For each category, two models with different functional forms were used to reach better results.

Sampling Approach

Bagging

Bagging is an ensemble approach to reduce the variance of an estimate by averaging multiple estimates together (Breiman, 1996). Figure 5.2 depicts the process of bagging. Two bagging-based methods (bagged tree and random forest) have been used in this study.

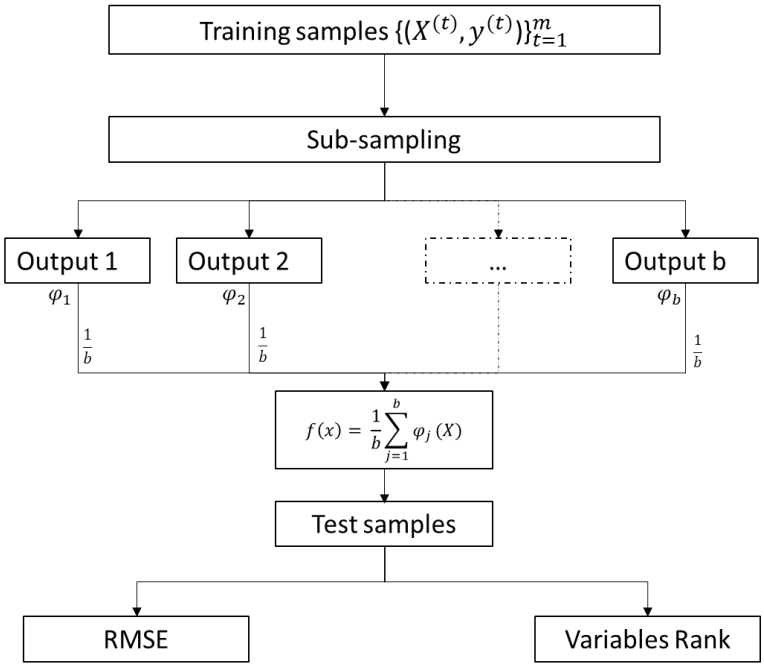


Figure 5.2. The process of bagging

Bootstrap Aggregating (Bagged tree) Bagged tree is conducted based on creating many random sub-samples of a dataset with replacement. The basic motivation of implementing bagged tree is to combine the predictions of several base learners to create more accurate output. This method is a procedure that can be utilized for prediction as well as ranking the variable importance. Bagged tree only has one parameter, which is the number of trees to include. Tuning the model is based on finding the optimal number of trees that minimize Out-Of-Bag (OOB) error. OOB is the part of the data that is not taken for each bagged sample. The performance of each model on its unsampled source when averaged can provide an estimated accuracy of the bagged models (Breiman, 1996). In this study, bagged tree was implemented using the "ipred" package in R Studio software (Team, 2015).

Random Forests (RF) The RF approach generates multiple decision trees in parallel. Every single tree draws a random sample from the primary dataset when generating its splits. In order to prevent overfitting, a further element of randomness is added. The main principle of parallel methods is to exploit independence between the basic decision trees since the error can be decreased significantly by averaging. Implementing this method requires determination of the models' hyperparameters, consisting of the number of trees to grow and set of variables randomly sampled to choose at each split. Hyperparameters were optimized to ensure the results are not strongly dependent on any individual feature and all potential predictive features are involved in the model. The basic difference between two introduced bagging approaches is the fact that in RF, only a portion of total features are randomly selected, and the best split feature from the subset is used to split each node in a tree; whereas in the bagged tree, all features are considered for splitting a node (Breiman, 2001). In this study, RF was implemented using the "randomForest" package in R Studio software (Team, 2015).

Boosting Method

Boosting is a family of methods that are able to transform weak learners into strong learners. The concept behind boosting is to fit a sequence of weak learners to weighted versions of the data. More weight is given to records that were misclassified by earlier iterations. To produce the final prediction, the predictions are combined through a weighted sum. The main difference between boosting and bagging is the fact that base learners are trained as a result of frequent iterations on a weighted form of the data (T. Chen & Guestrin, 2016; Schapire, 2003). Figure 5.3 shows the process of boosting.

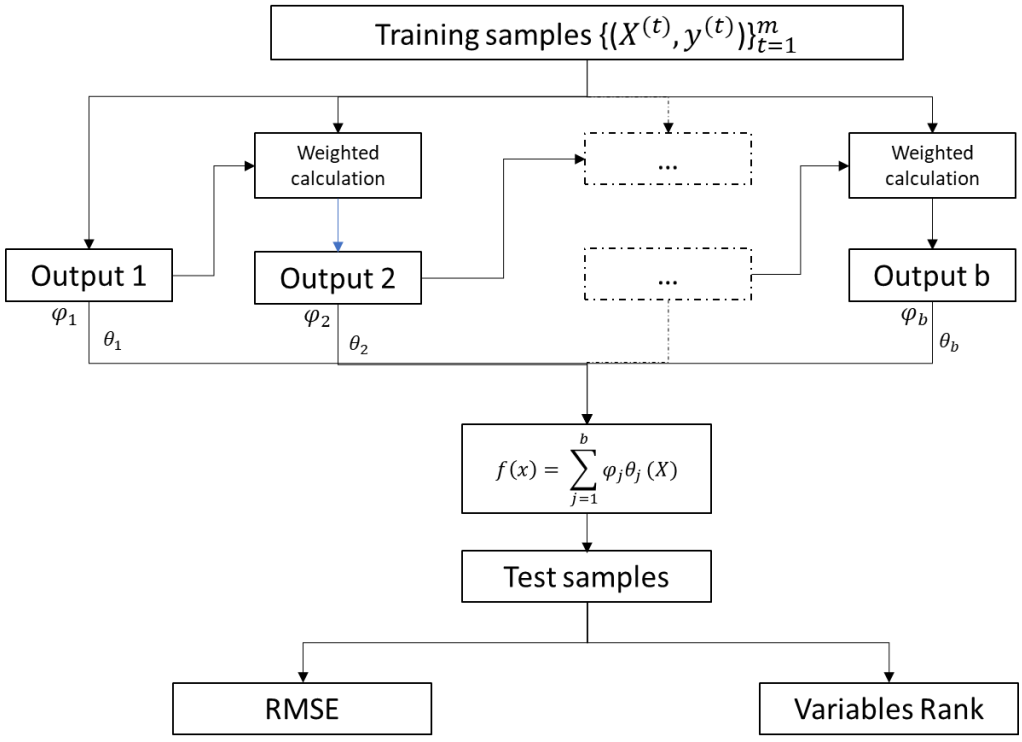


Figure 5.3. The process of boosting

Gradient Boosting Machine (GBM) GBM is a generalization of boosting that implements an additive weighting scheme to improve the prediction performance. Consider data defined as $\{(x_i, y_i)\}_{i=1}^{n}$ and a differentiable loss function $l(y_i, F(x))$ in which x_i are explanatory variables related to each EMS run, y_i is the associated response time of that run and i refers to the EMS run. The model initializes with a constant value (equation 5.5).

$$F_0(x) = argmin \sum_{i=1}^{n} l(y_i, \gamma) \qquad (5.5)$$

Where y_i refers to the observed value (EMS response time here) and γ represents the predicted value (equation 5.6).

$$For\ m = 1\ to\ M: r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad i = 1, \ldots, N \qquad (5.6)$$

Where m refers to individual trees, M is the number of trees and r_{im} represents to residual in EMS run i for tree m. The next stage is generating a regression tree to the r_{im} values and creating terminal regions $\{R_{jm}\}_1^J$ in which j is the index of each leaf in a tree. Furthermore, the value of the leaf nodes in the regression tree are estimated (equation 5.7).

$$For\ j\ = 1\ to\ j_m{:}\gamma_{jm}\ =\ argmin\ \sum_{x_i \in R_{jm}} l(y_i, F_{m-1}(x_i)\ +\ \gamma) \qquad (5.7)$$

According to equation 3, the final output is updated as it is shown in equation 5.8.

$$F_m(x)\ =\ F_{m-1}(x)\ +\ \vartheta \sum_{j=1}^{J_m} \gamma_{jm}\ (x \in\ R_{jm}) \qquad (5.8)$$

Where ϑ is the learning rate. A small learning rate reduces the effect each step has on the final prediction and this improves the accuracy in the long run. $F_M(x)$ is estimated iteratively (*30*). In this study, GBM was implemented using the "GBM" package in R Studio software (*26*).

XGBoost XGBoost is an ensemble method that is built upon iteratively growing weak learners (i.e., low-depth decision trees) to predict the dependent variable $\hat{y}_i$ based on K additive functions. Given a dataset with n EMS runs and independent variables x_i with m features ($x_i \in\ \mathbb{R}^m$) and their corresponding dependent variable y_i ($y_i \in \mathbb{R}$).

$$\hat{y}_i\ =\ \sum_{k=1}^{K} f_k(x_i)\,, f_k \in\ F \qquad (5.9)$$

Where f_k is an independent tree structure with leaf scores in the space of trees (F). The final prediction is equal to summing up the score in the corresponding leaves. The goal is to minimize objective function (equation 5.10) at iteration t:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (5.10)$$

$$\text{Where } \Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega_i\|^2$$

Where l is a differentiable convex loss function that measures residuals, T is the number of terminal leaves in a tree, γ is a user-definable penalty meant to encourage pruning. $\frac{1}{2}\lambda\|\omega_i\|^2$ is known as a regularization term, which helps to smooth the learning process to avoid overfitting. The main difference between GBM and XGBoost is in regularization term. Using second-order Taylor approximation at step t and simplified objective function results in equation 5.11.

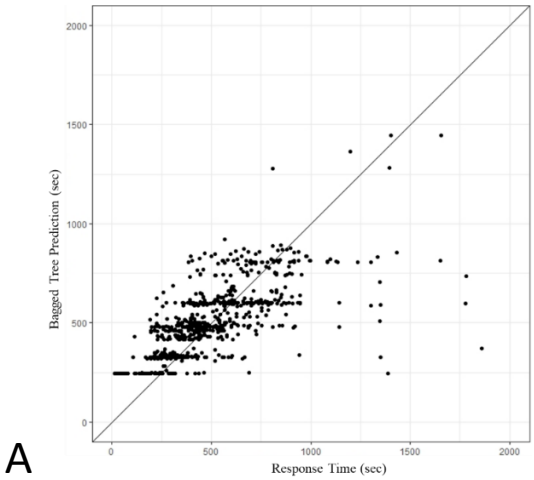
$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (5.11)$$

In simple terms, every single low-depth decision tree is generated to minimize a loss function. In each stage, the estimation allocates more weight to the runs that were incorrectly predicted by preceding trees. The ultimate model outcome is collectively determined by the results of all the developed trees (T. Chen & Guestrin, 2016). In this study, XGBoost was implemented utilizing the "XGboost" package in R Studio software (Team, 2015).

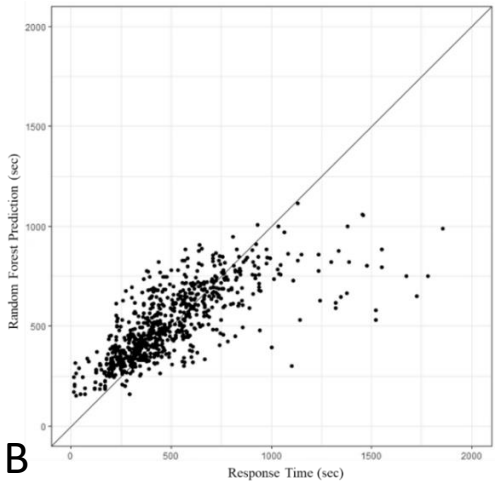
5.3.3 Results

In this section, the results of implementing the models are presented. Bagged tree, RF, GBM and XGBoost, as well as a regression method, were implemented and compared to predict EMS response time. Comparisons were conducted to find the most successful approach. Figure 5.4 shows the comparison between the EMS response time and the

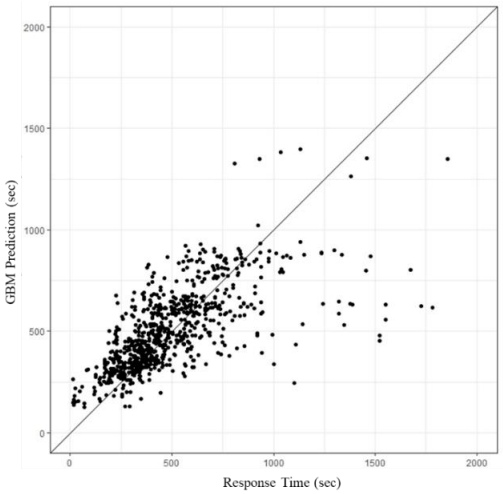
predicted response times using bagged tree, RF, GBM and XGBoost on the test set to evaluate the performance of trained models.



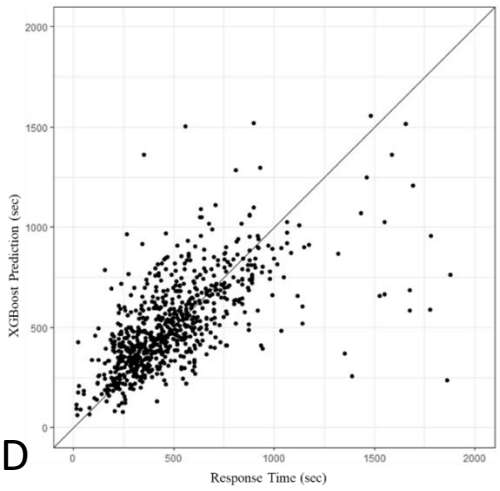
A



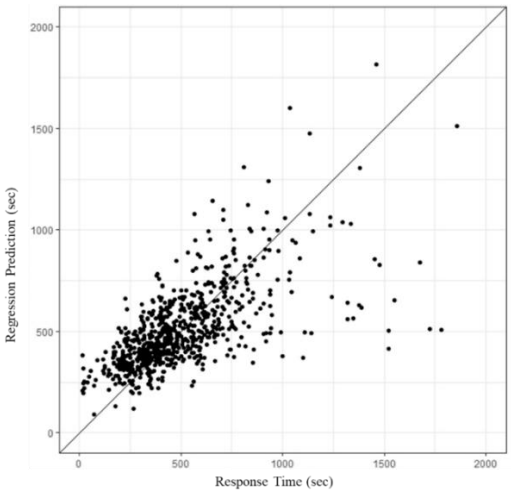
B



C



D



E

Figure 5.4. EMS response time vs predicted response time (A) Bagged tree (B) RF (C) GBM (D) XGBoost (E) Linear Regression

To compare the methods, Root Mean Square Error (RMSE), R-squared, adjusted R-squared and AIC were utilized as criteria. After running the model and assessing the predicted response times for each record in the test set, RMSE assessed them based on the difference between the results of the model response times and actual response times (i.e., residuals) for all models. Furthermore, R-squared, adjusted R-squared and AIC were calculated using the residual sum of squares, total sum of squares, number of independent variables and number of test set records.

Table 5.7. Comparison between the machine learning models

Models	Bagged Tree	RF	GBM	XGBoost	Linear Regression
RMSE	207.92	**204.00**	217.43	238.62	222.12
R-Sq	44.19	**50.72**	44.27	34.51	43.95
Adj R-Sq	41.14	**48.03**	41.22	30.93	41.20
AIC	3601.85	**3589.26**	3629.94	3692.88	3645.52

According to Table 5.7, RF was superior to other approaches in describing EMS response time. Therefore, RF results were used for further assessment of influential factors on EMS response time. Figure 5.5 shows the ranks of important factors in describing response time.

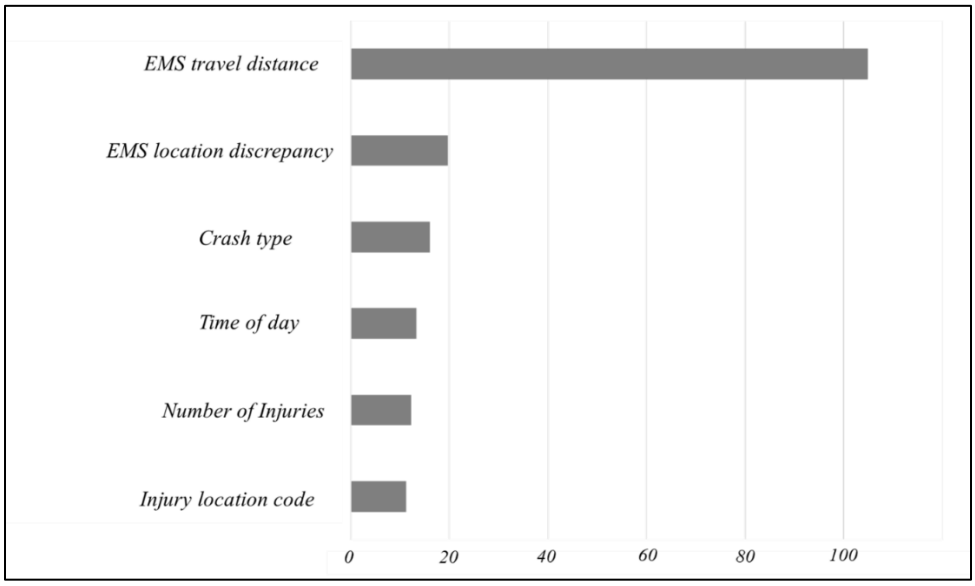


Figure 5.5. Most influential factors on EMS response time (RF results)

5.3.4 Discussion

The distance between the dispatch center and the event location was expected to be the most influential factor, which the findings of this study confirmed. Figure 5.6A shows the direct relationship between EMS travel distance and response time. Police/EMS location discrepancy was found as the second most influential factor. Increasing the distance between the locations resulted in increasing the EMS response time (Figure 5.6B). The location of the event is reported by the caller who could be a bystander, an individual involved in the crash, or authorities near the scene. Since police officers file the report on the crash scene, the authors believe crash locations on police-reported data are more likely to be accurate than EMS CAD data, which is the location reported by the caller. The results indicate reporting inaccurate locations could cause confusion for dispatchers and hinder first responders from providing aid to the injuries as soon as possible.

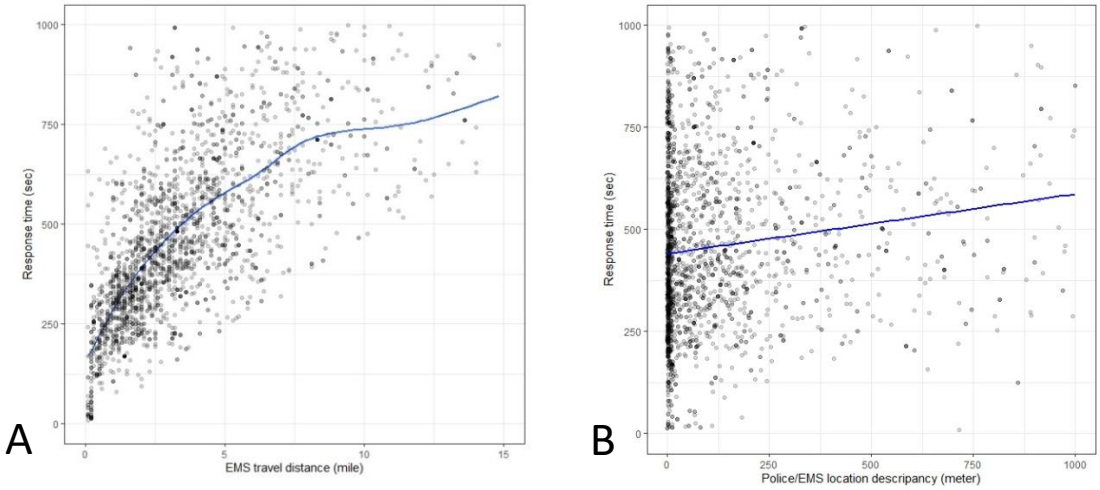
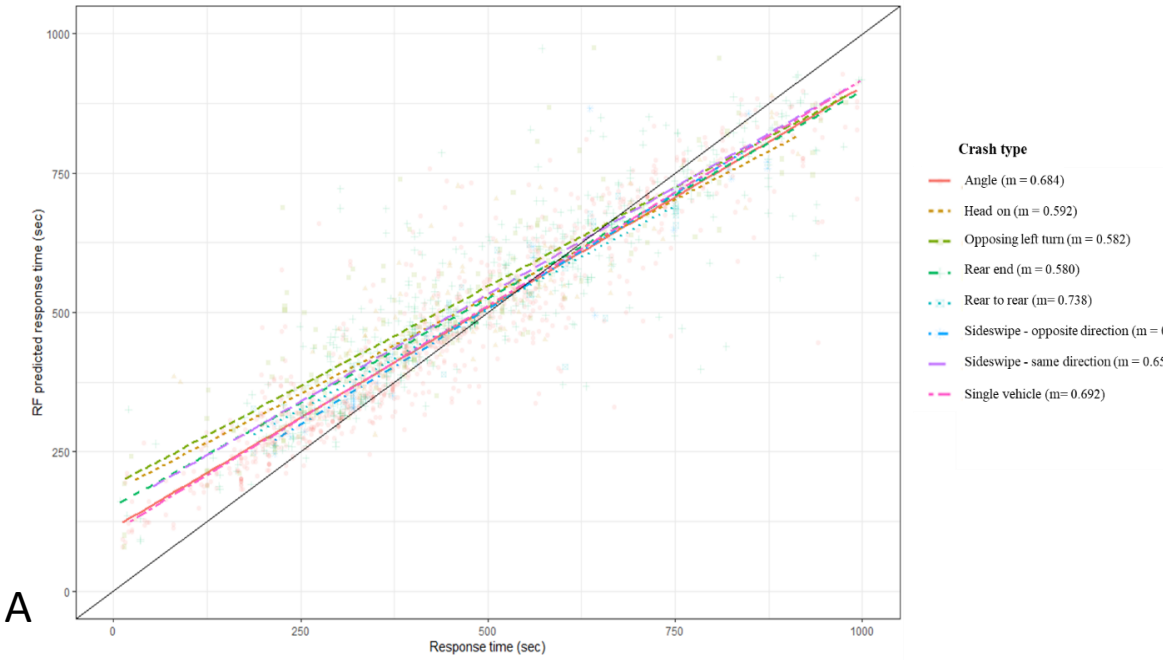


Figure 5.6. The relationship between response time and (A) EMS travel distance (B) Police/EMS location discrepancy



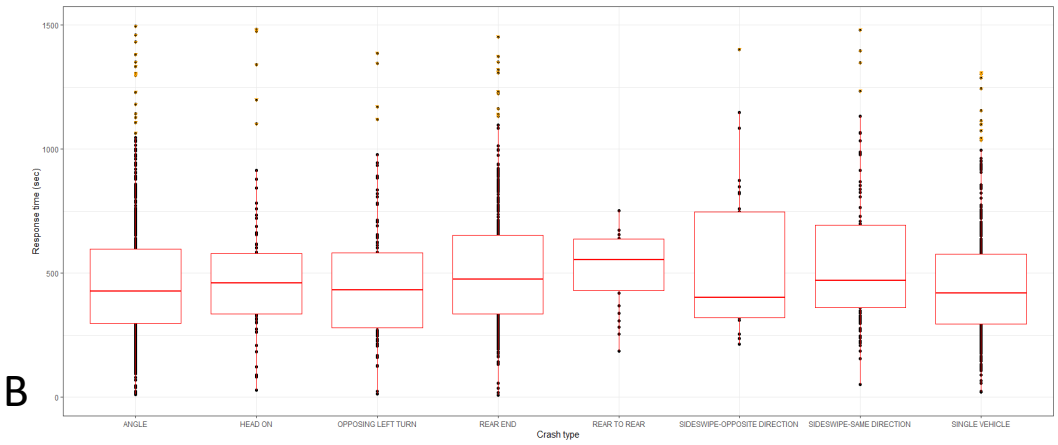


Figure 5.7. (A) Response time vs. RF predicted response time for different crash types (B) response time box and whisker plot in different crash types

The third most important variable was crash type in the RF model. According to Figure 5.7(A), it is discernible that in some crash types the model performed better; for instance, crash types of single vehicle, sideswipe – opposite direction and rear to rear were more successful in predicting response time compared to opposing left turn and rear-end. Moreover, according to Figure 5.7(B), crash types of single vehicle and sideswipe – opposite direction showed faster response times compared to rear-end and sideswipe – same direction.

Based on the results of Figure 5.8(B), as expected, the morning and evening peaks show higher response times. Figure 5.8(A) indicates that early morning was less successful than other times of day in predicting response time, probably because early mornings have more uncertainty in terms of fleet management and are harder to predict.

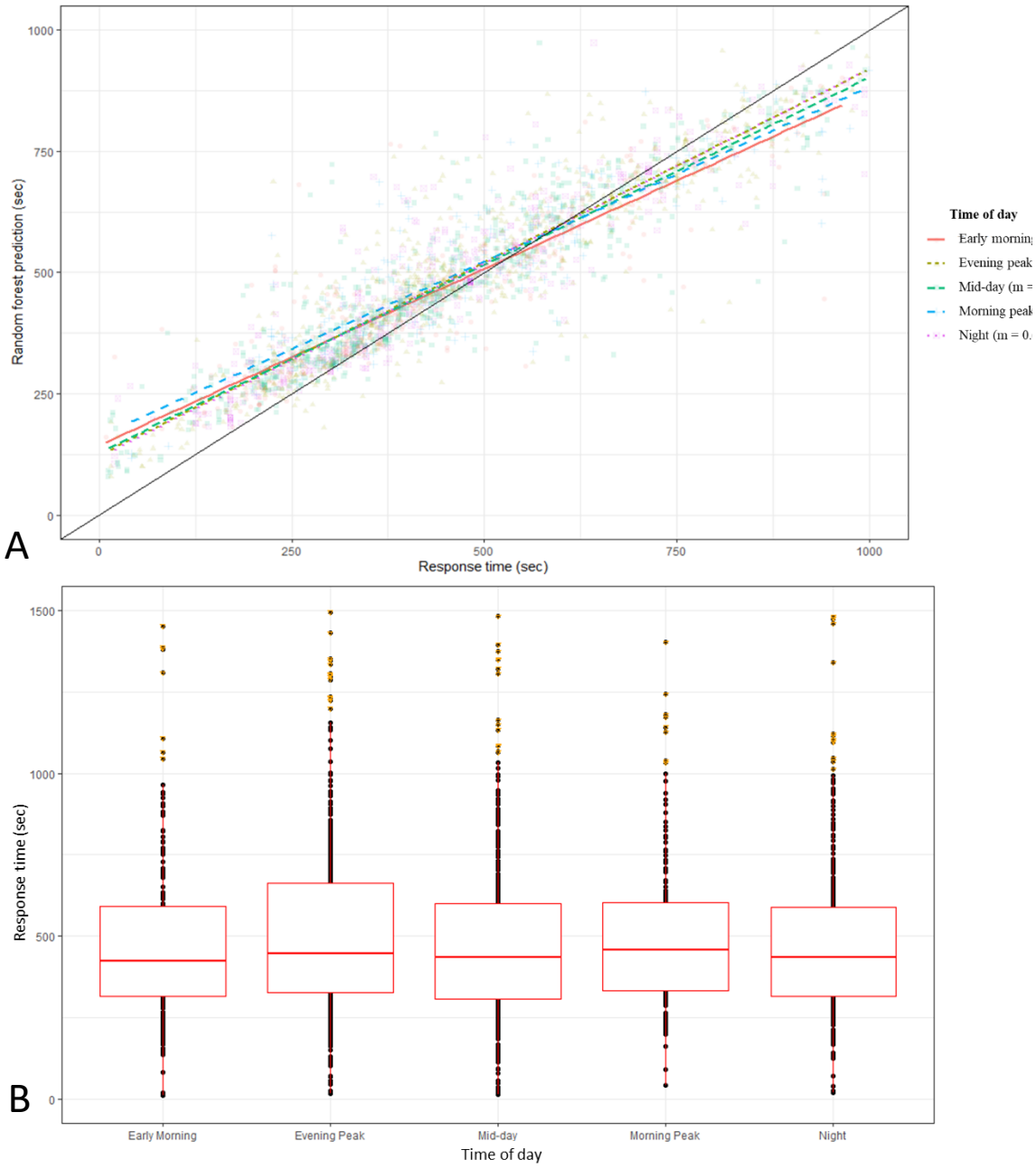


Figure 5.8. (A) Response time vs. RF predicted response time for different time of days (B) response time box and whisker plot in different time of days

On average, the response time reduced slightly as the number of injuries involved in a crash increased, shown in Figure 5.9(B). From a model fit perspective, Figure 5.9(A) shows that there is not a discernable difference in predicting EMS response time based on the number of injuries in a crash.

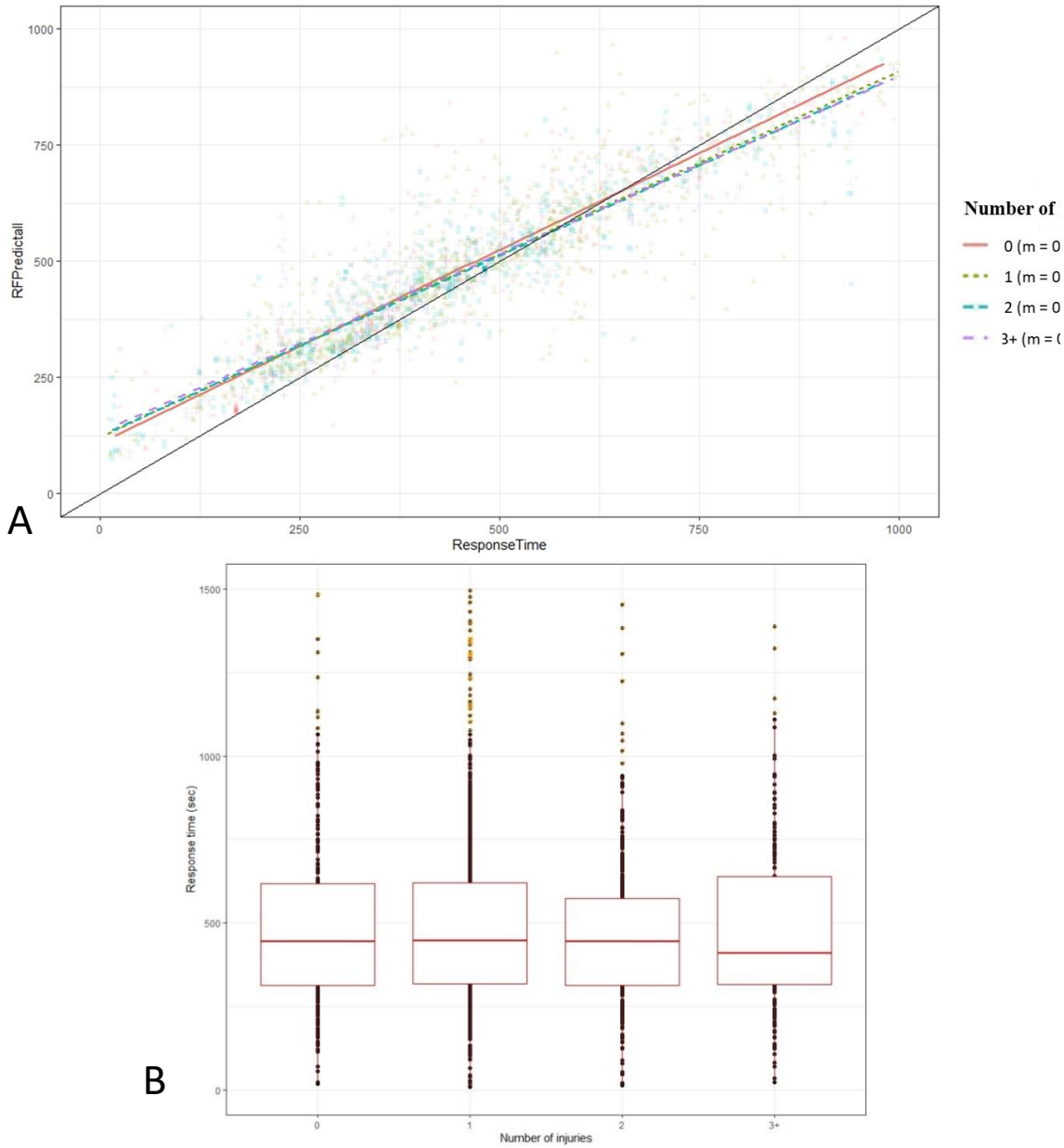


Figure 5.9. (A) Response time vs. RF predicted response time for different number of injuries (B) response time box and whisker plot in different number of injuries

The sixth variable was injury location. Injury location in police-reported crash data provide information about the part of the body that got injured in the crash. Figure 5.10 (B) shows the lower response time to injuries involving the legs and feet or the back. It seems that more apparent injuries resulted in faster response times. Any discernible differences are not recognized among the prediction of different injury locations (Figure 5.10 (A)).

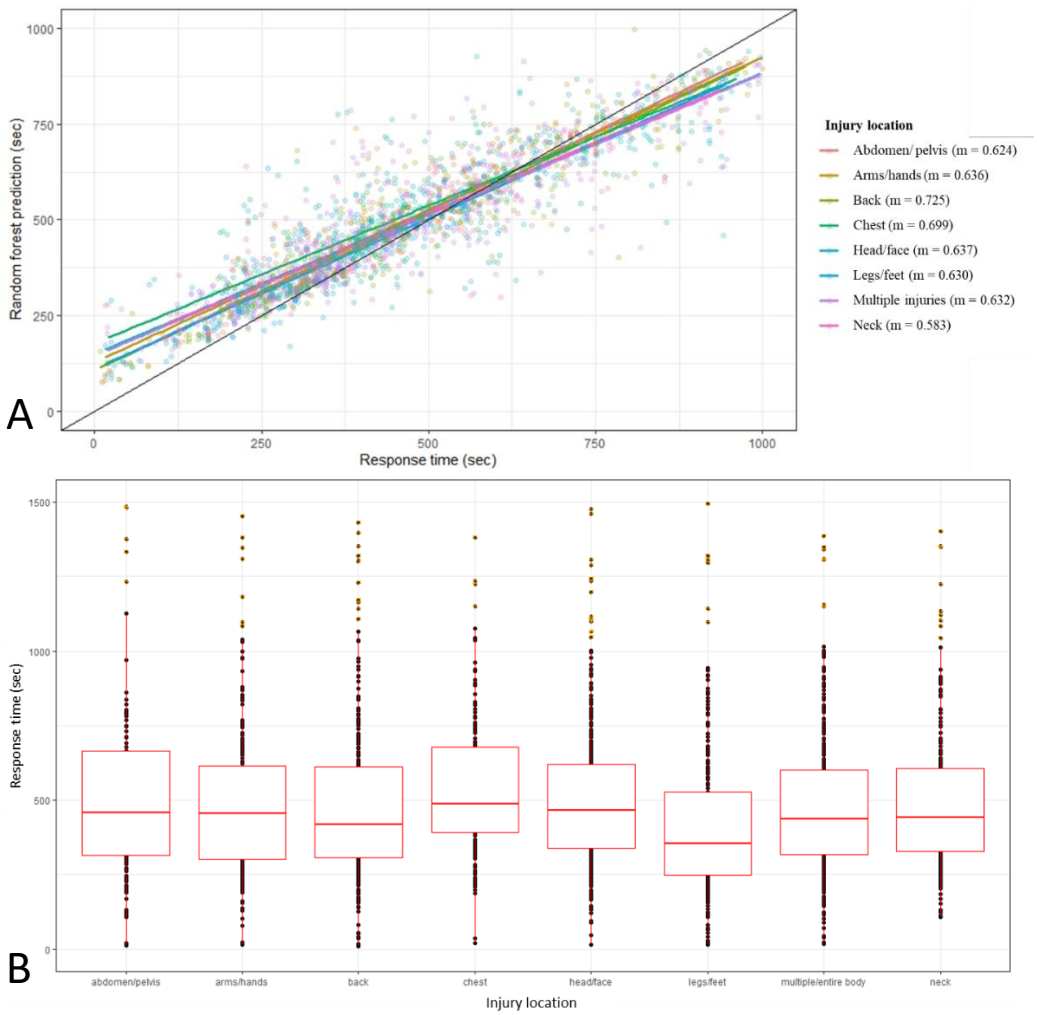


Figure 5.10. (A) Response time vs. RF predicted response time for different injury location (B) response time box and whisker plot in different injury location

The models resulted in some additional compelling findings. Injury severity was not found among high-importance factors in the model results. However, according to the outcome, crash types that were typically more severe (i.e., single vehicle) had slightly faster response. Surprisingly, disposition with light/siren was not among the high-importance factors in the RF model. Additionally, some studies found the impact of light/siren to increase the risk of emergency vehicle crashes (Bertholet et al., 2020; Watanabe et al., 2019). Adding these findings to the previous studies' results calls the functionality of light/siren in emergency vehicles into

question and suggest that further research should be done to identify best practices for their use. EMS priority also did not have a significant impact on response time. However, previous studies cast doubt about the accuracy of priority assigned by the dispatcher based on emergency medical technicians' evaluations on the scene (Palumbo et al., 1996; Slovis et al., 1985).

5.3.5 Conclusion

This study explores factors influencing EMS response time to crashes in Jefferson County, KY. Minimizing EMS response time was identified as one of the factors that can save lives and reduce injury severity in crashes due to the provision of immediate medical care. EMS travel distance, as expected, was identified as the most important factor in EMS response time. Police/EMS location discrepancy, crash type, time of day, number of injuries and injury location were also found to influence EMS response time. The priority of the run and disposition with/without light/siren were not found among top important factors.

Implementation of the study outcome in practice could help EMS to reach its goal of providing immediate care for injuries sustained in motor vehicle crashes. Discrepancies between EMS locations and crash report location suggest action is needed to improve the accuracy of the crash locations reported to EMS. Applying new emergency communication technology in the field of EMS could be a practical option to reduce errors. Considering the impact of time of day on EMS response time, optimizing ambulances' fleet management may help to provide shorter response times by EMS are closer to likely emergency scenes. A thorough investigation is needed to determine whether using light/siren in emergency vehicles is beneficial since, based on the findings of this paper, light/siren did not play a significant

role in providing a faster response. Also, the method to assign high/low priority emergency events should be reconsidered as it does not show a statistically significant difference in EMS response time for crashes.

5.4 Injury Severity Misclassification: Police Officers vs. Emergency Physicians Evaluation, What Drives the Difference?[6]

5.4.1 Introduction

Several issues highlight the importance of accurate crash injury reporting. First, inaccurate data may result in the wrong estimate of crash-related safety model parameters, and consequently, can lead to insufficient safety policies and inappropriately allocated road safety investments. Moreover, incorrect evaluation of injuries on the scene, especially underestimation of non-apparent injuries such as internal injuries and low visibility injuries, may result in a life-threatening injury not being treated. The results of these studies suggest that further investigation to distinguish factors associated with inaccurate crash severity classifications may help to address approaches to field evaluations.

While these issues are prevalent in crash data, police records are currently the most comprehensive source of information for monitoring road safety. In this regard, evaluating the misclassification records and identifying the influencing factors on the discrepancy of injury severity judgment is an important step to improve the quality. The objective of this chapter is to identify factors that contribute to the misclassification of injury severity in crash

[6] Sections from "Hosseinzadeh, A., Kuzel, A., Kluger, R. and Orthober, R. (2022). Injury Severity Misclassification: Police Officers vs. Emergency Physicians Evaluation, What Drives the Difference? Transportation Research Board" included in this sub-chapter.

reports and suggest focus areas where officer training may improve reporting quality. To accomplish those objectives, a panel of Emergency Department (ED) physicians, including medical doctors (MD) and doctors of osteopathic medicine (OD) reviewed detailed medical records of trauma registry patients that were successfully linked to a police report, and classified the severity of the injury according to KABCO scale definitions. To investigate the factors influencing the misclassification of injury severity, an ordered Probit model was employed. The contributing factors investigated in the model included individual-related, crash-related and trauma-related factors.

5.4.2 Method

Data Linkage

This study utilized linked data of police-reported crash data from the Kentucky State Police, emergency medical services (EMS) patient care reports from Louisville Metro Government's Department of Emergency Services and trauma registry data from the University of Louisville Hospital (ULH), the only Level One Trauma Center in Jefferson County, Kentucky. The merits of using the linked data were shown in crash analysis in some recent research (Hosseinzadeh et al., 2020; Ryan et al., 2020; Tainter et al., 2020). This data set analyzed 93 individuals who were involved in the motor vehicle crash, were transported by EMS to ULH and included in the trauma registry. The data included all types of road users and all types of crashes.

Within this dataset, inconsistency in injury classification can be seen by comparing the trauma records and the dispositions of the patients. Figure 5.11 compares the distribution of the severity of injuries sustained by using the crash severity, and Injury Severity Score (ISS)- a quantitative measurement-based value of the severity of injuries sustained in a traumatic event.

ISS uses Abbreviated Injury Scale (AIS) to categorize the severity of an injury by calculating the sum of squares of the three highest AISs. The ISS ranges from 1 to 75 (Greenspan et al., 1985). Based on Figure.5.11, 65% (31% ISS_{8-15} and 34% ISS_{1-7}) of A-level injuries according to crash records had an ISS below 16, which classifies as minor injuries.

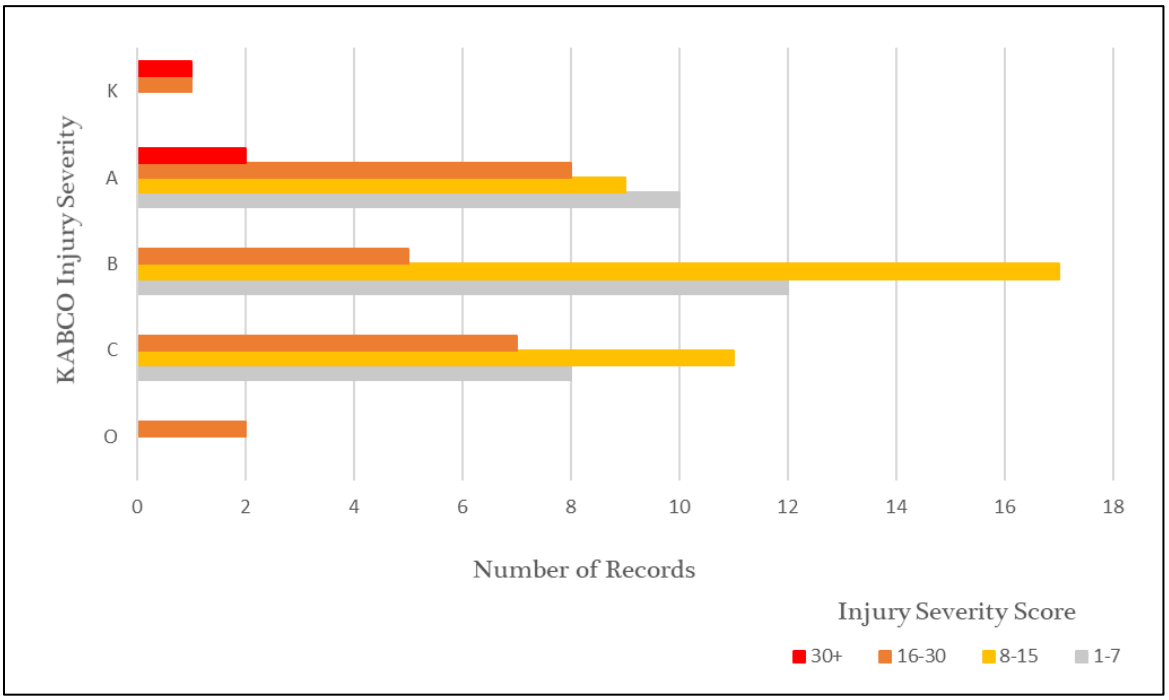


Figure 5.11. Distribution of ISS by KABCO injury severity

Data Preparation

The police-reported variables were used to identify factors that led to discordance between two injury severity ratings are presented in Table 5.8.

Table 5.8. Individual-related and crash-related explanatory variables in police-reported data

	Variable	Levels/ interval
Individual-related	Gender	(1) Female
		(2) Male
	Race	(1) Non-white
		(2) White

	Age	[14 – 90]
Crash-related	Time of day	1. Day 2. Night
	Weekend	1. No 2. Yes
	Number of injuries	1. 1- 2 2. 3+
	Crash type	1. Angle 2. Head on 3. Single vehicle 4. Others

In our study, a hypothesis was head injuries, internal injuries and the visibility of injuries were all likely to contribute to the misclassification of injuries in the field. These factors were quantified in the data by manually reviewing the injury disposition that was reported by the physician at ULH in the patient's medical record. Signs of both head and internal injuries among patients were tracked from the trauma registry data based on the patients' charts. Table 5.9 lists the specific injury dispositions that were mapped to head and internal injury indicators.

Table 5.9. Descriptions of head injury and internal injury

	Variables	Description	
1	Head Injury	head injury in trauma record:	
		concussion	neurological deficits
		GCS 14 or lower	lacerations to the head
		positive loss of consciousness	cerebral hemorrhage
		subdural hematoma	cerebral contusions
		skull fracture	diffuse axonal injury
		orbital floor fracture	orbital wall fracture
		subarachnoid hemorrhage	large scalp hematomas
		interventricular hemorrhage	epidural hematoma
2	Internal Injury	internal injury in trauma record:	
		kidney laceration	adrenal gland hemorrhage
		subdural hematoma	interventricular hemorrhage
		liver laceration	epidural hematoma
		bowel perforation	subarachnoid hemorrhage
		pneumomediastinum	open book pelvic fracture with hemorrhage

	hemothorax	mesenteric tear or laceration
	spleen laceration	diaphragm tear, laceration or rupture
	pneumothorax	bowel or mesentery hematoma
	hemopneumothorax	cerebral hemorrhage
	lung laceration	pancreatic rupture or laceration
	hemomediastinum	

An injury visibility factor was also proposed to describe the external manifestations of injuries sustained in motor vehicle collisions. The injury visibility factor was used to quantify the severity of the visible injuries sustained and to determine if officers on the scene of a motor vehicle crash rated injuries based on external injury visibility. The score was limited to charted injuries information from EMS records as well as documents from assessment within the trauma center. Those individuals with a lack of abrasions, lacerations, contusions, or blood present at the crash scene were afforded an injury visibility factor rating of 1 as there were no apparent injuries at the scene of the collision. The scoring scale increased based on documented external injuries, with 4 being the highest rating. An individual with a large laceration or multiple lacerations on multiple extremities, open fractures, joint instability, or evidence of contusions behind the ear or around the eye was afforded a score of 4. While the injury visibility rating is not an exact quantitative measurement, and much of the accuracy is limited to the charting performed by EMS or trauma center staff, it provided a rudimentary index to estimate the extent of external injuries that may influence an officer in their scoring of injury severity. Table 5.10 includes the criteria for rating the proposed injury visibility factor.

Table 5.10. Description of injury visibility factor levels

Levels	Description
level 1	No abrasions, lacerations, contusions, or blood present on the scene.
level 2	Small or mild abrasions, lacerations or contusions. No signs of seatbelt sign (contusions to the abdomen or on the neck seen in the distribution of the seatbelt.

level 3	Multiple lacerations, abrasions, and contusions. Seatbelt signs present. Moderate bloody appearance on face and extremities.
level 4	Large laceration and/or multiple lacerations on multiple body systems. Open fracture or obvious extremity deformity. Joint instability observed. Evidence of basilar skull fracture (contusions behind the ear or around the eye).

Physician Survey

Injury severity was evaluated through a survey by a panel of six emergency physicians in an effort to capture variance in the opinion surrounding injury severity. Physicians were provided with standard KABCO definitions from the Model Minimum Uniform Crash Criteria (MMUCC) (NHTSA, 2017) and verbal instructions to rate the severity of each injury in the linked dataset on the KABCO scale based on the charted diagnosis from trauma records. After the survey, each injury had injury ratings from six emergency physicians as well as the official rating from the police report.

The Discrepancy in Injury Severity Evaluation Modeling

Discrete choice models for estimating ordinal response data have been applied in exploring injury severity in the traffic safety area (Kockelman & Kweon, 2002). In this application, the difference in occupant injury severity between physician reviews and officer evaluations in police data was modeled using the ordered Probit model. Underlying the indexing in such models is a latent but continuous descriptor of the response. In the ordered Probit model, the random error associated with this continuous descriptor is assumed to follow a normal distribution. Equation 5.12 shows the ordered Probit model.

$$T_n^* = \beta' z_n + \varepsilon_n \qquad (5.12)$$

Where T_n^* is the latent and continuous measure of the difference in the evaluated injury severity of injured individual *n* in a crash, z_n is a vector of explanatory variables describing the

characteristics of individuals characteristics, crash features and detailed trauma registry data. β is a vector of parameters to be estimated, and ε_n is a random error term that is assumed to follow a standard normal distribution (Greene, 2000).

The mode of evaluated injury severities of physician surveys was used in this study. Physicians-surveyed injury severity ranged from 1 (=K) to 4 (=C), and police-reported injury severity ranged from 1 (=K) to 5 (=O). The difference (physician injury severity – police-reported injury severity) ranged from -3 to 1 in the mode of physician survey.

The observed and coded discrete injury severity variable, T_n, is determined as follows (equation 5.13):

$$\begin{aligned} T_n = \{ & 0 \; if \; -\infty \leq T_n^* \leq \mu_1 \; (PE - OE) = \{-2, -3\} \\ & 1 \; if \; \mu_1 \leq T_n^* \leq \mu_2 \; (PE - OE) = -1 \\ & 2 \; if \; \mu_2 \leq T_n^* \leq \mu_3 \; (PE - OE) = 0 \\ & 3 \; if \; \mu_3 \leq T_n^* \leq \infty \; (PE - OE) = 1\} \end{aligned} \qquad (5.13)$$

Where the μ_i represent threshold to be estimated, *PE* represents the physician evaluation of injury severity in the survey and *OE* shows the reporting officer evaluation of injury severity. For more information on the ordered Probit model specification, see (Greene, 2000). It's necessary to specify that only the variables of interest were kept in the analysis due to a limited number of observations. Even if the number of records was higher, it was not feasible to ask physicians to specifically go through each record's detailed information and rate the severity while maintaining consistency and quality reviews.

5.4.3 Results

The dependent variable was the difference between the mode of evaluated injury severity among physicians and police-reported injury severity. Table 5.11 shows the frequency of each observed discrepancy between the officer evaluation and mode of physician evaluation.

Table 5.11. Frequency of injury severity mode differences among physicians

PE-OE				
-2 & -3	-1	0	1	2 & 3
13	35	35	10	0

Table 5.12. Ordered Probit model results for the difference in injury severity

Variable	**Value**	**SD**	**t-value**	**OR**	**OR Lower bound**	**OR upper bound**
Gender	-0.032	0.248	-0.132	0.967	0.594	1.574
Age	**-0.014**	**0.006**	**-2.195**	**0.985**	**0.972**	**0.998**
Race - white	0.176	0.271	0.650	1.192	0.701	2.03
Crash type - head on	-0.179	0.348	-0.514	0.835	0.422	1.653
Crash type - single vehicle	0.154	0.302	0.510	1.167	0.645	2.114
Crash type - others	0.026	0.375	0.070	1.026	0.491	2.142
Time of day - night	-0.091	0.252	-0.360	0.912	0.555	1.498
Weekend - yes	0.074	0.266	0.280	1.077	0.639	1.816
Number of Injuries – 2+	0.169	0.287	0.590	1.185	0.674	2.083
Internal injury - yes	**-0.486**	**0.269**	**-1.711**	**0.614**	**0.361**	**1.041**
Head injury - yes	-0.009	0.288	-0.032	0.990	0.563	1.743
Injury visibility - linear	**-0.587**	**0.342**	**-1.115**	**0.555**	**0.282**	**1.084**
Injury visibility - quadratic	-0.215	0.285	-0.753	0.806	0.459	1.409
Injury visibility - cubic	-0.410	0.255	-1.607	0.663	0.401	1.092
Thresholds						
$\boldsymbol{\mu_1}$	**-1.718**	**0.513**	**-3.348**			
$\boldsymbol{\mu_2}$	**-0.499**	**0.488**	**-1.863**			
$\boldsymbol{\mu_3}$	**0.830**	**0.502**	**1.654**			
Residual deviance	**217.14**					
AIC	**251.14**					
LL	**-108.57**					
Likelihood odds ratio ($\boldsymbol{\rho^2}$)	**0.074**					

Table 5.12 presents the ordered Probit model results. The coefficients, significance, odds ratio, and 95% confidence interval are also included in Table 5.11. The results indicate age, internal injury, and injury visibility factor were significant in misclassification. The negative sign on

coefficients indicates physicians viewed those factors as contributing to injuries that are more severe than the officers (PE > OE), which are cases of underestimation among trauma patients. All three of the statistically significant factors indicated that police were likely to underestimate the trauma injuries. Thresholds reported in the model were found to be significant in all classes. Likelihood odds ratio indicates the ratio of maximum likelihood with the explanatory variable set divided by the maximum likelihood without the explanatory variable set. It's important to note that the results are only informative among injuries that warranted inclusion in the ULH Trauma Registry. In other words, this model has identified factors that may lead to underestimations of severity, which may extend to the entire population of crash-involved individuals, but the rate at which this occurs cannot be determined here.

5.4.4 Discussion and Practical Applications

According to Table 5.12, age was found to be a significant factor in injury severity difference. The outcome shows that as the age increases, police officers were more likely to underestimate injury severity showing that crashes may lead to more complex and unknown injuries among older people. Another possibility could be the fact that older individuals are more prone to have a preexisting health condition, which exacerbates their situations later in the ED when their diagnoses are entered by trauma center staff. This result is in contrast with literature that found officers' overestimation of 65 years old and older adults in Hong Kong (Tsui et al., 2009) and New Zealand (McDonald et al., 2009), though differences in officer field training, cultural differences or other factors between the US and those countries may be the root cause of those differences.

In addition, reports of internal injuries in hospital records were associated with underestimation of injuries. It seems that officers were not able to identify cases with likely internal injuries among the cases evaluated. The presence of internal injuries, on average, is 38.6% more likely to lead police officers to underestimate injury severity in this study. Internal injury is not generally available as a factor in hospital data, and it has not been considered as a contributing factor in other crash-hospital data linkage studies. Therefore, this finding may be extremely valuable and warrants further investigation, particularly surrounding how officers can be trained to recognize signs of internal injury.

The injury visibility factor was found as an index in underestimating injury severity. According to the outcome, the more visible injuries are more likely to result in officers' underestimation of injury severity. It is unclear why this is the case, but perhaps officers are too conservative about overestimating injuries or maybe the blood and/or swelling makes it hard to identify the severity of the injury. Again, the proposed injury visibility factor is limited to charting by hospital emergency personnel. Further study is needed to understand what visual cues officers may be using at the scene when filing injury information.

Neither crash type nor the number of injuries of the crash was found to be significant factors. Although there was a presumption that more severe crash types (e.g., single-vehicle) and more injuries sustained from a crash (e.g., two and more) have an impact on officers' misjudgment, the results did not show a significant relationship between them. The officer training surrounding the typical injuries based on crash type appears to be sufficient. Time of the day and weekday/weekend were the other two variables that did not show a significant effect on injury severity discordances in this study.

5.4.5 Conclusion, Limitations, and Future Work

This chapter aims to investigate the factors associated with injury severity discrepancy in motor vehicle crashes. Findings indicate that officers tended to underestimate injuries specifically associated with high injury visibility, increasing age, and the presence of an internal injury.

This study has several limitations that can be addressed with further research. Small sample size may influence the results. Specifically, the physician survey would be difficult to implement with a larger sample since it required a review of individual records and protected health information. Additionally, it would require a much larger sample of linked records with charting that met the standards of a level 1 trauma center. The other limitation is associated with taking the difference between physician and police-reported injuries. The difference implies that there is an equal "distance" between each level on the KABCO scale. However, a -1 value of the response variable could be an officer labeling B on a Physician-labelled A crash, or it could be an officer labeling C on a Physician-labelled O crash. Further research should also be devoted to developing field tests that support officer injury assessment. Also, results suggest that injury visibility is important and therefore should be investigated further for the purposes of reporting.

CHAPTER 6

SUMMARY, CONTRIBUTIONS, AND FUTURE DIRECTIONS

In this dissertation, a framework was developed to link crash-related datasets. Furthermore, the linked dataset, which is not available traditionally and is not commonly used for safety analysis, was used for ad-hoc analysis. This dissertation aims to propose a method for linking crash-related datasets, examining the adaptability of the proposed method on another dataset, comparing different linkage methods, and providing some showcases of what the linked dataset can add to safety research. This dissertation suggests a holistic research approach regarding improving safety analysis by incorporating several crash-related datasets and how crash outcome assessment can benefit from a linked dataset. Here each chapter is adapted based on the available data, but the theme is transferable to other crash-related datasets, other geographical contexts, and other applications.

In chapter 2, a review of the existing literature was conducted, and gaps were highlighted. In chapter 3, a heuristic framework is developed to match EMS run reports to crashes through time, location, and other indicators present in both datasets. Types of matches between EMS and crashes were classified. To investigate the fidelity of the matching approach, a manual review of a sample of data was conducted. A comparative bias analysis was implemented on several key variables. 72.2% of EMS run reports matched to a crash record, and 69.3% of trauma registry records matched with a crash record. Females, individuals between 11 to 20 years old, and individuals involved in single-vehicle or head-on crashes were more likely to be present in linked data sets. Using the linked data sets, relationships between EMS response time and reported injury in the crash report and between police-reported injury and injury severity score was examined. Linked crash - EMS CAD – PCR – trauma registry data provides a valuable opportunity to evaluate the impact of prehospital care and emergency

department care on crash outcomes. In general, policy steps could be taken to require cross-reporting and linkage of the data sets as the events occur to better monitor outcomes of injury crashes without requiring post hoc linkage. This method can also realistically be integrated into a tool or software to undergo record linkage automatically. In chapter 4 Bayesian record linkage method was implemented and the results were compared with the already developed heuristic algorithm. The linkage rate was compared, and consistent and inconsistent pairs matches were identified.

Chapter 5 highlights the applications of the linked data. Sub-chapter 5.2, utilized the linked data of police-reported crash data and EMS runs, including 2480 crash injuries that transferred to hospital. A random-effects ordered probit approach was implemented to identify effective factors on crash injury severity. Three models of (1) crash-related variables, (2) crash-related and EMS times and (3) crash-related, EMS times and interaction effects were estimated. The outcome could not find the impact of faster EMS times on injury severity. The highest scene time and the highest transport to hospital time categories resulted in a less severe outcome. Based on the outcome, the authors did not find a significant relationship between EMS times and injury severity. Adding EMS time and interaction effects of EMS times, based on different body injury locations to the model, improved the model quality marginally.

In sub-chapter 5.3, EMS response time was modeled and compared using four machine-learning approaches, as well as regression analysis. The most successful approach in terms of root means square error and goodness of fit was chosen to represent contributing factors. The results show variables such as emergency medical services travel distance, the discrepancy between crash location reported in police and emergency medical services data, and crash type were important factors in EMS response time. The study outcome can be used

to guide practice and help EMS reduce the time to care for individuals injured in motor vehicle crashes. EMS travel distance, as expected, was identified as the most important factor in EMS response time. Police/EMS location discrepancy, crash type, time of day, number of injuries, and injury location were also found to influence EMS response time. The priority of the run and disposition with/without light/siren was not found among the top important factors.

In sub-chapter 5.4, The discrepancy between police-reported injury severities and physicians' evaluations of corresponding trauma records was modeled. The trauma data were reviewed and classified by a panel of emergency physicians. An ordered probit model was used to model factors contributing to misclassification between police reports and emergency physicians. According to the results, age, internal injury, and injury visibility rating were found to be contributing factors to injury severity discrepancy. Internal injury and injury visibility ratings were among the trauma-related factors that were developed to explore their impact on injury severity discrepancy. Findings indicate officers tended to underestimate injuries associated with high injury visibility, increasing age, and the presence of an internal injury, specifically among trauma patients.

In summary, in this dissertation, an implementation of crash-related dataset was conducted as well as showcases of how and what analyzing these datasets can add to safety research. The first step was conducting the linkage. Since various agencies are responsible for gathering data, common identifiers are not available in the dataset; hence most of the times a deterministic linkage is not applicable. Therefore, a heuristic linkage framework or a probabilistic approach such as the Bayesian approach can be used for data linkage purposes. The next step is making sure that the linked data is a representative sample of each individual dataset to make sure the inferences are not biased. In chapter 3 of this dissertation, a

descriptive exploration was performed to ensure an unexpected bias was not imposed on the dataset. For instance, although about 65% of the crash-person records are no injury crashes, only 10% of the records ended up in the dataset. However, it's not an unexpected bias since crashes with less severity has lower chance to be available in EMS runs data and trauma registry. As the final step of this section, the proposed heuristic algorithm expanded and adapted across the state of Kentucky to examine the adaptability. There are three possible future directions for this section: first, incorporating more data sources such as roadway inventories, traffic operations data (e.g., Waze), EMS dispatch data, Census data, medical billing data and driver/vehicle records in the linkage process. Second, joint modeling of selectivity bias among the linked dataset and individual datasets to determine which variables are significantly influential in leading biased datasets. Third, spatial analysis of county-level linkage rate and exploring associated factors. The next chapter examined the fidelity of linkage by comparing the results of the heuristic algorithm with the outcome of implementing the Bayesian probabilistic record linkage.

The dissertation followed by the implications of the linked datasets to answer research questions that were not possible to answer without the linked dataset. The association between injury severity and EMS response time was investigated by analyzing police-reported crash data – EMS CAD - PCR data. Further exploration revealed the part of the injured body plays a role in the association of injury severity and EMS response time. The second question of interest was exploring the factors impacting EMS response time, including the demographics, weather-related factors and crash characteristics. As future steps, spatial analysis of EMS response time and transport time and modeling hospital coverage area and reasons behind choosing the facilities can be considered.

The dissertation was followed by an investigation on misclassification of injury severity according to crash rating and emergency physician survey based on the detailed information in trauma registry data. The factors affecting the misclassification were investigated. However, it's not possible to distinguish the misclassification is a result of police officers' over/under-estimation or the status of the injured individuals changed during the transfer to hospital. Future studies can incorporate more data sources to differentiate between officers' misjudgment and changing the status. For example, in Kentucky state data there is a field that states how the status of the injured individual has changed during the transport and upon arrival to the hospital.

Linking the crash-related datasets unlock incredible potential for safety analysis. Linking datasets have not been extensively used or involved in safety narratives. In this dissertation, we elaborate on a data linkage and showcase the linkage application. In general, policy steps could be taken to require cross-reporting and linkage of the data sets as the events occur to better monitor outcomes of injury crashes without requiring post-hoc linkage.

REFERENCES

Alsop, J., Langley, J., 2001. Under-reporting of motor vehicle traffic crash victims in New Zealand. Accident Analysis & Prevention 33, 353-359.

American Association of State Highway and Transportation Officials, 2010. Highway Safety Manual. Washington, DC: American Association of State Highway and Transportation Officials.

Amorim, M., Ferreira, S., Couto, A., 2014. Linking police and hospital road accident records: how consistent can it be? Transportation Research Record 2432, 10-16.

Amorim, M., Ferreira, S., Couto, A., 2019. How do traffic and demand daily changes define urban emergency medical service (uEMS) strategic decisions?: A robust survival model. Journal of Transport & Health 12, 60-74.

Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accident Analysis & Prevention 38, 627-635.

Archer, K.J., Hedeker, D., Nordgren, R., Gibbons, R.D., 2018. mixor: Mixed-Effects Ordinal Regression Analysis. R package version 1.0.4. https://CRAN.R-project.org/package=mixor

Baker, S.P., o'Neill, B., Haddon Jr, W. and Long, W.B., 1974. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. Journal of Trauma and Acute Care Surgery, 14(3), pp.187-196.

Bakhshi, A.K. and Ahmed, M.M., 2020. Practical advantage of crossed random intercepts under Bayesian hierarchical modeling to tackle unobserved heterogeneity in clustering critical versus non-critical crashes. Accident Analysis & Prevention, 149, p.105855.

Benavente, M., Knodler Jr, M.A., Rothenberg, H., 2006. Case study assessment of crash data challenges: Linking databases for analysis of injury specifics and crash compatibility issues. Transportation research record 1953, 180-186.

Bertholet, O., Pasquier, M., Christes, E., Wirths, D., Carron, P.N., Hugli, O. and Dami, F., 2020. Lights and Siren transport and the need for hospital intervention in nontrauma patients: a prospective study. *Emergency Medicine International*.

Boufous, S., Finch, C., Hayen, A., Williamson, A., 2008. Data linkage of hospital and police crash datasets in NSW. NSW Injury Risk Management Research Centre.

Boufous, S. and Williamson, A., 2006. Work-related traffic crashes: A record linkage study. Accident Analysis & Prevention, 38(1), pp.14-21.

Breiman, L., 1996. Bagging predictors. *Machine learning*, *24*(2), pp.123-140.

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

Burch, C., Cook, L., Dischinger, P., 2014. A comparison of KABCO and AIS injury severity metrics using CODES linked data. Traffic injury prevention 15, 627-630.

Burdett, B., Li, Z., Bill, A.R., Noyce, D.A., 2015. Accuracy of injury severity ratings on police crash reports. Transportation research record 2516, 58-67.

Ceklic, E., Tohira, H., Ball, S., Brown, E., Brink, D., Bailey, P., Whiteside, A. and Finn, J., 2021. Motor vehicle crash characteristics that are predictive of high acuity patients: an analysis of linked ambulance and crash data. Prehospital emergency care, 25(3), pp.351-360.

Chen, B., Maio, R.F., Green, P.E., Burney, R.E., 1995. Geographic variation in preventable deaths from motor vehicle crashes. Journal of Trauma and Acute Care Surgery 38 (2), 228-232.

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chen, X., Guyette, F.X., Peitzman, A.B., Billiar, T.R., Sperry, J.L. and Brown, J.B., 2019. Identifying patients with time-sensitive injuries: association of mortality with increasing prehospital time. *Journal of trauma and acute care surgery*, *86*(6), pp.1015-1022.

Chin, S.N., Cheah, P.K., Arifin, M.Y., Wong, B.L., Omar, Z., Yassin, F.M. and Gabda, D., 2017, April. Determinants of ambulance response time: A study in Sabah, Malaysia. In *AIP Conference Proceedings* (Vol. 1830, No. 1, p. 080003). AIP Publishing LLC.

Chitturi, M.V., Ooms, A.W., Bill, A.R., Noyce, D.A., 2011. Injury outcomes and costs for cross-median and median barrier crashes. Journal of safety research 42, 87-92.

Clark, D.E., 2004. Practical introduction to record linkage for injury research. *Injury Prevention* 10(3), 186-191.

Conderino, S., Fung, L., Sedlar, S., Norton, J.M., 2017. Linkage of traffic crash and hospitalization records with limited identifiers for enhanced public health surveillance. Accident Analysis & Prevention 101, 117-123.

Cong, H., Chen, C., Lin, P.S., Zhang, G., Milton, J. and Zhi, Y., 2018. Traffic incident duration estimation based on a dual-learning Bayesian network model. *Transportation research record*, *2672*(45), pp.196-209.

Conner, K.A., Smith, G.A., 2014. The impact of aggressive driving-related injuries in Ohio, 2004–2009. Journal of safety research 51, 23-31.

Cook, L.J., Thomas, A., Olson, C., Funai, T., Simmons, T., 2015. Crash Outcome Data Evaluation System (CODES): An Examination of Methodologies and Multi-State Traffic Safety Applications.

Copes, W.S., Champion, H.R., Sacco, W.J., Lawnick, M.M., Keast, S.L. and Bain, L.W., 1988. The injury severity score revisited. Journal of Trauma and Acute Care Surgery, 28(1), pp.69-77.

Couperthwaite, A.B.G., 2015. *Emergency Medical Services Response Time and Mortality in Paediatric Trauma Patients in the Urban Setting: A Cohort Study* (Master's thesis, Graduate Studies).

Couto, A., Amorim, M., Ferreira, S., 2016. Reporting road victims: assessing and correcting data issues through distinct injury scales. Journal of safety research 57, 39-45.

Cryer, P.C., Westrup, S., Cook, A.C., Ashwell, V., Bridger, P. and Clarke, C., 2001. Investigation of bias after data linkage of hospital admissions data to police road traffic crash reports. Injury prevention, 7(3), pp.234-241.

Dean, J.M., Vernon, D.D., Cook, L., Nechodom, P., Reading, J., Suruda, A., 2001. Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services. Annals of emergency medicine 37, 616-626.

Dharap, S., Kamath, S., Kumar, V., 2017. Does prehospital time affect survival of major trauma patients where there is no prehospital care? Journal of postgraduate medicine 63 (3), 169.

Ding, C., Ma, X., Wang, Y. and Wang, Y., 2015. Exploring the influential factors in incident clearance time: disentangling causation from self-selection bias. *Accident Analysis & Prevention*, *85*, pp.58-65.

Doggett, S., Ragland, D.R., Felschundneff, G., 2018a. Prehospital response time and traumatic injury—a review. Institute of Transportation Studies at UC Brekley.

Doggett, S., Ragland, D.R. and Felschundneff, G., 2018b. Evaluating Research on Data Linkage to Assess Underreporting of Pedestrian and Bicyclist Injury in Police Crash Data.

Doidge, J.C. and Harron, K.L., 2019. Reflections on modern methods: linkage error bias. International journal of epidemiology, 48(6), pp.2050-2060.

Dove A, Pearson J, Weston P. 1986. Data collection from road traffic accidents. Emergency Medicine Journal 3 (3), 193-198.

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis & Prevention 40 (3), 1033-1054.

Farmer C.M. 2003. Reliability of police-reported information for determining crash and injury severity. Traffic injury prevention, 4, 38-44.

Feero, S., Hedges, J.R., Simmons, E., Irwin, L., 1995. Does out-of-hospital ems time affect trauma survival? The American journal of emergency medicine 13 (2), 133-135.

Ferreira, S., Amorim, M., Couto, A., 2017. Risk factors affecting injury severity determined by the MAIS score. Traffic injury prevention 18, 515-520.

Ferreira, S., Amorim, M., Couto, A., 2019. The prehospital time impact on traffic injury from hospital fatality and inpatient recovery perspectives. Journal of Transportation Safety & Security, 1-21.

Ferreira, S., Falcão, L., Couto, A., Amorim, M., 2015. The quality of the injury severity classification by the police: An important step for a reliable assessment. Safety science 79, 88-93.

Gonzalez, R.P., Cummings, G., Mulekar, M. and Rodning, C.B., 2006. Increased mortality in rural vehicular trauma: identifying contributing factors through data linkage. Journal of Trauma and Acute Care Surgery, 61(2), pp.404-409.

Gonzalez, R.P., Cummings, G.R., Phelan, H.A., Mulekar, M.S. and Rodning, C.B., 2009. Does increased emergency medical services prehospital time affect patient mortality in rural motor vehicle crashes? A statewide analysis. *The American journal of surgery*, *197*(1), pp.30-34.

Greene W.H. 2000. Econometric analysis 4th edition. International edition, New Jersey: Prentice Hall, 201-215.

Greenspan, L., McLellan, B.A. and Greig, H., 1985. Abbreviated injury scale and injury severity score: A scoring chart. The Journal of trauma, 25(1), pp.60-64.

Haleem, K., Alluri, P., Gan, A., 2015. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accident Analysis & Prevention 81, 14-23.

Han, G.-M., Newmyer, A., Qu, M., 2017. Seatbelt use to save money: Impact on hospital costs of occupants who are involved in motor vehicle crashes. International emergency nursing 31, 2-8.

Harmsen, A., Giannakopoulos, G., Moerbeek, P., Jansma, E., Bonjer, H., Bloemers, F., 2015. The influence of prehospital time on trauma patients outcome: A systematic review. Injury 46 (4), 602-609.

Harron, K., Doidge, J.C. and Goldstein, H., 2020. Assessing data linkage quality in cohort studies. Annals of Human Biology, 47(2), pp.218-226.

Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B. and Goldstein, H., 2014. Evaluating bias due to data linkage error in electronic healthcare records. BMC medical research methodology, 14(1), pp.1-10.

He, Z., Qin, X., Renger, R. and Souvannasacd, E., 2019. Using spatial regression methods to evaluate rural emergency medical services (EMS). *The American Journal of Emergency Medicine*, *37*(9), pp.1633-1642.

Hojati, A.T., Ferreira, L., Washington, S. and Charles, P., 2013. Hazard based models for freeway traffic incident duration. *Accident Analysis & Prevention*, *52*, pp.171-181.

Hosseinzadeh, A., Karimpour, A., Kluger, R. and Orthober, R., 2020, January. A framework to link crashes to emergency medical service runs and trauma admissions: for improved highway safety monitoring and crash outcome assessment. In Transportation Research Board. 99th Annual Meeting Transportation Research Board. United States Washington DC.

Hosseinzadeh, A. and Kluger, R., 2021a. Do EMS times associate with injury severity?. Accident Analysis & Prevention, 153, p.106053.

Hosseinzadeh, A. and Kluger, R., 2021b. Data Linkage for Traffic Safety in Jefferson County, Kentucky. In International Conference on Transportation and Development (pp. 243-250).

Hosseinzadeh, A., Haghani, M. and Kluger, R., 2021a. Exploring Influencing Factors on Crash-related Emergency Response Time: A Machine Learning Approach (No. TRBAM-21-00614).

Hosseinzadeh, A., Karimpour, A., Kluger, R. and Orthober, R., 2020. A Framework to Link Crashes to Emergency Medical Service Runs and Trauma Admissions: For Improved Highway Safety Monitoring and Crash Outcome Assessment. In Transportation Research Board. 99th Annual Meeting Transportation Research Board.

Hosseinzadeh, A., Karimpour, A., Kluger, R., Orthober, R., 2022. Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records. *Journal of Safety Research*.

Hosseinzadeh, A., Moeinaddini, A. and Ghasemzadeh, A., 2021b. Investigating factors affecting severity of large truck-involved crashes: Comparison of the SVM and random parameter logit model. *Journal of safety research*, *77*, pp.151-160.

Hou, L., Lao, Y., Wang, Y., Zhang, Z., Zhang, Y. and Li, Z., 2013. Modeling freeway incident response time: A mechanism-based approach. *Transportation Research Part C: Emerging Technologies*, *28*, pp.87-100.

Hu, W., Dong, Q., Huang, B., 2017. Correlations between road crash mortality rate and distance to trauma centers. Journal of Transport & Health 6, 50-59.

Janstrup, K.H., Kaplan, S., Hels, T., Lauritsen, J. and Prato, C.G., 2016. Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. Traffic Injury Prevention, 17(6), pp.580-584.

Kamaluddin, N.A., Abd Rahman, M.F. and Várhelyi, A., 2019. Matching of police and hospital road crash casualty records–a data-linkage study in Malaysia. International journal of injury control and safety promotion, 26(1), pp.52-59.

Karmel, R., Anderson, P., Gibson, D., Peut, A., Duckett, S., Wells, Y., 2010. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC health services research 10, 41.

Katayama, Y., Kitamura, T., Kiyohara, K., Sado, J., Hirose, T., Matsuyama, T., Kiguchi, T., Izawa, J., Nakagawa, Y., Shimazu, T., 2019. Prehospital factors associated with death on hospital arrival after traffic crash in japan: A national observational study. BMJ open 9 (1), e025350.

Katz, M., 2006. Study design and statistical analysis: a practical guide for clinicians. Cambridge University Press.

Kentucky State Police. Kentucky collison for the Public. Available: http:// crashinformationky.org (accessed 19.07.06).

Khoda Bakhshi, A. and Ahmed, M.M., 2020. Utilizing black-box visualization tools to interpret non-parametric real-time risk assessment models. Transportmetrica A: Transport Science, pp.1-27.

Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: An application of ordered probit models. Accident Analysis & Prevention 34 (3), 313-321.

Kudryavtsev, A.V., Kleshchinov, N., Ermolina, M., Lund, J., Grjibovski, A.M., Nilssen, O. and Ytterstad, B., 2013. Road traffic fatalities in Arkhangelsk, Russia in 2005–2010: Reliability of police and healthcare data. Accident Analysis & Prevention, 53, pp.46-54.

Kumar, A., Abudayyeh, O., Fredericks, T., Kuk, M., Valente, M. and Butt, K., 2017. Trend analyses of emergency medical services for motor vehicle crashes: Michigan case study. *Transportation Research Record*, *2635*(1), pp.55-61.

Lam, S.S.W., Nguyen, F.N.H.L., Ng, Y.Y., Lee, V.P.X., Wong, T.H., Fook-Chong, S.M.C. and Ong, M.E.H., 2015. Factors affecting the ambulance response times of trauma incidents in Singapore. Accident Analysis & Prevention, 82, pp.27-35.

Laman, H., Yasmin, S. and Eluru, N., 2018. Joint modeling of traffic incident duration components (reporting, response, and clearance time): a copula-based approach. *Transportation research record*, *2672*(30), pp.76-89.

Langley, J.D., Dow, N., Stephenson, S. and Kypri, K., 2003. Missing cyclists. Injury prevention, 9(4), pp.376-379.

Lee, J., Abdel-Aty, M., Cai, Q., Wang, L., 2018. Effects of emergency medical services times on traffic injury severity: A random effects ordered probit approach. Traffic injury prevention 19 (6), 577-581.

Legler, J., Rojas Jr, J. and Mann, N.C., 2017. NEMSIS V3 StateDataSet: software developer technical guide.

Li, R., Pereira, F.C. and Ben-Akiva, M.E., 2018. Overview of traffic incident duration analysis and prediction. European transport research review, 10(2), pp.1-13.

Loo, B.P., Tsui, K., 2007. Factors affecting the likelihood of reporting road crashes resulting in medical treatment to the police. Injury prevention 13, 186-189.

Lovely, R., Trecartin, A., Ologun, G., Johnston, A., Svintozelskiy, S., Vermeylen, F., Thiel, D., Golden, D., Casos, S., Granet, J., 2018. Injury severity score alone predicts mortality when compared to ems scene time and transport time for motor vehicle trauma patients who arrive alive to hospital. Traffic injury prevention 19 (sup2), S167-S168.

Lu, Y., Davidson, A., 2017. Fatal motor vehicle crashes in texas: Needs for and access to emergency medical services. Annals of GIS 23 (1), 41-54.

Lujic, S., Finch, C., Boufous, S., Hayen, A. and 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), pp.28-33.

Ma, L., Zhang, H., Yan, X., Wang, J., Song, Z., Xiong, H., 2019. Smooth associations between the emergency medical services response time and the risk of death in road traffic crashes. Journal of Transport & Health 12, 379-391.

Ma, X., Ji, Y., Yuan, Y., Van Oort, N., Jin, Y., Hoogendoorn, S., 2020. A comparison in travel patterns and determinants of user demand between docked and dockless bike-sharing systems using multi-sourced data. Transportation Research Part A: Policy and Practice 139, 148-173.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research 1, 1-22.

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, 184-190.

McGlincy, M.H., 2004. A Bayesian record linkage methodology for multiple imputation of missing links, ASA Proceedings of the Joint Statistical Meetings. Citeseer, pp. 4001-4008.

McGlincy, M.H., 2006. Using test databases to evaluate record linkage models and train linkage practitioners. Proceedings of the 29th American Statistical Association, Survey Research Method Section, Seattle, WA, 3404-3410.

Medrano, N.W., Villarreal, C.L., Price, M.A., Mackenzie, E., Nolte, K.B., Phillips, M.J., Stewart, R.M., Eastridge, B.J., 2019. Multi-institutional multidisciplinary injury mortality investigation in the civilian pre-hospital environment (mimic): A methodology for reliably measuring prehospital time and distance to definitive care. Trauma surgery & acute care open 4 (1), e000309.

Milani, J., Kindelberger, J., Bergen, G., Novicki, E., Burch, C., Ho, S., West, B., 2015. Assessment of characteristics of state data linkage systems.

Mitchell, R.J., Senserrick, T., Bambach, M.R. and Mattos, G., 2015. Comparison of novice and full-licenced driver common crash types in New South Wales, Australia, 2001–2011. Accident Analysis & Prevention, 81, pp.204-210.

Möller, A., Hunter, L., Kurland, L., Van Hoving, D.J., 2018. The association between hospital arrival time, transport method, prehospital time intervals, and in-hospital mortality in trauma patients presenting to khayelitsha hospital, cape town. African Journal of Emergency Medicine 8 (3), 89-94.

Moore, M.A., 1998. Comparison of young and adult driver crashes in Alaska using linked traffic crash and hospital data. US Department of Transportation, National Highway Traffic Safety Administration.

Morris, A., Mackay, M., Wodzin, E., Barnes, J., 2003. Some injury scaling issues in uk crash research. Monash University Accident Research Center, Report No. CR 199.

Naik, B., Tung, L.-W., Zhao, S., Khattak, A.J., 2016. Weather impacts on single-vehicle truck crash injury severity. Journal of safety research 58, 57-65.

Nasser, A.A., Nederpelt, C., El Hechi, M., Mendoza, A., Saillant, N., Fagenholz, P., Velmahos, G., Kaafarani, H.M., 2020. Every minute counts: The impact of pre-hospital response time and scene time on mortality of penetrating trauma patients. The American Journal of Surgery 220 (1), 240-244.

National Highway Traffic Safety Administration, 2017. MMUCC Guideline: Model Minimum Uniform Crash Criteria Fifth Edition. (Report No. DOT HS 812 433). Washington, DC: National Highway Traffic Safety Administration.

Newgard, C.D., Schmicker, R.H., Hedges, J.R., Trickett, J.P., Davis, D.P., Bulger, E.M., Aufderheide, T.P., Minei, J.P., Hata, J.S., Gubler, K.D., 2010. Emergency medical

services intervals and survival in trauma: Assessment of the "golden hour" in a north american prospective cohort. Annals of emergency medicine 55 (3), 235-246. e4.

Olsen, C.S., Thomas, A.M., Cook, L.J., 2014. Hospital charges associated with motorcycle crash factors: a quantile regression analysis. Injury prevention 20, 276-280.

Paixão, L.M.M.M., Gontijo, E.D., Mingoti, S.A., Costa, D.A.D.S., Friche, A.A.D.L. and Caiaffa, W.T., 2015. Urban road traffic deaths: data linkage and identification of high-risk population sub-groups. Cadernos de saude publica, 31, pp.92-106.

Palumbo, L., Kubincanek, J., Emerman, C., Jouriles, N., Cydulka, R. and Shade, B., 1996. Performance of a system to determine EMS dispatch priorities. *The American journal of emergency medicine*, *14*(4), pp.388-390.

Picado-Aguilar, G. and Aguero-Valverde, J., 2020. Emergency response times and crash risk: an analysis framework for Costa Rica. *Journal of Transport & Health*, *16*, p.100818.

Pons, P.T. and Markovchick, V.J., 2002. Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome?. The Journal of emergency medicine, 23(1), pp.43-48.

Popkin C.L, Campbell B, Hansen A.R, Stewart R. 1991. Analysis of the accuracy of the existing KABCO injury scale. University of North Carolina Highway Safety Research Center, Chapel Hill, NC.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ryan, A., Page, M., Christofa, E., Fitzpatrick, C. and Knodler, M., 2020. Linking geospatial crash and citation data to inform equitable enforcement decisions. *Journal of Transportation Safety & Security*, pp.1-24.

Ryan, A., Tainter, F., Fitzpatrick, C., Gazzillo, J., Riessman, R. and Knodler, M., 2020. The impact of sex on motor vehicle crash injury outcomes. *Journal of Transportation Safety & Security*, pp.1-25.

Sánchez-Mangas, R., García-Ferrrer, A., De Juan, A., Arroyo, A.M., 2010. The probability of death in road traffic accidents. How important is a quick medical response? Accident Analysis & Prevention 42 (4), 1048-1056.

Schapire, R.E., 2003. The boosting approach to machine learning: An overview. Nonlinear estimation and classification, pp.149-171.

Sciortino, S., Vassar, M., Radetsky, M., Knudson, M.M., 2005. San Francisco pedestrian injury surveillance: mapping, under-reporting, and injury severity in police and hospital records. Accident Analysis & Prevention 37, 1102-1113.

Shen, S., Neyens, D.M., 2015. The effects of age, gender, and crash types on drivers' injury-related health care costs. Accident Analysis & Prevention 77, 82-90.

Short, J., Caulfield, B., 2016. Record linkage for road traffic injuries in Ireland using police hospital and injury claims data. Journal of safety research 58, 1-14.

Slovis, C.M., Carruth, T.B., Seitz, W.J., Thomas, C.M. and Elsea, W.R., 1985. A priority dispatch system for emergency medical services. *Annals of emergency medicine*, *14*(11), pp.1055-1060.

Stiell, I.G., Wells, G.A., Field III, B.J., Spaite, D.W., De Maio, V.J., Ward, R., Munkley, D.P., Lyver, M.B., Luinstra, L.G., Campeau, T. and Maloney, J., 1999. Improved out-of-hospital cardiac arrest survival through the inexpensive optimization of an existing defibrillation program: OPALS Study Phase II. *Jama*, *281*(13), pp.1175-1181.

Stutts, J.C. and Hunter, W.W., 1999. Motor vehicle and roadway factors in pedestrian and bicyclist injuries: an examination based on emergency department data. Accident analysis & prevention, 31(5), pp.505-514.

Tainter, F., Fitzpatrick, C., Gazillo, J., Riessman, R. and Knodler Jr, M., 2020. Using a novel data linkage approach to investigate potential reductions in motor vehicle crash severity–An evaluation of strategic highway safety plan emphasis areas. Journal of Safety Research, 74, pp.9-15.

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y. and Huang, H., 2020. Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. Analytic Methods in Accident Research, 27, p.100123.

Tarko, A. and Azam, M.S., 2011. Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data. Accident Analysis & Prevention, 43(5), pp.1689-1695.

Tin, S.T., Woodward, A., Ameratunga, S., 2013a. Completeness and accuracy of crash outcome data in a cohort of cyclists: a validation study. BMC public health 13, 420.

Tin, S.T., Woodward, A. and Ameratunga, S., 2013b. Incidence, risk, and protective factors of bicycle crashes: Findings from a prospective cohort study in New Zealand. Preventive medicine, 57(3), pp.152-161.

Tsui K, So F, Sze N.N, Wong S, Leung T.F. 2009. Misclassification of injury severity among road casualties in police reports. Accident Analysis & Prevention 41 (1), 84-89.

Wen, H., Xue, G., 2020. Injury severity analysis of familiar drivers and unfamiliar drivers in single-vehicle crashes on the mountainous highways. Accident Analysis & Prevention 144, 105667.

Watanabe, B.L., Patterson, G.S., Kempema, J.M., Magallanes, O. and Brown, L.H., 2019. Is use of warning lights and sirens associated with increased risk of ambulance crashes?

A contemporary analysis using National EMS Information System (NEMSIS) data. *Annals of emergency medicine*, *74*(1), pp.101-109.

Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. Accident Analysis & Prevention 83, 18-25.

Wilson, S.J., Begg, D.J. and Samaranayaka, A., 2012. Validity of using linked hospital and police traffic crash records to analyse motorcycle injury crash characteristics. Accident Analysis & Prevention, 49, pp.30-35.

Winkler, W.E., 2002. Methods for record linkage and bayesian networks (pp. 2659-2665). Technical report, Statistical Research Division, US Census Bureau, Washington, DC.

World Health Organization. Dept. of Violence, Injury Prevention, World Health Organization. Violence, Injury Prevention and World Health Organization, 2009. *Global status report on road safety: time for action*. World Health Organization.

World Health Organization, 2015. *Global status report on road safety 2015*. World Health Organization.

Yannis, G., Papadimitriou, E., Chaziris, A., Broughton, J., 2014. Modeling road accident injury under-reporting in Europe. European transport research review 6, 425-438.

Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety science 63, 50-56.

Zhan, Z.Y., Yu, Y.M., Chen, T.T., Xu, L.J., An, S.L. and Ou, C.Q., 2020. Effects of hourly precipitation and temperature on ambulance response time. *Environmental Research*, *181*, p.108946.

Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J., Hao, W., 2019. Analyzing freeway crash severity using a bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accident Analysis & Prevention 127, 87-95.

CURRICULUM VITAE

NAME: Aryan Hosseinzadeh

ADDRESS: Department of Civil and Environmental Engineering, University of Louisville, W.S. Speed, Louisville, KY, 40292, USA

EDUCATION:

Ph.D., Civil and Environmental Engineering
University of Louisville, Louisville, KY, 2022

M.S., Civil and Environmental Engineering
Tehran Polytechnic, Tehran, Iran, 2016

B.S., Civil and Environmental Engineering
K.N. Toosi University of Technology, Tehran, Iran, 2013

JOURNAL PUBLICATIONS:

A Hosseinzadeh, A Karimpour, R Kluger "Exploring the Impact of Temporal Factors on Micromobility Modes of Trips", Transportation Research Part D, 100, p.103047, 2021.

A Hosseinzadeh, R Kluger "Do EMS times associate with injury severity?" Accident Analysis & Prevention, 153, p.106053, 2021.

A Hosseinzadeh, A Karimpour, R Kluger, R Orthober "Data Linkage for Transportation Safety: Crash, Emergency Medical Services and Trauma Registry" Journal of Safety Research, 2022.

A Hosseinzadeh, M Algomaiah, R Kluger, Z Li "Spatial Analysis of E-Scooter Trips", Journal of Transport Geography 92, 103016, 2021.

A Hosseinzadeh, M Algomaiah, R Kluger, R Li "E-scooters and sustainability: Investigating the relationship between the density of E-scooter trips and characteristics of sustainable urban development" Sustainable Cities and Society, 102624, 2021.

A Hosseinzadeh, M Moeinodini, A Ghasemzadeh "Investigating Factors Affecting Severity of Truck-involved

Crashes: Comparison of the SVM and Logit Model", Journal of Safety Research, 2021.

S Dibaj, **A Hosseinzadeh**, M Mlandenovic and R Kluger "Where Have Shared E-Scooters Taken Us So Far? A Review of Mobility Patterns, Usage Frequency, and Personas", Sustainability, 2021

A Hosseinzadeh "What affects how far individuals walk?", SN Applied Science, 3, 2021.

Y Hatamzadeh, **A Hosseinzadeh** "Toward a Deeper Understanding of Elderly Walking Mode Choice Behavior: An Analysis across Genders in a Case Study of Iran", Journal of Transport and Health, 19, p.100949, 2020.

A Hosseinzadeh, A Baghbani "Walking Trip Generation and Built Environment: A Comparative Study on Trip Purposes", International Journal for Traffic and Transport Engineering,10(3), 2020.

M Habibian, **A Hosseinzadeh** "Walkability Index Across Trip Purposes", Sustainable Cities and Society 42, 216-225, 2018.

CONFERENCE PROCEEDINGS:

A Hosseinzadeh, A Kuzel, R Kluger, R Orthober "Injury Severity Misclassification: Police Officers vs. Emergency Physicians Evaluation, What Drives the Difference?", Transportation Research Board 101st annual meeting, 2022.

A Hosseinzadeh, M Haghani, R Kluger "Exploring Influencing Factors on Emergency Response Time: A Machine Learning Approach", Transportation Research Board 100th annual meeting, 2021.

A Hosseinzadeh, R Kluger "Analyzing the Impact of COVID-19 Pandemic on Micromobility Transportation" in International Conference on Transportation and Development, 2021.

N Abdoli, **A Hosseinzadeh** "Assessing Spatial Equity of Public Transit Demand amid COVID-19" in International Conference on Transportation and Development, 2021.

A Hosseinzadeh, R Kluger "Data Linkage for Traffic Safety in Jefferson County, Kentucky" in International Conference on Transportation and Development, 2021.

A Hosseinzadeh, A Karimpour, R Kluger, R Orthober "A Framework to Link Crashes to Emergency Medical Service Runs and Trauma Admissions for Improved Highway Safety Monitoring and Crash Outcome Assessment" Transportation Research Board 99th Annual Meeting, 2020.

Y Hatamzadeh, **A Hosseinzadeh** "Toward a Deeper Understanding of Elderly Walking Mode Choice Behavior: An Analysis across Genders in a Case Study of Iran", 15th World Conference on Transport Research, 2018.

M Habibian, Z Avaz, and **A Hosseinzadeh**. "Sociological Study of Influence of Citizen's Traffic Ethics on Driving Violations: Case Study of Tehran, Iran." Transportation Research Board 94th Annual Meeting, 2015.

HONORS AND AWARDS:

Graduate Deans' Citation, 2022

Awarded top 40 lifesaver safety scholar, 2021

3MT (Three Minute Thesis) finalist, Fall 2019/Spring 2020

1st place at National Student Steel Bridge (NSSB) competition, May 2012